Ch8. Large-Sample Estimation

In previous chapters, you learned about the probability distributions of random variables and the sampling distributions of several statistics that, for large sample sizes, can be approximated by a normal distribution according to the Central Limit Theorem. This chapter presents a method for estimating population parameters and illustrates the concept with practical examples. The Central Limit Theorem and the sampling distributions presented in Chapter 7 play a key role in evaluating the reliability of the estimates.

윤 연 옥

Introduction

- **Parameters** are numerical descriptive measures for populations.
  - For the normal distribution, the location and shape are described by μ and σ.
  - For a binomial distribution consisting of n trials, the location and shape are determined by $\boldsymbol{p}$.
- Often the values of parameters that specify the exact form of a distribution are unknown.
- You must rely on the sample to learn about these parameters.

Types of Inference

- **Estimation:**

  - Estimating or predicting the value of the parameter

  - "What is (are) the most likely values of μ or p ?"

Ex) A consumer wants to estimate the average price of similar homes in her city before putting her home on the market.

Estimation: Estimate μ, the average home price.

- **Hypothesis Testing:**

  - Deciding about the value of a parameter based on some preconceived idea.

  - "Did the sample come from a population with $\mu = 5$ or $p = .2$?"

Ex) A manufacturer wants to know if a new type of steel is more resistant to high temperatures than an old type was.

Hypothesis test: Is the new average resistance, μ_N equal to the old average resistance, μ_0?

Types of Inference

- Whether you are estimating parameters or testing hypotheses, statistical methods are important because they provide:
  - Methods for making the inference
  - A numerical measure of the goodness or reliability of the inference

Definition

- An **estimator** is a rule, usually a formula, that tells you how to calculate the estimate based on the sample.

- **Point estimation:** A single number is calculated to estimate the parameter.
  - ✓ point estimator: the rule or formula that describes this calculation
  - ✓ point estimate: the resulting number calculated from sample

  ex) $\bar{X}$: estimator of μ $\bar{x}$: estimate of μ

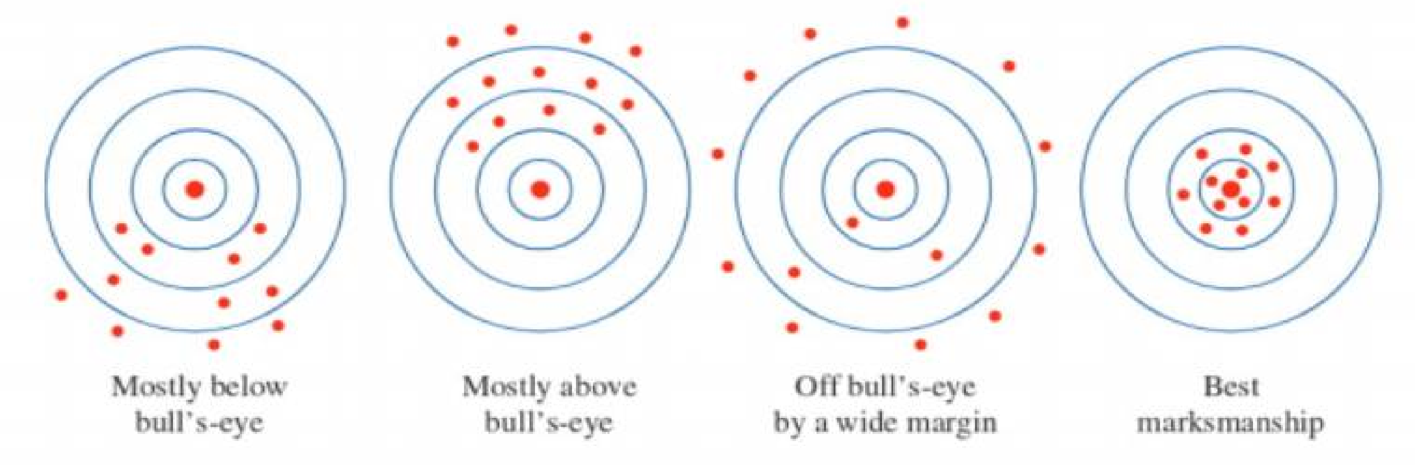
- **Interval estimation:** Two numbers are calculated to create an interval within which the parameter is expected to lie.
  - ✓ interval estimator :the rule or formula that describes this calculation
  - ✓ interval estimate or confidence interval: resulting pair of numbers calculated from sample

Properties of Point Estimators

- Since an estimator is calculated from sample values, it varies from sample to sample according to its **sampling distribution.**

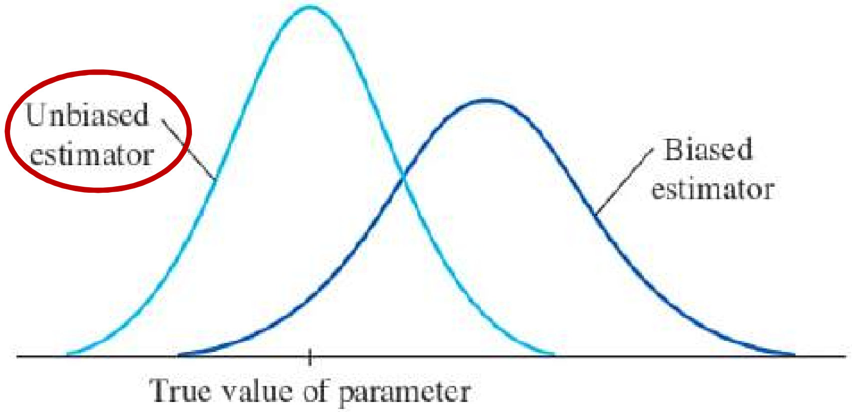
- An **estimator** is **unbiased** if the mean of its sampling distribution equals the parameter of interest.
  - It does not systematically overestimate or underestimate the target parameter.

