STA250 Probability and Statistics

Chapter 8 Notes

Sampling Distributions

Asst. Prof. Abdullah YALÇINKAYA

Ankara University, Faculty of Science, Department of Statistics

ayalcinkaya@ankara.edu.tr

2023

STA250 Probability and Statistics

Reference Book

This lecture notes are prepared according to the contents of

"PROBABILITY & STATISTICS FOR ENGINEERS & SCIENTISTS by Walpole, Myers, Myers and Ye"

Populations & Samples

- **A population is the set (possibly infinite) of all possible observations of interest.**
  - Examples: All likely voters in the next election
    All parts produced today
    All sales receipts for November
- **A sample is a subset of a population.**
  - Examples: 1000 voters selected at random for interview
    A few parts selected for destructive testing
    Random receipts selected for audit

- Our goal is to make inferences about the population based on an analysis of the sample.
- Observations in a random sample are made independently and at random. Here, random variables $X_1, X_2, \ldots, X_n$ in the sample all have same distribution as the population, X.

Populations & Samples

- **What would happen if we took many samples of 10 subjects from the population?** Here's how to answer this question:

  - Take a large number of samples of size 10 from the population
  - Calculate the sample mean $\bar{x}$ for each sample.
  - Make a histogram of the values of $\bar{x}$.
  - Examine the shape, center, and spread of the distribution displayed in the histogram.

