Marmara University, **2021**

Probability and Statistics

Subject 8

Large Sample Estimation

Mujdat Soyturk, Ph.D.
Associate Professor

Contents

Most parts of the slides are derived from the textbook: "Mendenhall, Beaver, Beaver, *Introduction to Probability and Statistics*, 14th Ed., Brooks/Cole, Cengage Learning, 2013"

Introduction

- Populations are described by their probability distributions and parameters.
  - For quantitative populations, the location and shape are described by μ and σ.
  - For a binomial populations, the location and shape are determined by p.
- **If the values of parameters are unknown, we make inferences about them using sample information.**

Types of Inference

- **Estimation:**
  - Estimating or predicting the value of the parameter
  - "What is (are) the most likely values of μ or p?"
- **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 $p = .2$?"

Types of Inference

Examples:

- 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.

- 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_O?

Types of Inference

- A statistical problem, which involves planning, analysis, and inference making, is incomplete without a measure of the goodness of the inference. That is,
  - how accurate or reliable is the method you have used?
- 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

Definitions

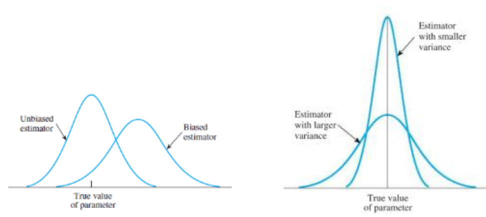
- To estimate the value of population parameter, you can use information from the sample in the form of estimator.
  - Estimators are calculated using information from the sample observations.
- An **estimator** is a rule, usually a formula, that tells you how to calculate the estimate based on the sample.
  - **Point estimation:** A single number is calculated to estimate the parameter.
  - **Interval estimation:** Two numbers are calculated to create an interval within which the parameter is expected to lie.

Properties of Point Estimators

- Since an estimator is calculated from sample values, it varies from sample to sample according to its **sampling distribution**.
- Sampling distributions provide information that can be used to select the **best estimator**.
- An **estimator** of a parameter is **unbiased** if the mean of its distribution is equal to the true value of the parameter.
  - It does not systematically overestimate or underestimate the target parameter.

Properties of Point Estimators

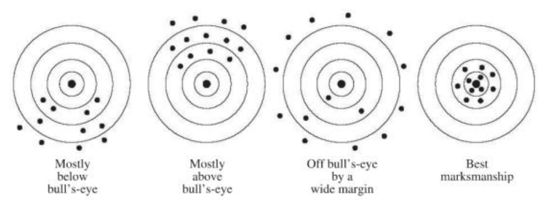
Of all the **unbiased** estimators, we prefer the estimator whose sampling distribution has the **smallest spread** or **variability**.



Properties of Point Estimators

Which one is best?

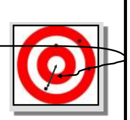
- Unbiasedness? (closeness to the true value)
- The spread? (variance of the distribution)



Measuring the Goodness of an Estimator

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

The distance between the bullet and the bull's-eye.



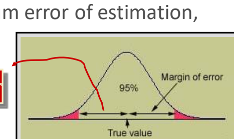
- In this chapter, the sample sizes are large, so that our ***unbiased*** estimators will have **normal** distributions.

Because of the Central Limit Theorem.

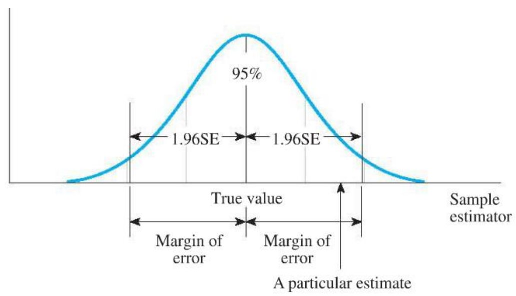
The Margin of Error

- For ***unbiased*** estimators with normal sampling distributions, 95% of all point estimates will lie within 1.96 standard deviations of the parameter of interest (The Empirical Rule).
- **Margin of error:** The maximum error of estimation, calculated as

$1.96 \times$ std error of the estimator



The Margin of Error



Estimating Means and Proportions

- For a quantitative population,

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

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

- For a binomial population,

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

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

Example

A homeowner randomly samples 64 homes similar to her own and finds that the average selling price is $252,000 with a standard deviation of $15,000. Estimate the average selling price for all similar homes in the city.