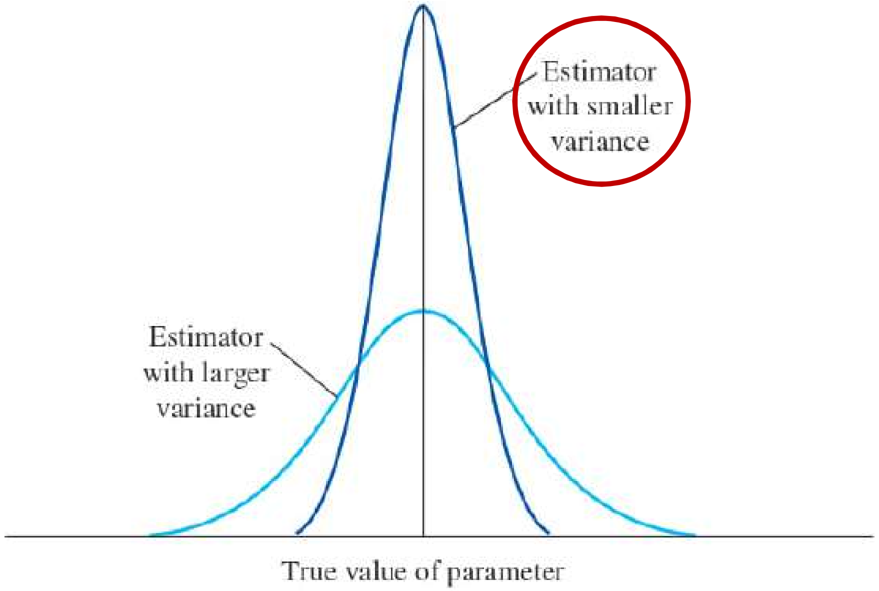
Ex) Firing a gun at a target. Which one is best?



Properties of Point Estimators

- Of all the **unbiased** estimators, we prefer the estimator whose sampling distribution has the **smallest spread** or **variability**.





Measuring the Goodness of an Estimator

- The distance between an estimate and the true value of the parameter is the **error of estimation.**

- In this chapter, the sample sizes are large, so that our ***unbiased*** estimators will have **normal** distributions.
  - Because of the Central Limit Theorem.

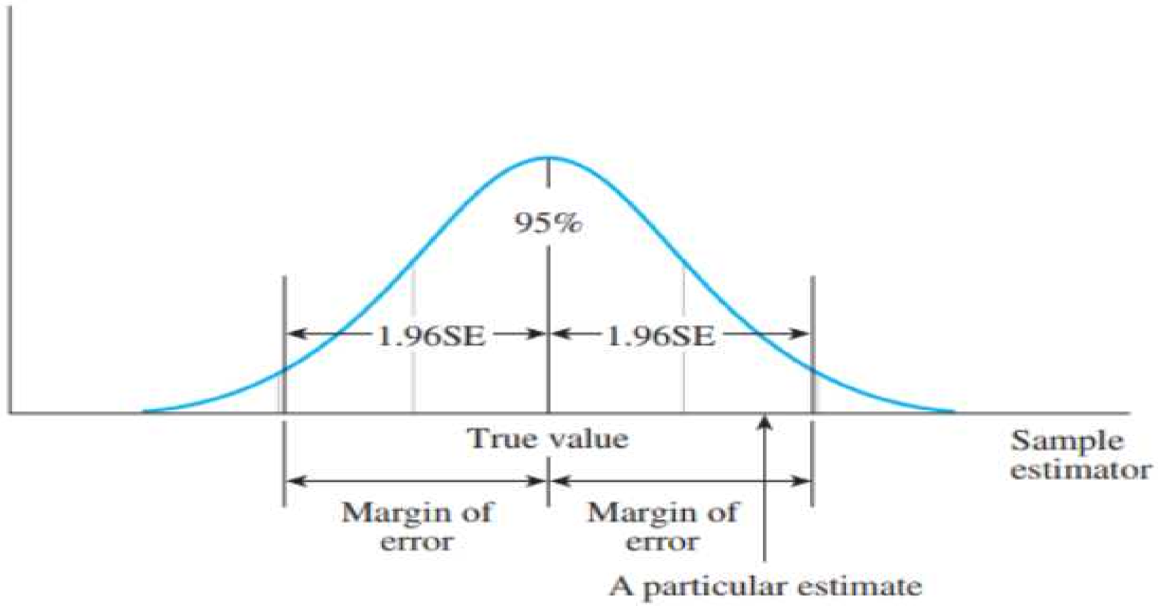
Margin of Error

- For ***unbiased*** estimators with normal sampling distributions, 95% of all point estimates will lie within 1.96 standard deviations of the parameter of interest.

- **95% margin of error(or margin of error)**

: The maximum error of estimation calculated as

$$1.96 \times \text{ std error of the estimator}$$



Estimating Means and Proportions

- For a quantitative population,

Point estimator of population mean $\mu: \bar{x}$

95% Margin of error$(n \geq 30)$: $\pm 1.96 \frac{s}{\sqrt{n}}$

SE

- For a binomial population,

Point estimator of population proportion $p: \hat{p} = x/n$

95% Margin of error $(n \geq 30)$: $\pm 1.96 \sqrt{\frac{\hat{p}\hat{q}}{n}}$

Assumption: $n\hat{p} > 5$ *and* $n\hat{q} > 5$

SE

Example

- A random sample of n=50 polar bears produced an average weight of 980 pounds with a standard deviation of 105 pounds.
- Estimate the average weight of all Arctic polar bears and margin of error for the estimate.

Sol)

- The point estimate of μ , the average weight of all Arctic polar bears, is $\bar{x}$ = 980 pounds.

- The margin of error is estimated as

$$1.96\ \ SE\ = 1.96\left(\frac{s}{\sqrt{n}}\right) = 1.96\left(\ \frac{105}{\sqrt{50}}\ \right) = 29.1\ \ \ \approx 29\ pounds$$

- We can be fairly confident that the sample estimate of 980 pounds is within ± 29 pounds of the population mean.

Example

- In a random sample of n=100 adults, 73% of the sample indicated that global warming is a very serious problem.
- Estimate the true population proportion of adults who believe that global warming is a very serious problem, and find the margin of error for the estimate.