Populations & Samples

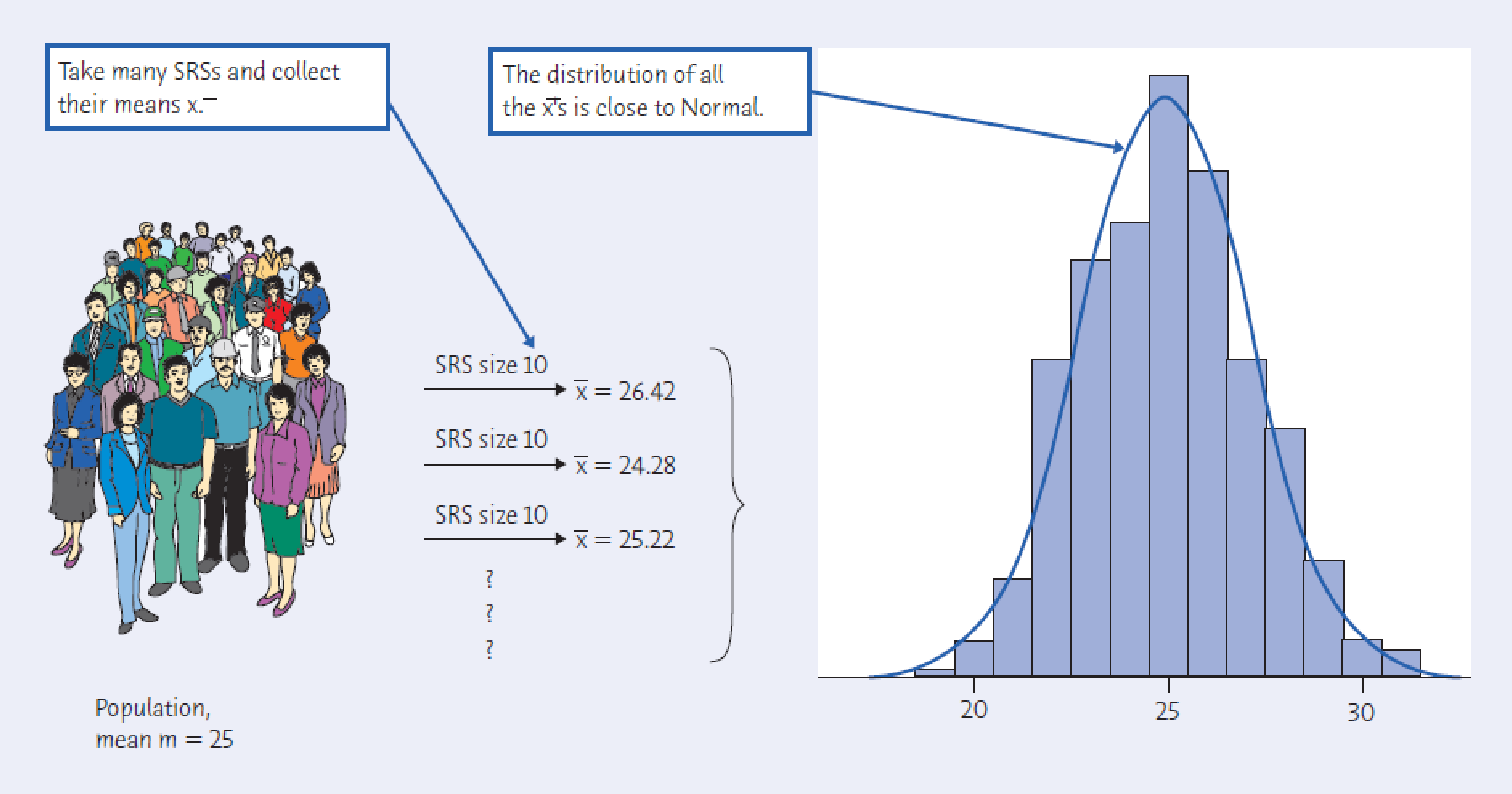


FIGURE 11.2

The idea of a sampling distribution: take many samples from the same population, collect the $\bar{x}$'s from all the samples, and display the distribution of the $\bar{x}$'s. The histogram shows the results of 1000 samples.

Source: The basic Practice of Statistics, Moore,D.,Notz, W.,Flıgner, M.,New York, Sixth Edition.

Populations & Samples

- **What can we say about the shape, center and spread of this distribution?**
- **Shape:** It looks Normal! Detailed examination confirms that the distribution of x from many samples is very close to Normal.
- **Center:** The mean of the 1000 x's is 24.95. That is, the distribution is centered very close to the population mean 25.
- **Spread:** The standard deviation of the 1000 x's is 2.217, notably smaller than the standard deviation 7 of the population of individual subjects.

Populations & Samples

Why Sample?

- **Less time consuming** than census
- **Less costly** to administer than a census
- It is possible to obtain statistical results of a sufficiently **high precision** based on samples.

Simple Random Samples

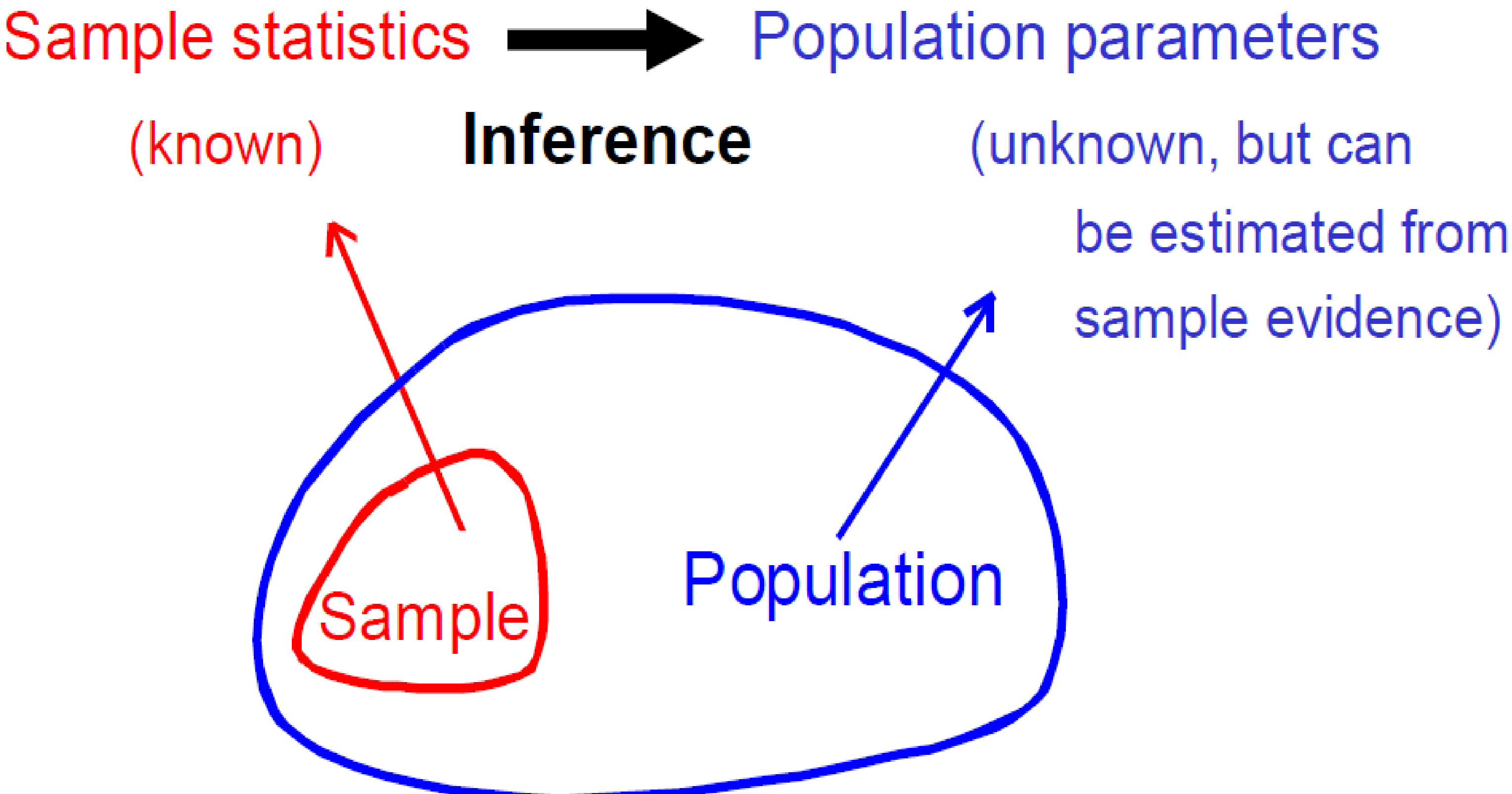
- Every object in the population has an equal chance of being selected
- Objects are selected independently
- Samples can be obtained from a table of random numbers or computer random number generators
- A simple random sample is the ideal against which other sample methods are compared