Point estimator of μ: $\bar{x} = 252{,}000$

Margin of error : $\pm 1.96 \frac{s}{\sqrt{n}} = \pm 1.96 \frac{15{,}000}{\sqrt{64}} = \pm 3675$

Example

A quality control technician wants to estimate the proportion of soda cans that are underfilled. He randomly samples 200 cans of soda and finds 10 underfilled cans.

$n = 200 \quad p = \text{proportion of underfilled cans}$

Point estimator of $p : \hat{p} = x/n = 10/200 = .05$

Margin of error: $\pm 1.96 \sqrt{\frac{\hat{p}\hat{q}}{n}} = \pm 1.96 \sqrt{\frac{(.05)(.95)}{200}} = \pm .03$

Interval Estimation

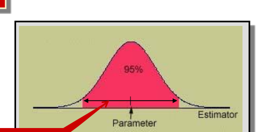
- Interval estimator is a rule for calculating two numbers – say, a and b – to create an interval that you are *fairly certain* contains the parameter of interest.
- The concept of "***fairly certain***" means "***with high probability***".
- We measure this probability using the ***confidence coefficient***, designated by $1 - \alpha$.

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-α.**

Usually, $1-\alpha = .90, .95, .98, .99$

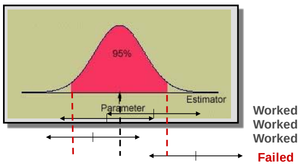
- Suppose 1-α = .95 and that the estimator has a normal distribution.



Parameter ± 1.96SE

Interval Estimation

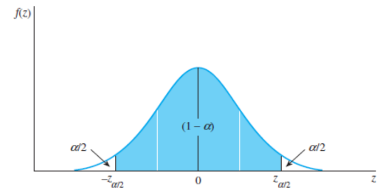
Since we don't know the value of the parameter, consider **Estimator ± 1.96SE** which has a variable center.



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

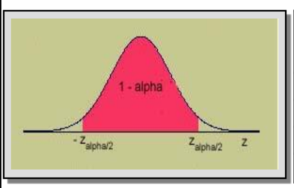
Confidence Coefficient

Since the remaining area in the two tails is α, each tail contains area $\alpha/2$. The value of z that has "tail area" $\alpha/2$ to its right is called $z_{\alpha/2}$, and the area between $-z_{\alpha/2}$ and $z_{\alpha/2}$ is the **confidence coefficient (1- α).**



To 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-α)	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,

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 :

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

Confidence Interval

- A 95% confidence interval tells you that, if you were to construct many of these intervals (all of which would have slightly different endpoints), 95% of them would enclose the population mean.
- You cannot be absolutely sure that any one particular interval contains the mean μ.
- A good confidence interval has two desirable characteristics:
  - It is narrow as possible.
    - More exactly to locate the estimated parameter.
  - It has large confidence coefficient.
    - More likely the interval will contain the estimated parameter.

– To increase the **confidence** without increasing the width of the interval is to increase the **sample size**, ***n***.

Example

A random sample of n = 50 males showed a mean average daily intake of dairy products equal to 756 grams with a standard deviation of 35 grams. Find a 95% confidence interval for the population average μ.

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

$$\text{or } 746.30 < \mu < 765.70 \text{ grams.}$$

Example

Find a 99% confidence interval for μ, the population average daily intake of dairy products for men.

$$\bar{x} \pm 2.58\frac{s}{\sqrt{n}} \Rightarrow 756 \pm 2.58\frac{35}{\sqrt{50}} \Rightarrow 756 \pm 12.77$$
$$\text{or } 743.23 < \mu < 768.77 \text{ grams.}$$

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

Example

Of a random sample of n = 150 college students, 104 of the students said that they had played on a soccer team during their K-12 years. Estimate the proportion of college students who played soccer in their youth with a 98% confidence interval.

$$\hat{p} \pm 2.33\sqrt{\frac{\hat{p}\hat{q}}{n}} \Rightarrow \frac{104}{150} \pm 2.33\sqrt{\frac{.69(.31)}{150}}$$
$$\Rightarrow .69 \pm .09 \qquad \text{or } .60 < p < .78.$$

Estimating the Difference between Two Population 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.
- To make this comparison,

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 Population 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-μ_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 $= \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.

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 : $\pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

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

$$(\bar{x}_1 - \bar{x}_2) \pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$\Rightarrow (756 - 762) \pm 1.96\sqrt{\frac{35^2}{50} + \frac{30^2}{50}} \quad \Rightarrow \quad -6 \pm 12.78$$

$$\text{or } -18.78 < \mu_1 - \mu_2 < 6.78.$$

Example, continued

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

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 Bin. 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.
- To make this comparison,

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 Bin. 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 $\hat{p}_1 - \hat{p}_2$

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 $SE = \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$.