Sol)

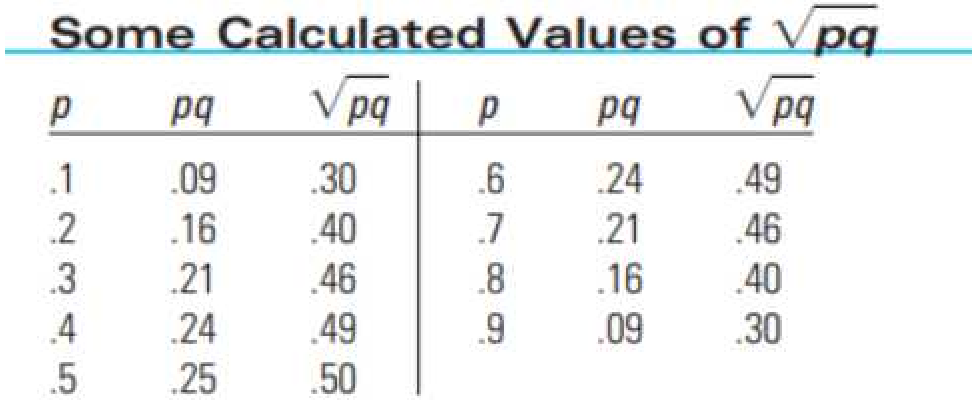
- p : the proportion of individuals in the population who believe that global warming is a very serious problem.
- $\hat{p} = .73$
- The margin of error is estimated as

$$1.96 \ SE = 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}} = 1.96\sqrt{\frac{.73(.27)}{100}} = .09$$

- We can be fairly confident that the estimate of .73 is within$\pm .09$ of the true value of p .

Maximum margin of error in Proportion estimation

- The margin of error for $\hat{p}$ will be a maximum when p= .5.



Some Calculated Values of $\sqrt{pq}$

p	pq	$\sqrt{pq}$	p	pq	$\sqrt{pq}$
.1	.09	.30	.6	.24	.49
.2	.16	.40	.7	.21	.46
.3	.21	.46	.8	.16	.40
.4	.24	.49	.9	.09	.30
.5	.25	.50			

- When p=.5,

$$1.96\,SE = 1.96\sqrt{\frac{.5\,(.5)}{n}} \approx 2\sqrt{\frac{.5\,(.5)}{n}} = \sqrt{\frac{1}{n}}$$

ex) n=1000, margin of error for $\hat{p}$ is .031 : ± 3 % points of the true population

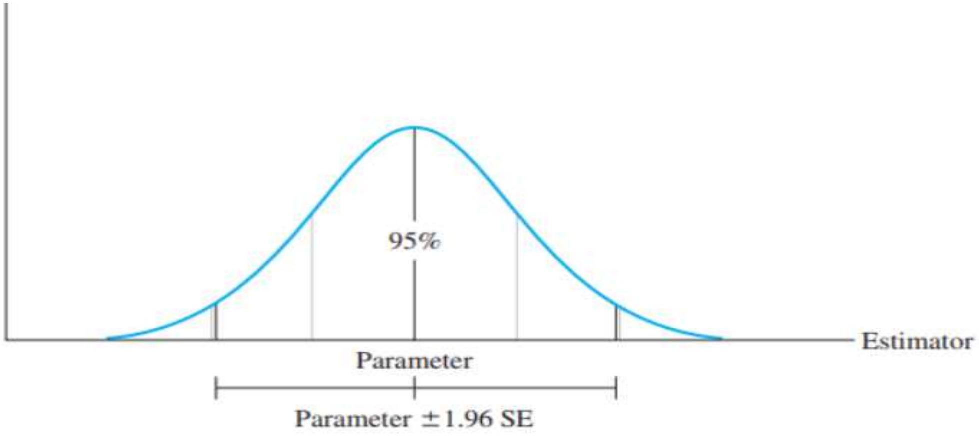
Interval Estimation

- Create an interval (a, b) so that you are fairly sure that the parameter lies between these two values.
- "Fairly sure" is means "with high probability", measured using the confidence coefficient, $1 - \alpha$.

  ◆ Def: The probability that a confidence interval will contain the estimated parameter is called the confidence coefficient.

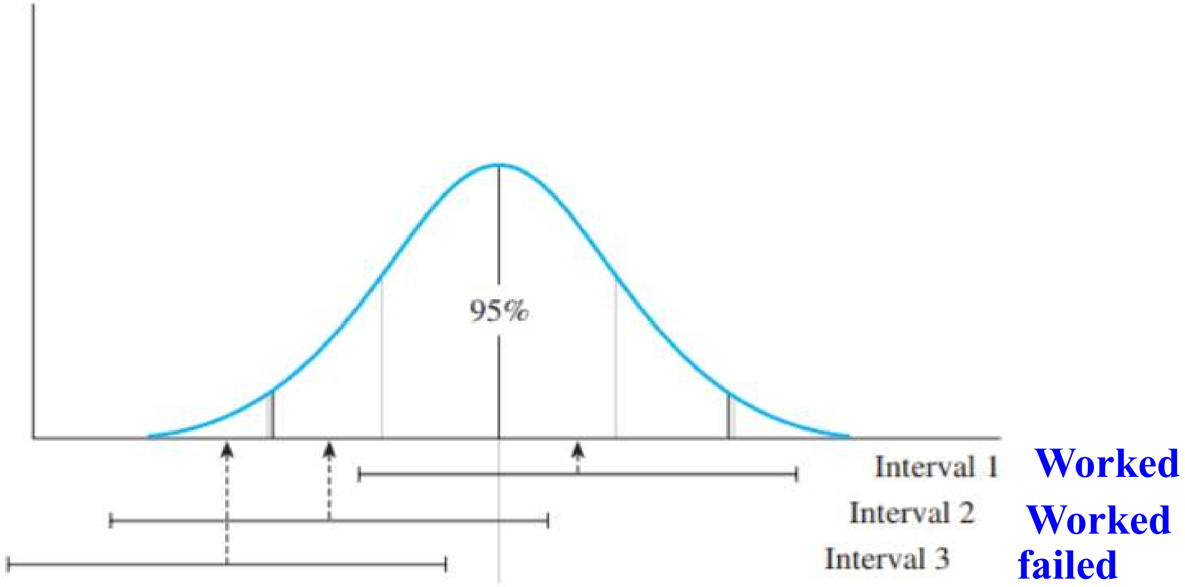
  ✓ Usually, 1-α = .90, .95, .99

- Suppose 1-α = .95 and that the estimator has a normal distribution.