Other Sampling Methods

- In a **systematic** sample, a random starting point is selected, and then every *k*th item thereafter is selected for the sample
- In a **stratified** sample, the population is divided into several groups, called strata, and then a random sample is selected from each stratum
- In **clustered** sampling, the population is divided into primary units, then samples are drawn from the primary units

Inferential Statistics

- Drawing conclusions and/or making decisions concerning a population based only on sample data
- Our goal is to make <u>inferences</u> about the population based on an analysis of the sample.



Sample Statistics

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

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

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

Sampling Distributions

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

  - The sample consists of independent and identically distributed (i.i.d.) observations X_1, X_2, …, X_n from the population.

  - Based on the sampling distributions of $\bar{x}$ 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 $\bar{x}$ by taking a large number of random samples of size n and plotting the distribution of the $\bar{x}$ values.

Sampling Distribution of the Sample Mean

- **A probability distribution of all possible sample means of a given sample size.**

  - For a given sample size, the mean of all possible sample means selected from a population is equal to the population mean

  - There is less variation in the distribution of the sample mean than in the population distribution

  - The sampling distribution of the sample mean tends to become bell-shaped

Expected Value and Standart Error of the Sample Mean

- **Let X_1, X_2, . . . X_n represent a random sample from a population**
  - The sample mean value of these observations is defined as

$$\boxed{\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i}$$

- **Standard Error of the Mean**
  - Different samples of the same size from the same population will yield different sample means

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

  - 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

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

Proof

Proof: Let X_1, X_2, ,X_n be independent random samples from a population having mean μ and variance σ^2. Then by using rules of expectation

$$\mu_{\bar{X}} = E(\bar{X}) = E(\Sigma\ Xi/n)$$

$$= (1/n)E(X_1+X_2+\ldots..+X_n)$$

$$= (1/n)[E(X_1)+E(X_2)+\ldots..+E(X_n)]$$

$$=(1/n)n\mu = \mu$$

Proof

Proof: Let X_1, X_2, ,X_n be independent random samples from a population having mean μ and variance σ^2. Then by using rules of expectation

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}(\Sigma\, Xi/n)$$

$$= (1/n^2)\mathrm{Var}(X_1+X_2+\ldots\ldots +X_n)$$

$$= (1/n^2)[\mathrm{Var}(X_1)+\mathrm{Var}(X_2)+\ldots\ldots +\mathrm{Var}(X_n)]$$

