Introduction to Probability and Statistics Twelfth Edition

Robert J. Beaver • Barbara M. Beaver • William Mendenhall

Presentation designed and written by:
Barbara M. Beaver

Edited by: Dr. Worapan Kusakunniran and Dr. Rob Egrot

Introduction to Probability and Statistics Twelfth Edition

Chapter 8
Large-Sample Estimation

Introduction

- Populations are described by their **probability distributions** and **parameters**.

  (parameter = a numerical descriptive measure that characterizes a population)

  – For normal populations, the location and shape are described by μ and σ.

  – For binomial populations, the location and shape are determined by ***n*** and ***p***.

- If the values of parameters are unknown, we make inferences about them using **sample** information.

Statistical Inference

What is a statistical inference?

- It is a process of
  - Drawing a conclusion
  - Making a prediction
  - Making a decision

about an unknown parameter of a population from information contained in sample.

Statistical Inference

Methods for making statistical inferences are

1. Estimation

2. Hypothesis testing

The main idea is that we use statistics calculated from samples to estimate or test hypotheses about the values of parameters of populations.

Methods of Inference

- **Estimation:**
  - **Estimating** or predicting the value of the parameter

  Example: "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.

  Example: "Did the sample come from a population with $\mu < 5$?"

Methods 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 higher than the old average resistance, μ_O?

Methods 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**

Types of Estimators

- An **estimator** is a **rule**, usually a **formula**, that tells you how to calculate the estimate based on the sample.
- Estimators are used in two different ways
  - **Point estimation:** A single number is calculated to estimate the parameter.
    - **E.g. use sample mean to estimate population mean.**
  - **Interval estimation:** Two numbers are calculated to create an interval within which the parameter is expected to lie.

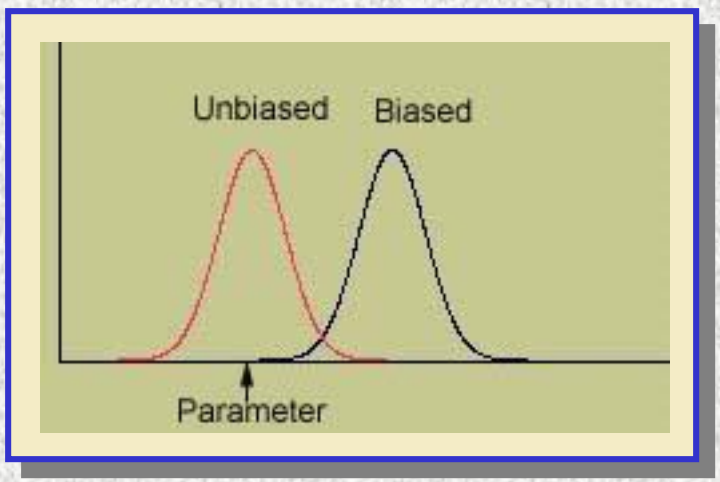
Point Estimators

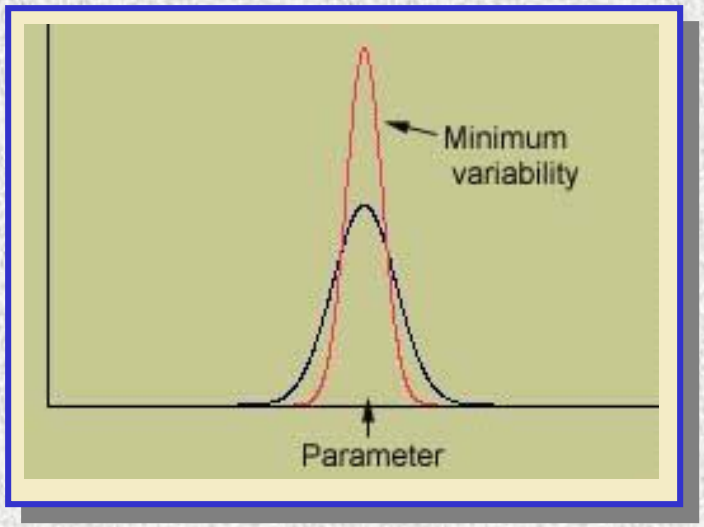
- Since an estimator is calculated from sample values, it varies from sample to sample according to its **sampling distribution.**