Interval Estimation

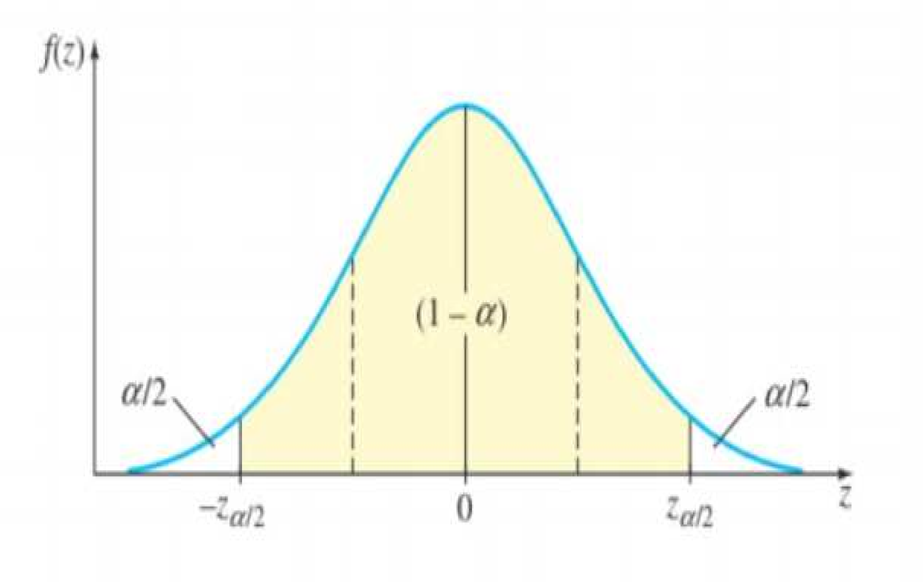
- Since we don’t know the value of the parameter, consider **Estimator ± 1.96SE** which has a variable center.

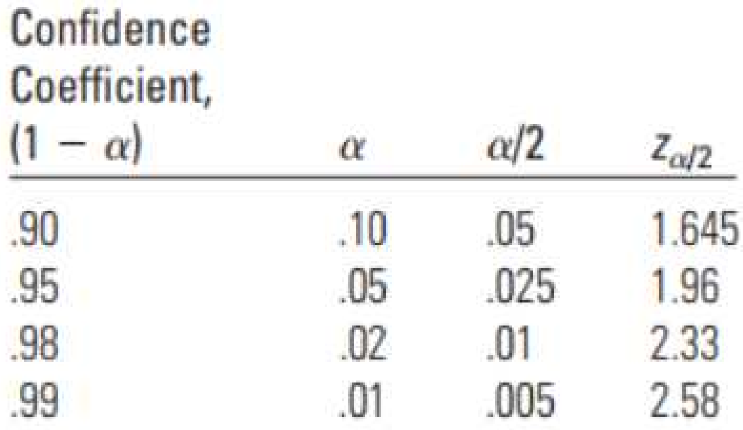


- Only if the estimator falls in the tail areas will the interval fail to enclose the parameter. This happens only 5% of the time.

Change the Confidence Level

- To change to a general confidence level, 1-α, pick a value of *z* that puts area 1-α in the center of the *z* distribution.





Confidence Coefficient, $(1-\alpha)$	α	$\alpha/2$	$z_{\alpha/2}$
.90	.10	.05	1.645
.95	.05	.025	1.96
.98	.02	.01	2.33
.99	.01	.005	2.58

100(1-α)% Confidence Interval: Estimator ± $z_{\alpha/2}$SE

Confidence Intervals for Means and Proportions

- For a quantitative population(large sample),

Confidence interval for a population mean μ:

$$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$$

- For a binomial population,

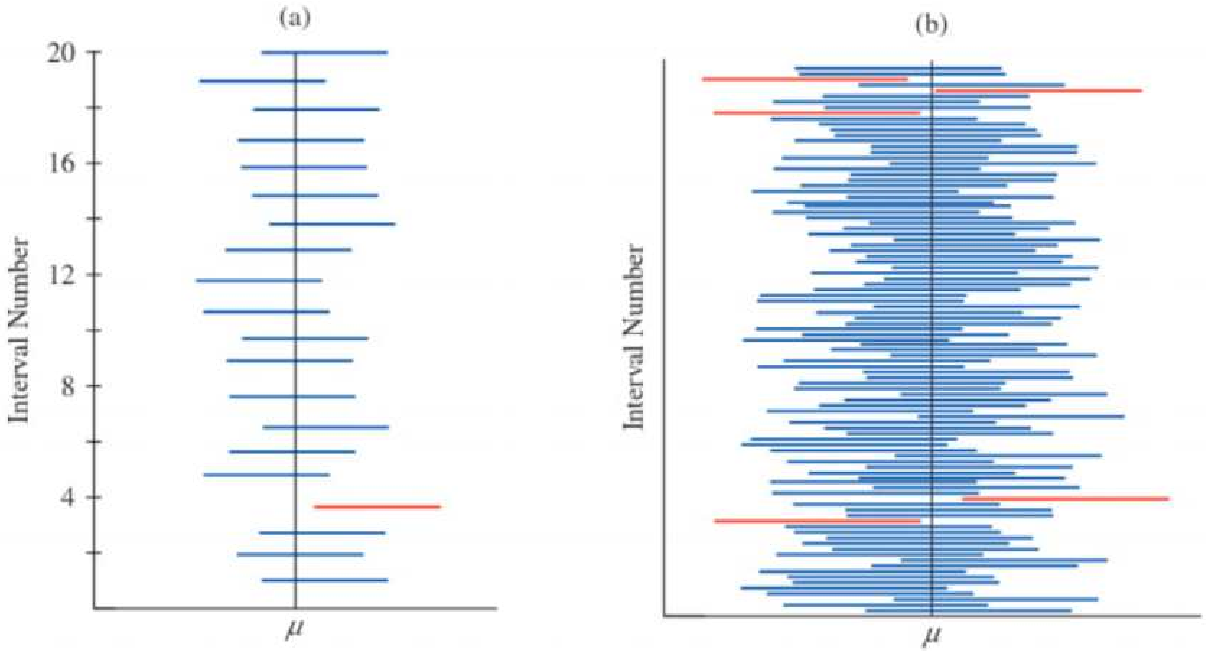
Confidence interval for a population proportion p:

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

✓ Check large sample, $n\hat{p} > 5$ and $n\hat{q} > 5$

Interpreting the Confidence Interval

- What does it mean to say you are “95% confident” that the true value of the population mean μ is within a given interval?



(a) 20 intervals

(b) 100 intervals

- A 95% confidence interval tells you that, if you were to construct many of these intervals, 95% of them would enclose the population mean.

Example

- A dietician selected a random sample of n=50 male adults and found that their average daily intake of dairy products was $\bar{x} = 756$ grams per day with a standard deviation of $s = 35$ grams per day.
- Construct a 95% confidence interval for the mean daily intake of dairy products for men.