$$=(1/n^2)n\sigma^2 = \sigma^{2/}n$$

$$\Rightarrow \qquad \sigma_{\bar{X}} = \sigma/\sqrt{n}$$

If the Population is not Normal

- **We can apply the Central Limit Theorem:**

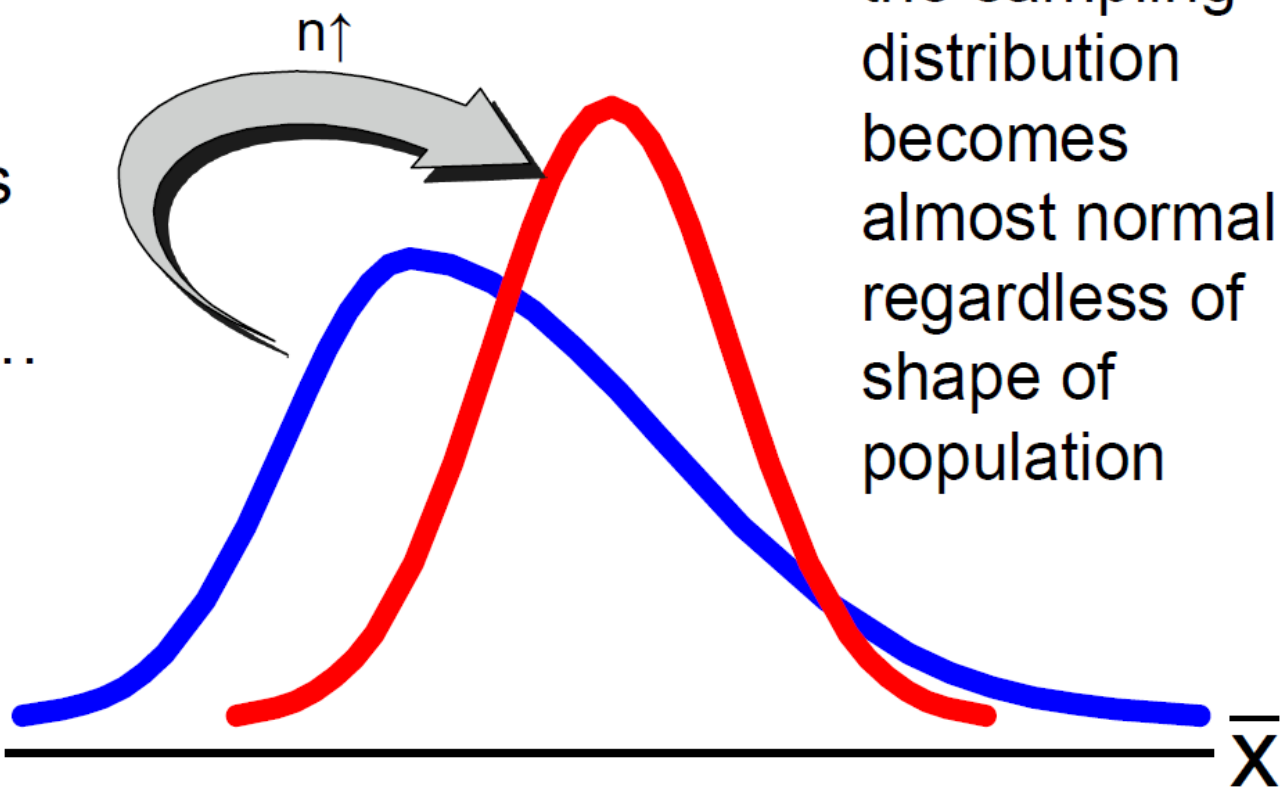
  - If samples of a particular size are selected from any population, the sampling distribution of the sample mean is approximately a normal distribution. The approximation improves with larger samples.

  - If $\bar{x}$ 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 $\bar{x}$ approaches a normal distribution with mean and standard deviation,

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

Central Limit Theorem

As the sample size gets large enough…



the sampling distribution becomes almost normal regardless of shape of population