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

Point Estimators

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





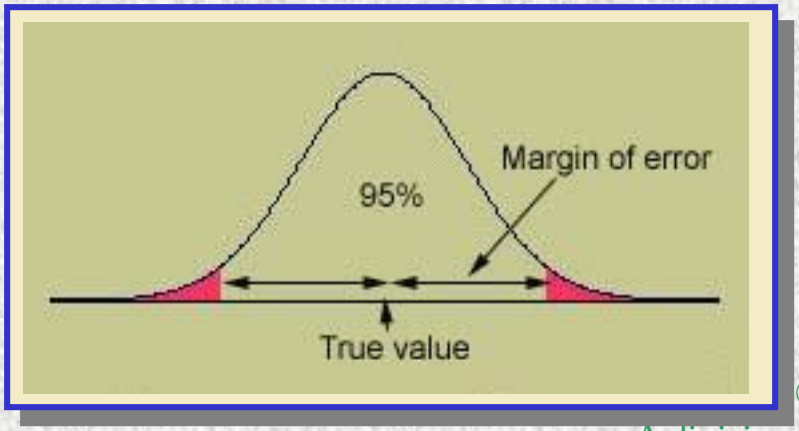
Measuring the Goodness of an Estimator

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

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.
- **Margin of error (for 95% confidence):**

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



Calculating the Margin of Error

- Most of the time we don't know the standard deviation of the population (σ) so for now we use the standard deviation of the sample (s) instead.
- This is a bit of a cheat because they may not be equal. Later we will see how to adjust the formula for better accuracy.

Estimating Means and Proportions

- For a normal population,

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

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

- For a binomial population,

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

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

This is an estimate of the standard deviation of the distribution of $\hat{p}$.

Example

- A homeowner randomly samples 64 homes similar to her own and finds that the average selling price is $252,000. The standard deviation of the sample is $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 : $\pm 1.96 \dfrac{s}{\sqrt{n}} = \pm 1.96 \dfrac{15,000}{\sqrt{64}} = \pm 3675$

Example

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

$n = 200$ $p =$ **proportion of under-filled cans**

Point estimator of p: $\hat{p} = \frac{x}{n} = \frac{10}{200} = 0.05$

Margin of error: $\pm 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}} = \pm 1.96\sqrt{\frac{(0.05)(0.95)}{200}} = \pm 0.03$

Class Activity/ Homework 10

1. Calculate the margin of error (95% confidence) in estimating a population mean μ for these values:
   a) $n = 36, \sigma = 0.2$
   b) $n = 36, \sigma = 0.9$
   c) $n = 36, \sigma = 1.5$
   d) What do these results tell us about the relationship between standard deviation and the margin of error?

Class Activity/ Homework 10

2. Calculate the margin of error in estimating a population mean μ for these values:
   a) $n = 36, \sigma^2 = 4$
   b) $n = 3600, \sigma^2 = 4$
   c) $n = 360000, \sigma^2 = 4$
   d) What do these results tell us about the relationship between sample size and the margin of error?

Class Activity/ Homework 10

3. A random sample of n = 100 observations from a population produced $\bar{x}$ = 29.7 and s^2 = 16. Give the best point estimate for the population mean μ and calculate the margin of error (95% confidence).

Class Activity/ Homework 10

4. A random sample of n = 900 observations from a binomial population produced x = 655 successes. Estimate the binomial proportion *p* and calculate the margin of error.