Sol)

- Since the sample size of n=50 is large, the distribution of the sample mean $\bar{x}$ is approximately normally distributed

- The approximate 95% confidence interval is

$$\bar{x} \pm 1.96 \frac{s}{\sqrt{n}} = 756 \pm 1.96 \frac{35}{\sqrt{50}} = 756 \pm 9.70$$

✓ The 95% confidence interval for μ is from 746.30 to 765.70 grams per day.

Example –continued

- Construct a 99% confidence interval for the mean daily intake of dairy products for men(μ).

Sol)

- The approximate 99% confidence interval is

$$\bar{x} \pm 2.58 \frac{s}{\sqrt{n}} = 756 \pm 2.58 \frac{35}{\sqrt{50}} = 756 \pm 12.77$$

✓ The 99% confidence interval for μ is from 743.23 to 768.77 grams per day.

The interval must be wider to provide for the increased confidence that is does indeed enclose the true value of μ.

Example

- A random sample of 985 “likely” were surveyed. Of those surveyed, 592 indicated that they intended to vote for the Republican candidate in the upcoming election.
- Construct a 90% confidence interval for p , the proportion of likely voters in the population who intend to vote for the Republican candidate.

Sol)

- The point estimate for p is $\hat{p} = x/n = \frac{592}{985} = .601$
- 90% confidence interval for p is

$$\hat{p} \pm z_{.05}\sqrt{\frac{\hat{p}\hat{q}}{n}} = .601 \pm 1.645\sqrt{\frac{(.601)(.399)}{985}} = .601 \pm .026$$

✓ The 90% confidence interval for the percentage of likely voters who intend to vote for the Republican candidate is between 57.5% and 62.7%.

Estimating the Difference between Two Means

- Sometimes we are interested in comparing the means of two populations.

  - The average growth of plants fed using two different nutrients.

  - The average scores for students taught with two different teaching methods.

- Let define

  - A random sample of size n_1 drawn from population 1 with mean μ_1 and variance σ_1^2.

  - A random sample of size n_2 drawn from population 2 with mean μ_2 and variance σ_2^2.

Estimating the Difference between Two Means

- We compare the two averages by making inferences about μ_1-μ_2, the difference in the two population averages.
  - If the two population averages are the same, then μ_1-$\mu_2 = 0$.
  - The best estimate of μ_1-μ_2 is the difference in the two sample means, $\bar{x}_1 - \bar{x}_2$

The Sampling Distribution of $\bar{x}_1 - \bar{x}_2$

1. The mean of $\bar{x}_1 - \bar{x}_2$ is $\mu_1 - \mu_2$, the difference in the population means

2. The standard deviation of $\bar{x}_1 - \bar{x}_2$ is $SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

3. If the sample sizes are large, the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is approximately normal, and SE can be estimated as $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ (by CLT)

4. If the sampled populations are normally distributed, then sampling distribution of $\bar{x}_1 - \bar{x}_2$ is exactly normally distributed, regardless of sample size

Estimating μ$_1$-μ$_2$

- For large samples, point estimates and their margin of error as well as confidence intervals are based on the standard normal (*z*) distribution.

Point estimate for $\mu_1 - \mu_2$: $\bar{x}_1 - \bar{x}_2$

95% Margin of Error : $\pm\, 1.96\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$

Confidence interval for $\mu_1 - \mu_2$:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Example

- Compare the average daily intake of dairy products of men and women using a 95% confidence interval.

Avg Daily Intakes	Men	Women
Sample size	50	50
Sample mean	756	762
Sample Std Dev	35	30

$$(\bar{x}_1 - \bar{x}_2) \pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = (756 - 762) \pm 1.96\sqrt{\frac{35^2}{50} + \frac{30^2}{50}}$$

$$= -6 \pm 12.78$$

$$\Rightarrow \quad -18.78 < \mu_1 - \mu_2 < 6.78$$

Example(continued)

$$\boxed{-18.78 < \mu_1 - \mu_2 < 6.78}$$

- Could you conclude, based on this confidence interval, that there is a difference in the average daily intake of dairy products for men and women?
- The confidence interval contains the value μ_1-μ_2= **0.** Therefore, it is possible that $\mu_1 = \mu_{2.}$

  You would not want to conclude that there is a difference in average daily intake of dairy products for men and women.

Estimating the Difference between Two Proportions

- Sometimes we are interested in comparing the proportion of "successes" in two binomial populations.
  - The germination rates of untreated seeds and seeds treated with a fungicide.
  - The proportion of male and female voters who favor a particular candidate for governor.

- Let define
  - A random sample of size n_1 drawn from binomial population 1 with parameter p_1.
  - A random sample of size n_2 drawn from binomial population 2 with parameter p_2.

Estimating the Difference between Two Proportions

- We compare the two proportions by making inferences about p_1-p_2, the difference in the two population proportions.
  - If the two population proportions are the same, then p_1-$p_2 = 0$.
  - The best estimate of p_1-p_2 is the difference in the two sample proportions,

$$\hat{p}_1 - \hat{p}_2 = \frac{x_1}{n_1} - \frac{x_2}{n_2}$$

The Sampling Distribution of p_1-p_2

1. The mean of $\hat{p}_1 - \hat{p}_2$ is $p_1 - p_2$, the difference in the population proportions.

2. The standard deviation of $\hat{p}_1 - \hat{p}_2$ is $SE = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$

3. If the sample sizes are large, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal, and SE can be estimated as $SE = \sqrt{\frac{\widehat{p_1}\widehat{q_1}}{n_1} + \frac{\widehat{p_2}\widehat{q_2}}{n_2}}$ (by CLT)

✓ $\hat{p}_1$ and $\hat{p}_2$ should be approximately normal :
$n_1\hat{p}_1 > 5,\ n_1\hat{q}_1 > 5$ *and* $n_2\hat{p}_2 > 5,\ n_2\hat{q}_2 > 5$

Estimating p_1-p_2

- For large samples, point estimates and their margin of error as well as confidence intervals are based on the standard normal (z) distribution

Point estimate for $p_1 - p_2$: $\hat{p}_1 - \hat{p}_2$

95% *Margin of Error:* $\pm\ 1.96\sqrt{\dfrac{\hat{p}_1\hat{q}_1}{n_1}+\dfrac{\hat{p}_2\hat{q}_2}{n_2}}$