Central Limit Theorem

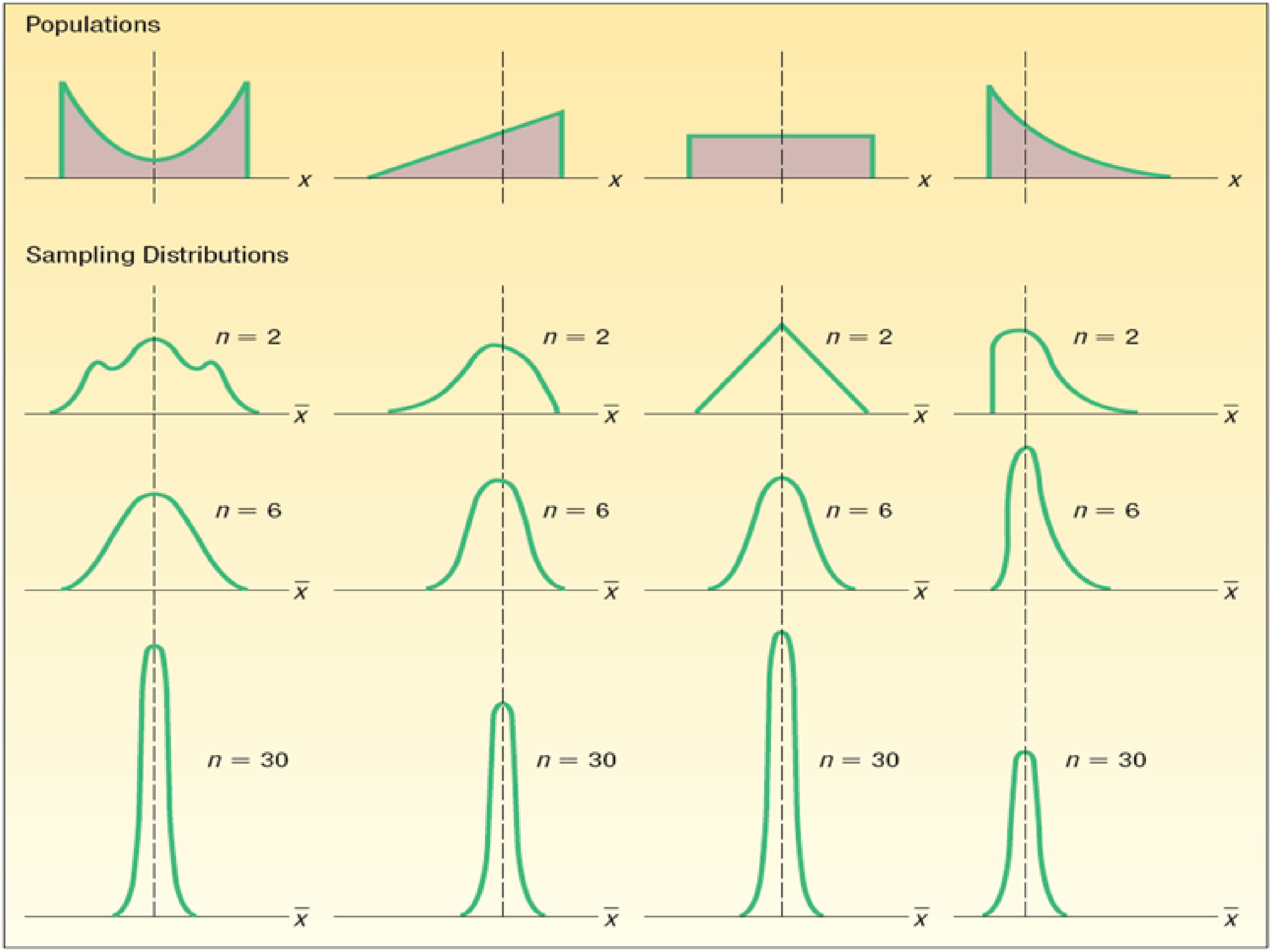
- **How large is large enough!**

  - For most distributions, n > 30 will give a sampling distribution that is nearly normal

  - For normal population distributions, the sampling distribution of the mean is always normally distributed

Central Limit Theorem

- **How large is large enough!**



Z-value for Sampling Distribution of the Mean

- **Z-value for the sampling distribution of $\overline{x}$:**

$$Z = \frac{(\overline{x} - \mu)}{\sigma_{\overline{x}}} = \frac{(\overline{x} - \mu)}{\frac{\sigma}{\sqrt{n}}}$$

Where:

$$\bar{x} = sample\ mean$$
$$\mu = population\ mean$$
$$\sigma = population\ standard\ devation$$
$$n = sample\ size$$

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 μ = 6 cm in diameter, and suppose σ is known to be .15 cm. If a random sample of 80 parts has $\bar{x}$ = 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?

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

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

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

Example 1

An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed, with mean equal to 800 hours and a standard deviation of 40 hours. Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours.

***Solution*:** The sampling distribution of $\bar{X}$ will be approximately normal, with $\mu_{\bar{X}} = 800$ and $\sigma_{\bar{X}} = 40/\sqrt{16} = 10$. The desired probability is given by the area of the shaded

Example 1

Solution: The sampling distribution of $\bar{X}$ will be approximately normal, with $\mu_{\bar{X}} = 800$ and $\sigma_{\bar{X}} = 40/\sqrt{16} = 10$. The desired probability is given by the area of the shaded region in Figure 8.2.

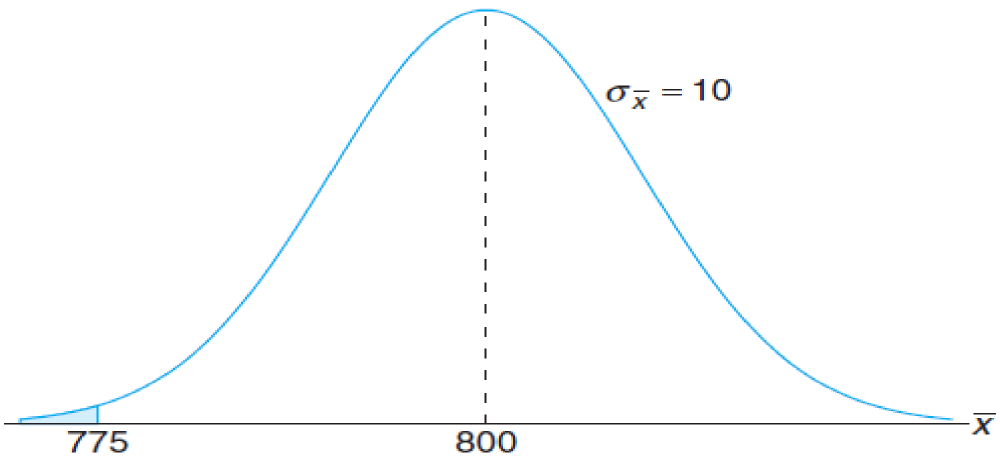


Figure 8.2: Area for Example 8.4.