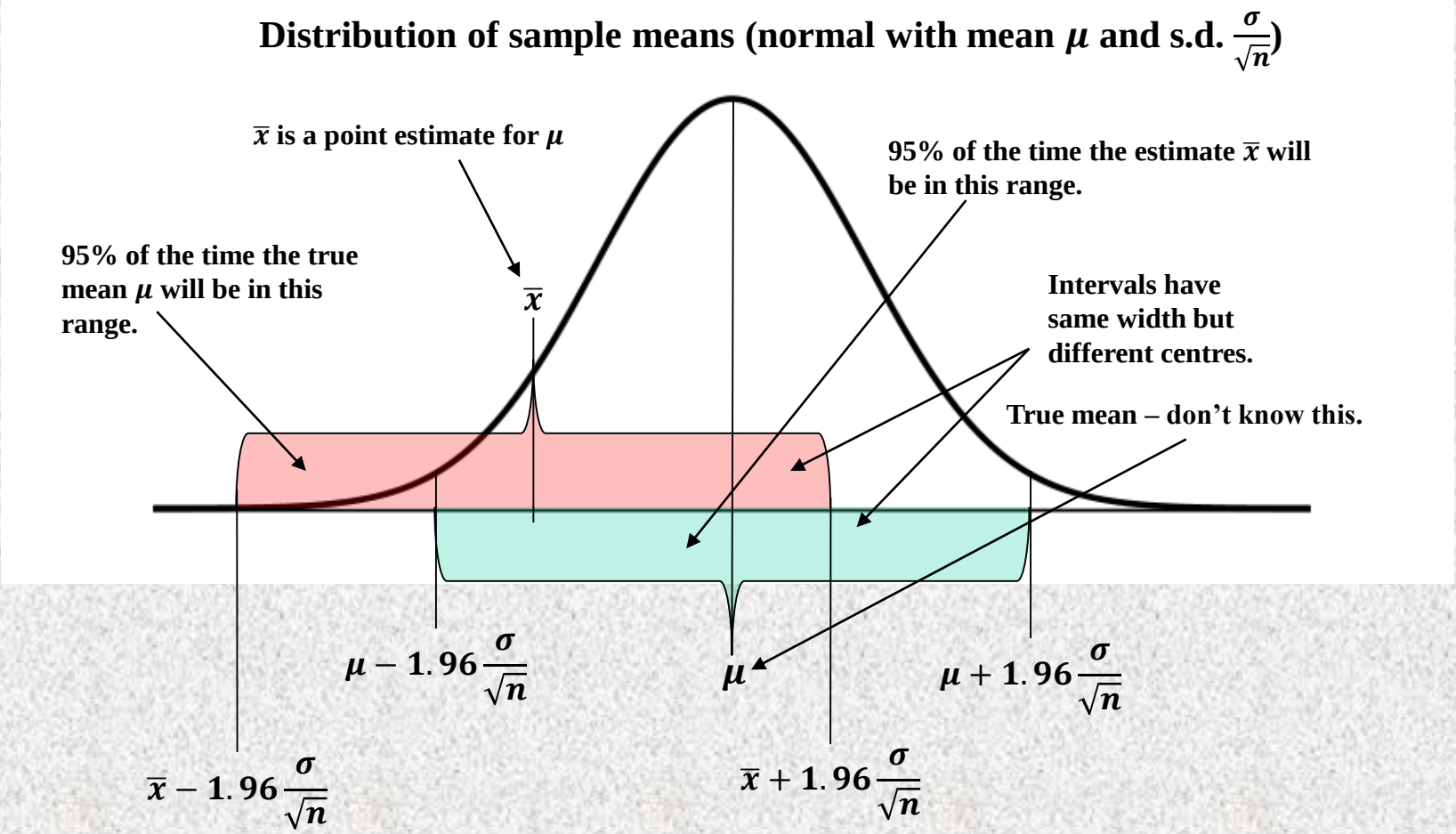
Interval Estimation

- 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-α = .90, .95, .98, .99

- Suppose 1-α = .95 and that the point estimator has a normal distribution.
- For example the sample mean for estimating the population mean after applying the CLT.

95% confidence interval for mean estimation



Confidence Interval for a Population Mean μ

Assume that the sample size n is large ($n \geq 30$).

A $(1-\alpha)100\%$ Confidence Interval for μ :

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

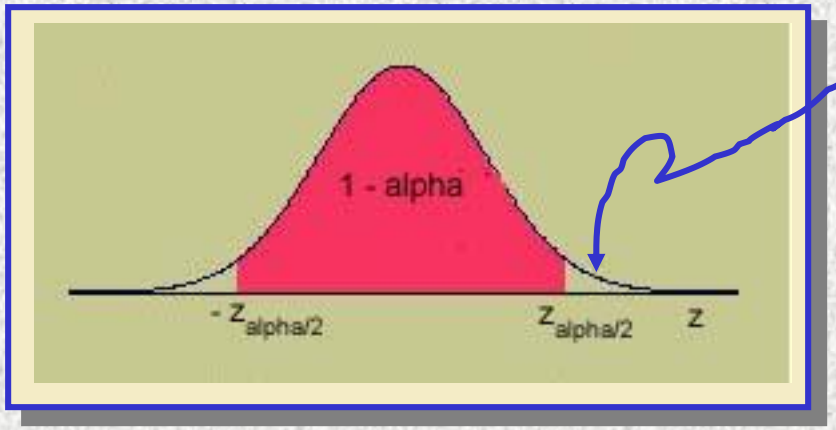
The Z value that has a tail of $\frac{\alpha}{2}\%$

The interval is based on the following probability:

$$P\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right) = 1 - \alpha$$

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.



Tail area	$z_{\alpha/2}$	100(1-α)%
.05	1.645	90%
.025	1.96	95%
.01	2.33	98%
.005	2.58	99%

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

Note about σ

- Note that we need to know the standard deviation of the original distribution to calculate confidence intervals with this method.
- This is unrealistic, as if we know the standard deviation we probably know the mean, which is what we're supposed to be estimating!
- Later we will see a way to find confidence intervals when we don't know σ.
- Until then we will use the sample standard deviation in place of σ, but this is a cheat and we shouldn't rely on it.