Confidence interval for $p_1 - p_2$:

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}+\frac{\hat{p}_2\hat{q}_2}{n_2}}$$

Example

- Compare the proportion of male and female college students who said that they had played on a soccer team during their K-12 years using a 99% confidence interval

Youth Soccer	Male	Female
Sample size	80	70
Played soccer	65	39

$$(\hat{p}_1 - \hat{p}_2) \pm 2.58\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

$$= \left(\frac{65}{80} - \frac{39}{70}\right)) \pm 2.58\sqrt{\frac{.81(.19)}{80} + \frac{.56(.44)}{70}} = .25 \pm .19$$

or $.06 < p_1 - p_2 < .44$

Example (continued)

$$.06 < p_1 - p_2 < .44$$

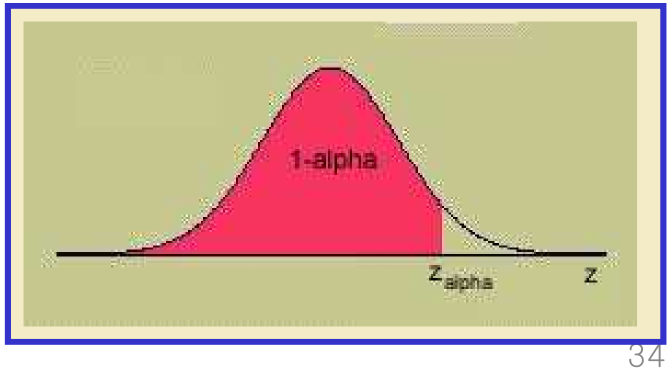
- Could you conclude, based on this confidence interval, that there is a difference in the proportion of male and female college students who said that they had played on a soccer team during their K-12 years?
- The confidence interval does not contains the value $\boldsymbol{p_1\text{-}p_2} = \mathbf{0}$**.** Therefore, it is not likely that $\boldsymbol{p_1 = p_2}$. You would conclude that there is a difference in the proportions for males and females.

- A higher proportion of males than females played soccer in their youth.

One Sided Confidence Bounds

- Confidence intervals are by their nature **two-sided** since they produce upper and lower bounds for the parameter.
- **One-sided bounds** can be constructed simply by using a value of *z* that puts α rather than α/2 in the tail of the *z* distribution.

$$\text{LCB: Estimator} - z_{\alpha} \times (\text{Std Error of Estimator})$$
$$\text{UCB: Estimator} + z_{\alpha} \times (\text{Std Error of Estimator})$$



Choosing the Sample Size

- The total amount of relevant information in a sample is controlled by two factors:

  - The **sampling plan** or **experimental design**: the procedure for collecting the information

  - The **sample size** ***n***: the amount of information you collect.

- In a statistical estimation problem, the accuracy of the estimation is measured by the **margin of error** or the **width of the confidence interval.**

Choosing the Sample Size

1. Determine the size of the **margin of error**, ***B***, that you are willing to tolerate.

2. Choose the sample size by solving for n (or $n = n_1 = n_2$ for two-sample) in the inequality: **1.96 SE ≤ *B*,** where SE is a function of the sample size n.

3. For quantitative populations, estimate the population standard deviation using a previously calculated value of ***s*** or the range approximation **$\sigma \approx$ Range / 4.**

4. For binomial populations, use the conservative approach and approximate p using the value ***p* = .5**.

Example

- A producer of PVC pipe wants to survey wholesalers who buy his product in order to estimate the proportion who plan to increase their purchases next year. What sample size is required if he wants his estimate to be within .04 of the actual proportion with probability equal to .95?

Sol)

$$1.96\sqrt{\frac{pq}{n}} \le .04 \quad \Rightarrow \quad 1.96\sqrt{\frac{.5(.5)}{n}} \le .04$$

$$\Rightarrow \quad \sqrt{n} \ge \frac{1.96\sqrt{.5(.5)}}{.04} = 24.5$$

$$\Rightarrow n \ge 24.5^2 = 600.25$$

✓ *He should survey at least 601 wholesalers.*

Example

A personnel director wishes to compare the effectiveness of two methods of training industrial employees to perform a certain assembly operation. Employees are to be divided into two equal groups: the first receiving training method 1 and the second training method 2. Each will perform the assembly operation, and the length of assembly time will be recorded. It is expected that the assembly times for both groups will have a range of approximately 8 minutes. For the estimate of the difference in mean times to assemble to be correct to within 1 minute with a probability equal to .95, how many workers must be included in each training group?