Corresponding to $\bar{x} = 775$, we find that

$$z = \frac{775 - 800}{10} = -2.5,$$

and therefore

$$P(\bar{X} < 775) = P(Z < -2.5) = 0.0062.$$

Example 2

Traveling between two campuses of a university in a city via shuttle bus takes, on average, 28 minutes with a standard deviation of 5 minutes. In a given week, a bus transported passengers 40 times. What is the probability that the average transport time was more than 30 minutes? Assume the mean time is measured to the nearest minute.

Solution: In this case, $\mu = 28$ and $\sigma = 3$. We need to calculate the probability $P(X > 30)$ with $n = 40$. Since the time is measured on a continuous scale to the nearest minute, an $\bar{x}$ greater than 30 is equivalent to $\bar{x} \geq 30.5$. Hence,

$$P(\bar{X} > 30) = P\left(\frac{\bar{X} - 28}{5/\sqrt{40}} \geq \frac{30.5 - 28}{5/\sqrt{40}}\right) = P(Z \geq 3.16) = 0.0008.$$

Example 2

There is only a slight chance that the average time of one bus trip will exceed 30 minutes. An illustrative graph is shown in Figure 8.4.

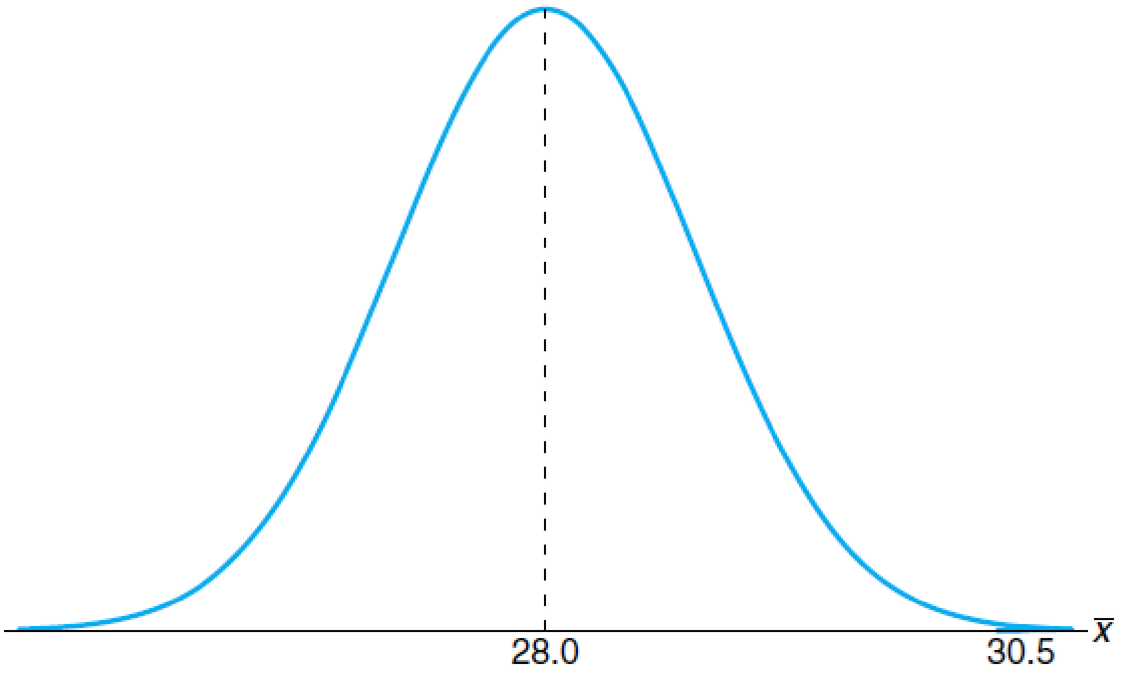
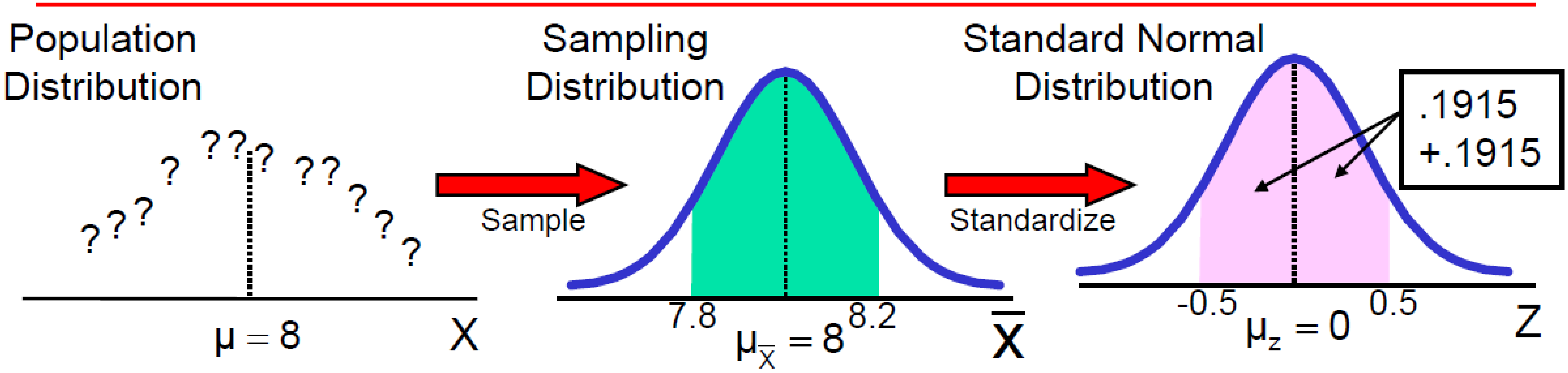