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 : $\pm 1.96\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$

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

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

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

$$\text{or } .06 < p_1 - p_2 < .44.$$

Example, continued

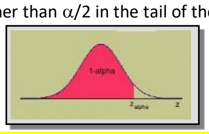
$$.06 < p_1 - p_2 < .44$$

- 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 $\boldsymbol{p_1 - p_2 = 0}$. 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 (of interest, such as μ, p, $\mu_1 - \mu_2$, or $p_1 - p_2$).
- **One-sided bounds** can be constructed simply by using a value of z that puts α rather than $\alpha/2$ in the tail of the z distribution.



Lower Confidence Bound, LCB : Estimator $- z_\alpha \times$ (Std Error of Estimator)
Upper Confidence Bound, UCB : Estimator $+ z_\alpha \times$ (Std Error of Estimator)

Choosing the Sample Size

- The amount (or quality) of information of a sample varies depending on how and where the information is collected.
- The total amount of relevant information in a sample is controlled by two factors:
  - The <u>sampling plan</u> or <u>experimental design</u>: the procedure for collecting the information
  - The <u>**sample size *n***</u>: the amount of information you collect.
- You can increase the amount of information you collect by increasing the sample size, or changing the type of sampling plan or experimental design.
- For a sampling plan (e.g. random sampling), **what should be the sample size**? How many measurements should be included in the sample?
- To answer these, you have to specify these firstly:
  - The reliability you wish to achieve,
  - The accuracy needed for your estimate.

Choosing the Sample Size

- In a statistical estimation problem, the accuracy of the estimation is measured by the **margin of error** or the **width of the confidence interval.**
- Since both of these measures are a function of <u>**the sample size**</u>, specifying the *reliability* and *accuracy* allows you determine the necessary sample size.
- For instance, suppose you want to estimate the daily average of a process and you need the margin of error less than amount ***X***. This means that:
  - Approximately, 95% of the time in repeated sampling, the "<u>**reliability**</u>", the distance between the sample mean $\bar{x}$ and the population mean $\boldsymbol{\mu}$ will be less than 1.96 SE.
  - You want this quantity to be less than amount ***X*** (the "<u>**accuracy**</u>"), $1.96\ SE\ < \boldsymbol{X}$.
- So what should be the sample size, n?

Choosing the Sample Size

$$1.96\ SE\ < \boldsymbol{X}$$

$$\rightarrow \quad 1.96\left(\frac{\sigma}{\sqrt{n}}\right) < \boldsymbol{X}$$

$$\rightarrow \quad n > \left(\frac{1.96}{X}\right)^2 \sigma^2$$

- If you know the population standard deviation, σ, you can solve the above formula for ***n***.
- If you don't know, use the best approximation;
  - An estimate s obtained from a previous sample.
  - A range estimate based on knowledge of the largest and smallest possible measurements, $\sigma \approx Range/4$

Choosing the Sample Size

- Sometimes, researchers request a different reliability or confidence level other than the 95% confidence specified by the margin of error.
- In this case, half-width of the confidence interval provides the accuracy measure for your estimate, that is the bound ***B*** on the error of your estimate is;

$$z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < B$$

Choosing the Sample Size

Determine **the parameter to be estimated** and the **standard error** of its point estimator. Then proceed as follows:

1. Choose ***B***, the bound on the error of your estimate (size of the **margin of error**), and a confidence coefficient $(1-\alpha)$.
2. For a one-sample problem, solve the equation for the sample size n:
$$z_{\alpha/2} \times (\text{Standard error of the estimator}) \leq B$$
where $z_{\alpha/2}$ is the value of ***z*** having area $\boldsymbol{\alpha/2}$ to its right.
3. For two-sample problem, set $n_1=n_2=n$ and solve the equation in step 2.
4. For quantitative populations, you can estimate the population standard deviation using a previously calculated value of ***s*** or the range approximation $\sigma \approx$ **Range / 4.**
5. 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?

$$1.96\sqrt{\frac{pq}{n}} \leq .04 \qquad \Rightarrow 1.96\sqrt{\frac{.5(.5)}{n}} \leq .04$$

$$\Rightarrow \sqrt{n} \geq \frac{1.96\sqrt{.5(.5)}}{.04} = 24.5 \qquad \Rightarrow n \geq 24.5^2 = 600.25$$

He should survey at least 601 wholesalers.

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 $z_{a/2}$ to z_a.