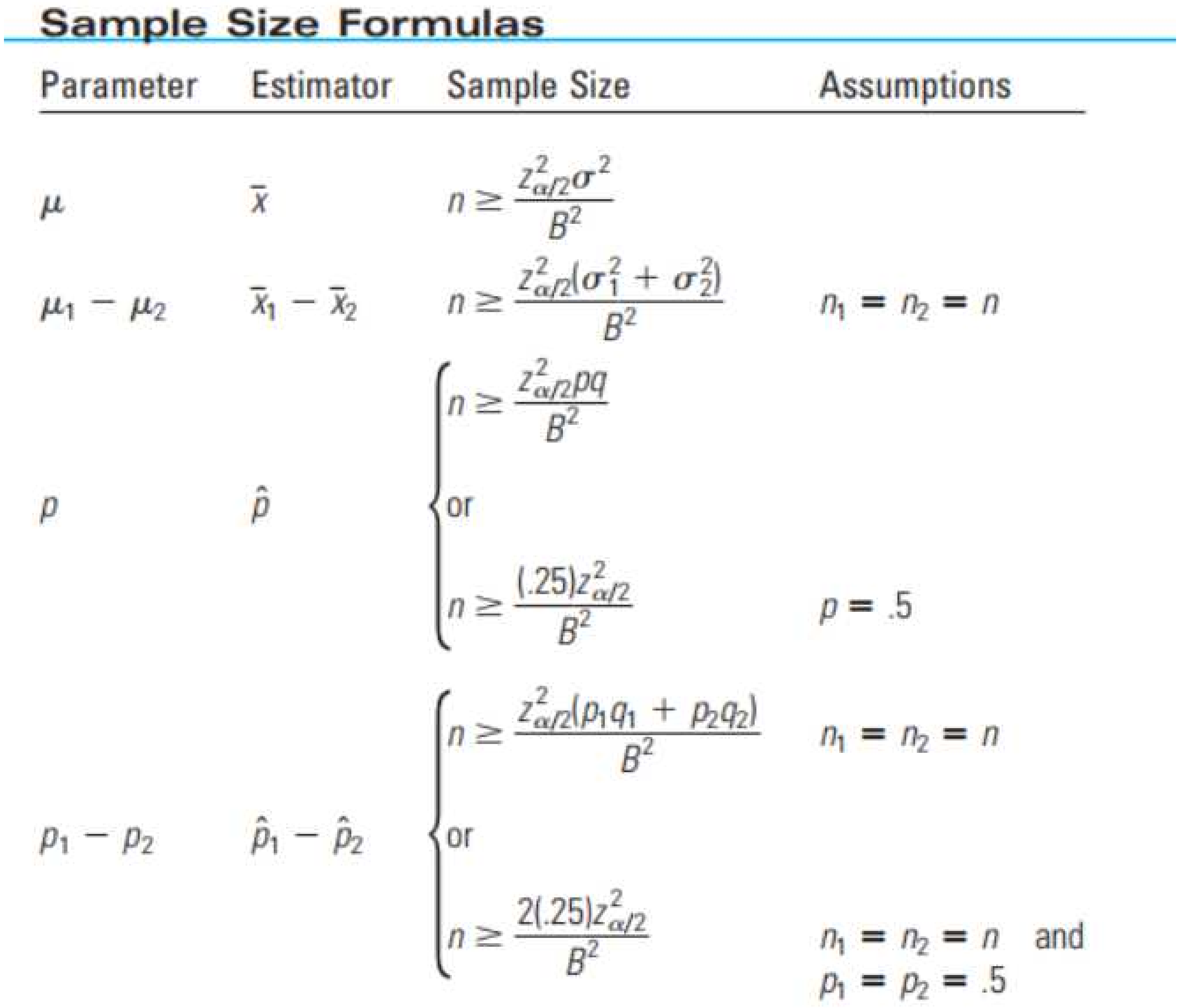
Sol) B=1 minute, SE = $1.96\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \leq 1$, $4\sigma \approx 8$ $\sigma \approx 2$

want to use two equal groups and common variance
i.e. , let $n = n_1 = n_2$, $\sigma_1^2 = \sigma_2^2 = \sigma^2$

$$1.96\sqrt{\frac{2^2}{n}+\frac{2^2}{n}} \leq 1 \Rightarrow 1.96\sqrt{\frac{8}{n}} \leq 1 \Rightarrow \sqrt{n} \geq 1.96\sqrt{8}$$

$$\Rightarrow n \geq 31$$

✓ each group should contain at least n=31 employees.



Sample Size Formulas

Parameter	Estimator	Sample Size	Assumptions
μ	$\bar{x}$	$n \geq \frac{z^2_{\alpha/2}\sigma^2}{B^2}$	
$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$n \geq \frac{z^2_{\alpha/2}(\sigma^2_1 + \sigma^2_2)}{B^2}$	$n_1 = n_2 = n$
p	$\hat{p}$	$n \geq \frac{z^2_{\alpha/2}pq}{B^2}$	
		or	
		$n \geq \frac{(.25)z^2_{\alpha/2}}{B^2}$	$p = .5$
$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$n \geq \frac{z^2_{\alpha/2}(p_1q_1 + p_2q_2)}{B^2}$	$n_1 = n_2 = n$
		or	
		$n \geq \frac{2(.25)z^2_{\alpha/2}}{B^2}$	$n_1 = n_2 = n$ and $p_1 = p_2 = .5$

Key Concepts

I. Types of Estimators

1. **Point estimator**: a single number is calculated to estimate the population parameter.

2. **Interval estimator**: two numbers are calculated to form an interval that contains the parameter.

II. Properties of Good Estimators

1. **Unbiased**: the average value of the estimator equals the parameter to be estimated.

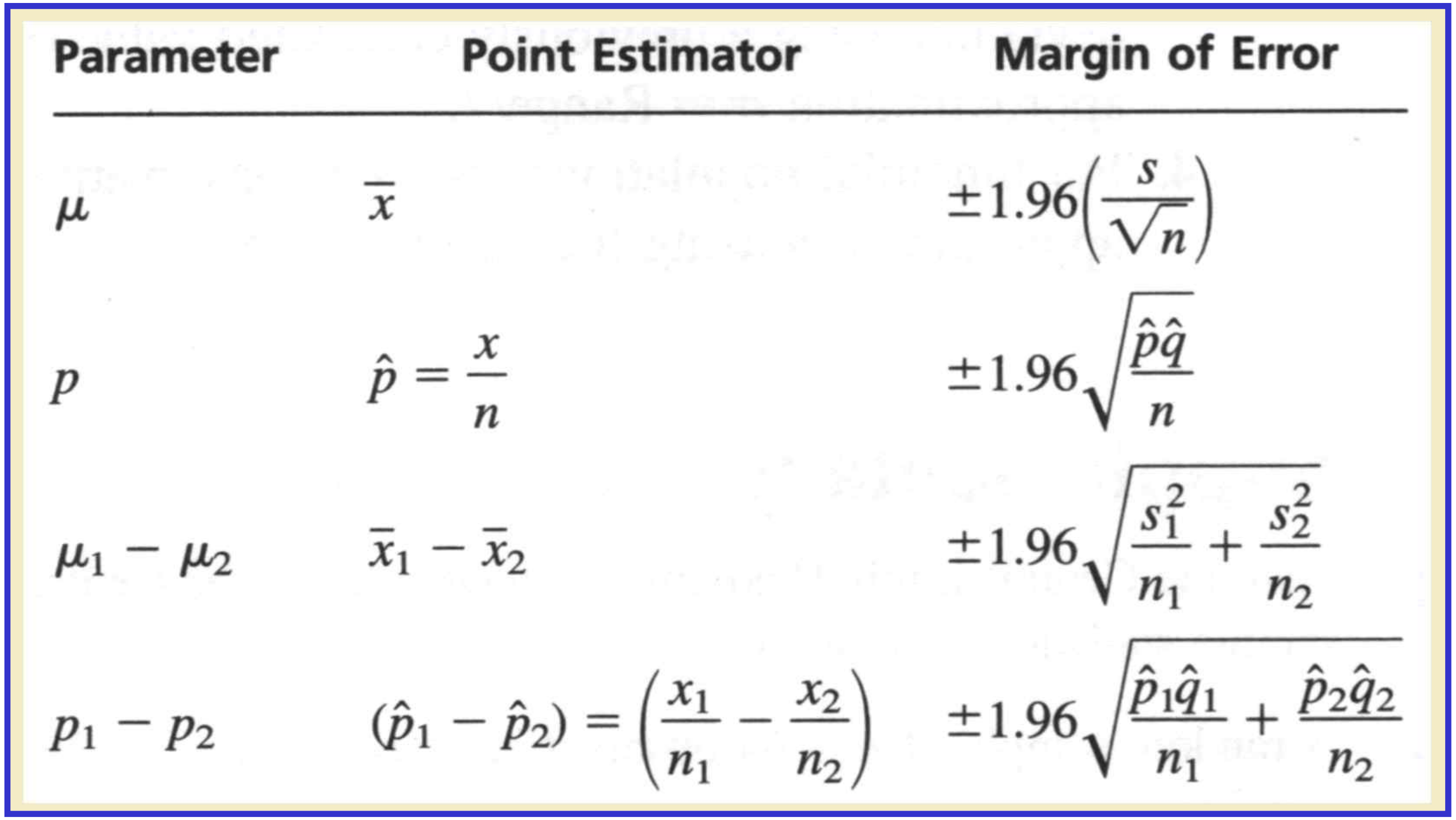
2. **Minimum variance**: of all the unbiased estimators, the best estimator has a sampling distribution with the smallest standard error.