Figure 8.4: Area for Example 8.5.

Summary


Population
Distribution
? ? ? ? ? ? ? ? ? ? ? ? ? ?
$\mu = 8$
X
Sample
Sampling
Distribution
7.8
8.2
$\mu_{\bar{X}} = 8$
$\bar{X}$
Standardize
Standard Normal
Distribution
.1915
+.1915
-0.5
0.5
$\mu_z = 0$
Z

Example 3

Shipping weights of packed cartons of hand-crafted wood furniture have a mean weight of 215 Kg. The distribution is normal and the standard deviation is 20 Kg.

a- Find the probability of any one carton reaching a <u>weight</u> of 212 kg or more.

Since X(weight of a caton)~ Normal

$$P(X \geq 212\ Kg) = P(Z \geq \frac{212-215}{20})$$

$$= P(Z \geq -0.15) = 0.5 + 0.0596 = 0.5596$$

Example 3 continue…

b- Find the probability that a group of 10 randomly selected cartons reaching a mean weight of 212 Kg. or more.

Since X~ Normal → $\bar{X}$ ~ Normal

$$\sigma_{\bar{X}} = \sigma/\sqrt{n} = 20/\sqrt{10} = 6.32$$

$$P(\bar{X} \geq 212 \text{ Kg}) = P\left(Z \geq \frac{212-215}{6.32}\right)$$

$$= P(Z \geq -0.47) = 0.5 + 0.1808 = 0.6808$$

Sampling Distribution for Sample Proportion

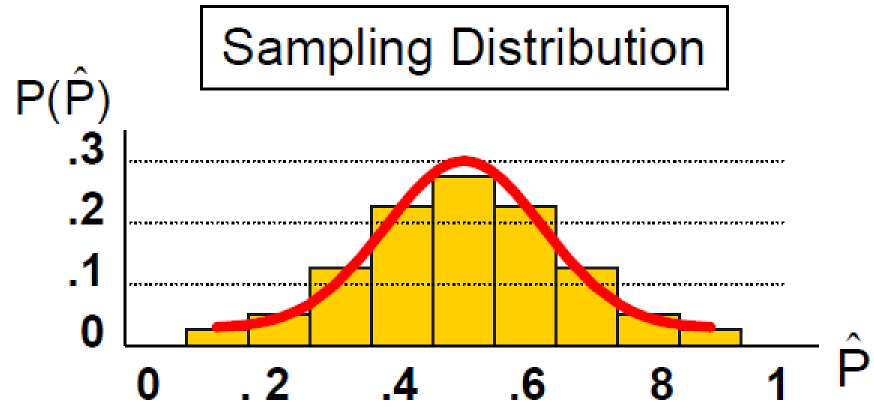
- P = the proportion of the population having some characteristic(success)
- Sample proportion ($\hat{p}$) is an estimate of P:

- $$\boxed{\hat{p} = \frac{X}{n} = \frac{\text{number of items in the sample having the characteristic of interest}}{\text{sample size}}}$$

- $0 \leq \hat{p} \leq 1$
- $\hat{p}$ has a binomial distribution, but can be approximated by a normal distribution when $min[nP, n(1-P)] \geq 5$

Sampling Distirbution for Sample Proportion

- Normal approximation:



Properties:

$$E(\hat{P}) = p$$

and

$$\sigma^2_{\hat{P}} = \mathrm{Var}\left(\frac{X}{n}\right) = \frac{P(1-P)}{n}$$

(where P = population proportion)

Sampling Distribution for Sample Proportion

Proof: Let random variable X distributed Binomial with parameters n and p. Then

$$\mu_{\hat{p}} = E(\hat{p}) = E(X/n) = (1/n)E(X) = (1/n)np = p$$

$$\sigma_{\hat{p}}^2 = Var(\hat{p}) = Var(X/n) = (1/n^2)Var(X)$$

$$= (1/n^2)np(1-p) = \frac{p(1-p)}{n}$$

$$\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$$

Z-Value for Proportions

Standardize $\hat{P}$ to a Z value with the formula:

$$Z = \frac{\hat{P} - P}{\sigma_{\hat{P}}} = \frac{\hat{P} - P}{\sqrt{\dfrac{P(1-P)}{n}}}$$

Example (1 of 1)

- If the true proportion of voters who support Proposition A is P = .4, what is the probability that a sample of size 200 yields a sample proportion between .40 and .45?

- i.e.: **if P = .4 and n = 200, what is**

  $$P(.40 \le \hat{P} \le .45)\ ?$$

Example (1 of 2)

- if $P = .4$ and $n = 200$, what is $P(.40 \leq \hat{P} \leq .45)$?

Find $\sigma_{\hat{P}}$:

$$\sigma_{\hat{P}} = \sqrt{\frac{P(1-P)}{n}} = \sqrt{\frac{.4(1-.4)}{200}} = .03464$$

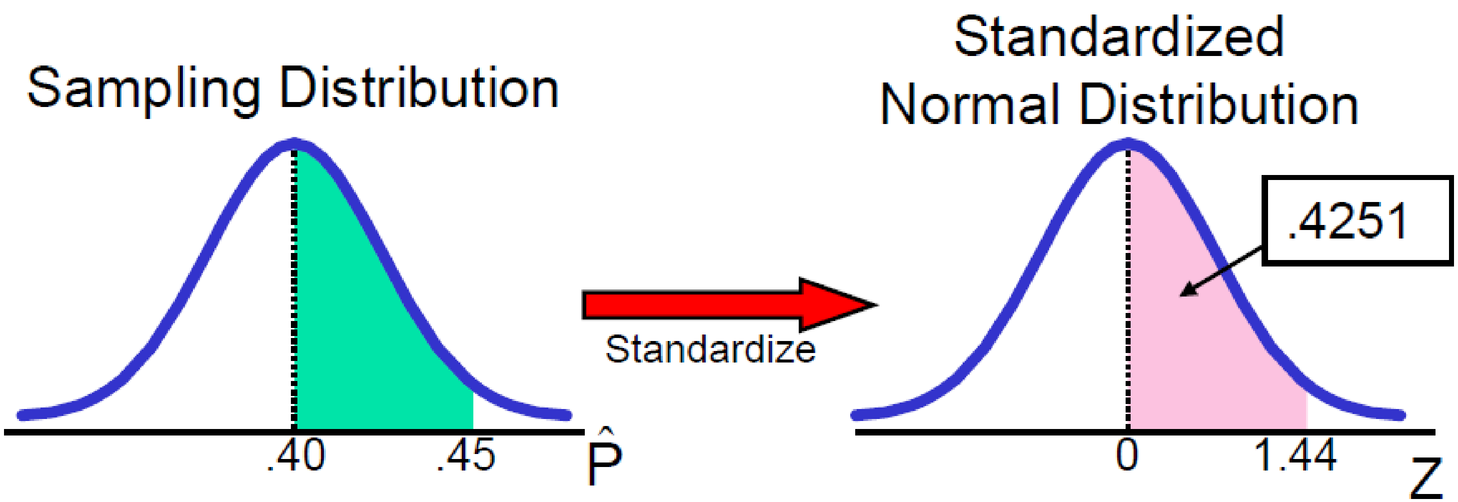
Convert to standard normal:

$$P(.40 \leq \hat{P} \leq .45) = P\left(\frac{.40-.40}{.03464} \leq Z \leq \frac{.45-.40}{.03464}\right)$$
$$= P(0 \leq Z \leq 1.44)$$

Example (1 of 3)

- if $p = .4$ and $n = 200$, what is $P(.40 \le \hat{P} \le .45)$?

Use standard normal table: $P(0 \le Z \le 1.44) = \boxed{.4251}$



Sampling Distribution Summary

- **Normal distribution: Sampling distribution of $\bar{x}$ 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 $\bar{x}$ 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.**

Next Lesson

- **Estimation**

See you☺