3. The **margin of error** measures the maximum distance between the estimator and the true value of the parameter.

Key Concepts

III. Large-Sample Point Estimators

To estimate one of four population parameters when the sample sizes are large, use the following point estimators with the appropriate margins of error.

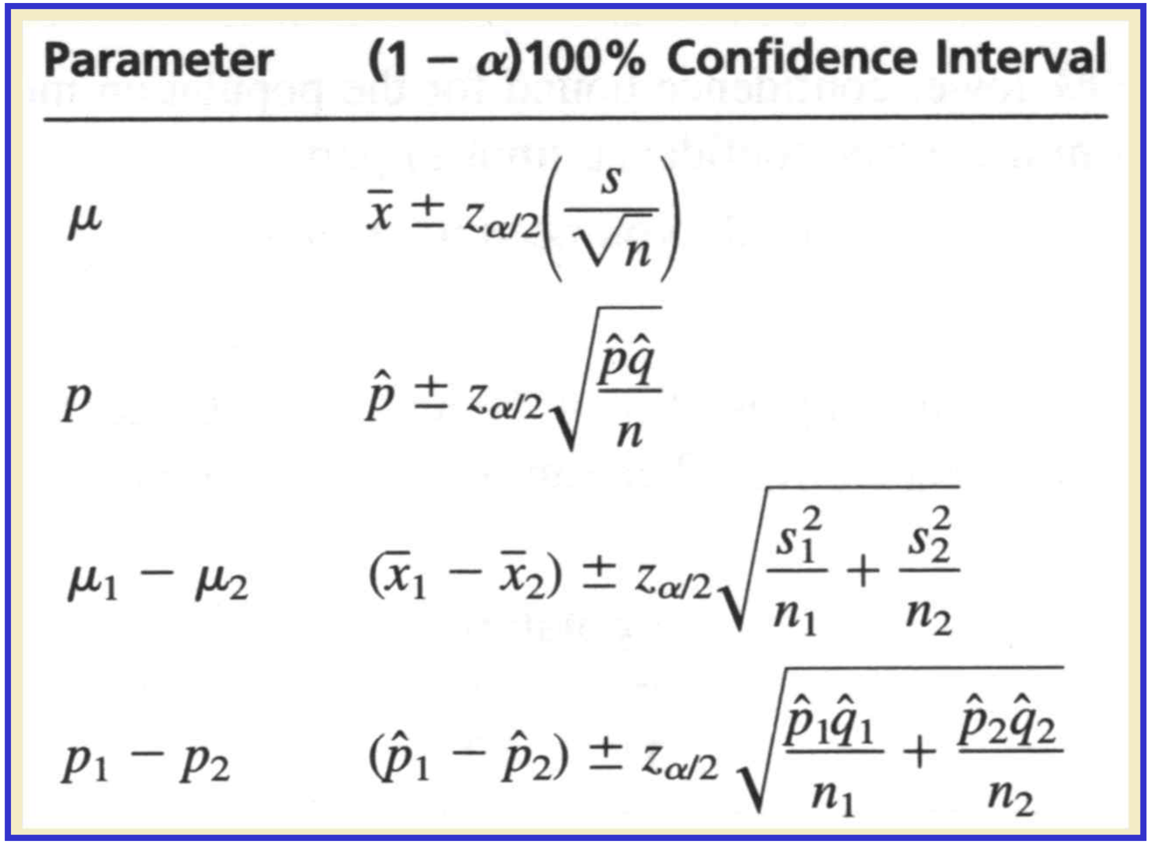


Parameter	Point Estimator	Margin of Error
μ	$\bar{x}$	$\pm 1.96\left(\frac{s}{\sqrt{n}}\right)$
p	$\hat{p} = \frac{x}{n}$	$\pm 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}$
$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$\pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
$p_1 - p_2$	$(\hat{p}_1 - \hat{p}_2) = \left(\frac{x_1}{n_1} - \frac{x_2}{n_2}\right)$	$\pm 1.96\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$

Key Concepts

IV. Large-Sample Interval Estimators

To estimate one of four population parameters when the sample sizes are large, use the following interval estimators.



Parameter	$(1-\alpha)100\%$ Confidence Interval
μ	$\bar{x} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$
p	$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$
$\mu_1 - \mu_2$	$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
$p_1 - p_2$	$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$

Key Concepts

1. All values in the interval are possible values for the unknown population parameter.
2. Any values outside the interval are unlikely to be the value of the unknown parameter.
3. To compare two population means or proportions, look for the value 0 in the confidence interval. If 0 is in the interval, it is possible that the two population means or proportions are equal, and you should not declare a difference. If 0 is not in the interval, it is unlikely that the two means or proportions are equal, and you can confidently declare a difference.

V. One-Sided Confidence Bounds

Use either the upper (+) or lower (–) two-sided bound, with the critical value of z changed from $z_{\alpha/2}$ to z_{α}.