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.

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

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

Thus, μ is estimated to be between 746.3 and 765.7 grams.

Example

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

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

$$or \quad 743.23 < \mu < 768.77 \quad \text{grams}$$

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

Error of Estimation

- Given $\varepsilon > 0$, we want to choose a sample size so that the (1-α)100% confidence interval for μ is no bigger than $\bar{x} \pm \varepsilon$.

$$Z_{\alpha/2}\frac{s}{\sqrt{n}} \leq \varepsilon \quad \Leftrightarrow \quad n \geq \frac{(Z_{\alpha/2})^2 s^2}{\varepsilon^2}$$

- ε is called the maximum error of the estimate of μ or **margin of error**.
- In other words, we want to find the minimum possible sample size so that the m.o.e. is at most ε.
- We will need to take a preliminary estimate of s.

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. Estimate a 95% margin of error in estimating μ.

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

Example (Continued)

What sample size would be needed for the maximum error of the estimate (margin of error) of μ to be $\varepsilon = 8$ with 95% confidence?

$$n \geq \frac{z_{\alpha/2}^2}{\varepsilon^2} s^2 = \frac{(1.96)^2(35)^2}{8^2} = 73.53$$

So $n \geq 74$.

Class Activity/ Homework 10

5. A homeowner randomly samples 64 homes similar to her own and finds that the average selling price is $250,000 with a standard deviation of $15,000. Find

   a) A point estimate of the selling price for all similar homes in the city.

   b) A 95% margin of error.

   c) Find the sample size n that would be needed for the 95% CI for μ to have length of at most 7000 (assuming same $\bar{x}$ and s).

Estimating Population Proportion, *p*

- Recall that p is the probability of success in a binomial distribution.
- $\hat{p}$ is the sample proportion.
- $\hat{p}$ is an unbiased point estimator for p.

$$\hat{p} = \frac{x}{n} = \frac{\text{the number of sucesses}}{\text{the sample size, } n}$$

Confidence Interval for a Population Proportion *p*

Assume that the sample size n is large.

It is recommended that $n\hat{p} > 5$ and $n\hat{q} > 5$.

A (1-α)100% Confidence Interval for *p* :

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

Estimating Population Proportion, *p*

- The point estimator of p is $\hat{p}$
- The standard error of $\hat{p}$ is estimated as

$$\text{SE} = \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

- The (1-α)100% margin of error when n is large is estimated as

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

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.

$$\hat{p} \pm 2.33\sqrt{\frac{\hat{p}\hat{q}}{n}} = \frac{104}{150} \pm 2.33\sqrt{\frac{0.69(0.31)}{150}}$$

or $0.60 < p < 0.78$

Estimating the Difference between Two Population Means

- Sometimes we are interested in comparing the means of two independent 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 difference between sample means $\bar{x}_1 - \bar{x}_2$ is a random variable, and it follows a distribution with mean $\mu_1 - \mu_2$ and variance $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$.**

Variance sum law says
$var(X \pm Y) = var(X) + var(Y)$
for independent X and Y.

Combinations of normal random variables

- FACT: If X and Y are independent normal random variables, then a linear combination of X and Y is also normal.
- In particular, the random variable $\bar{x}_1 - \bar{x}_2$ will be normally distributed if $\bar{x}_1$ and $\bar{x}_2$ are.
- So, if the sample sizes are big enough for the CLT to apply to both $\bar{x}_1$ and $\bar{x}_2$, then we can assume $\bar{x}_1 - \bar{x}_2$ is normally distributed.

Estimating μ_1-μ_2

So, for large samples ($n_1 \geq 30$ and $n_2 \geq 30$), we can use $\bar{x}_1 - \bar{x}_2$ as an unbiased point estimator for $\mu_1 - \mu_2$ and we can assume we know what distribution it follows.

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

Margin of Error : $\pm Z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

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

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

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

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

$$= -6 \pm 12.78$$

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

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 $\mu_1 - \mu_2 = 0$. 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.
- 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 Proportions

- We compare the two proportions by making inferences about p_1-p_2, the difference in the two population proportions.
  - The best estimate of p_1-p_2 is the difference in the two sample proportions.

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

  - This is an unbiased point estimator.

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 : $\hat{p}_1 - \hat{p}_2$

Margin of Error : $\pm Z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$

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

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

- 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.

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

$$= \frac{65}{80} - \frac{39}{70} \pm 2.58\sqrt{\frac{0.81(0.19)}{80} + \frac{0.56(0.44)}{70}} \Rightarrow .25 \pm .19$$

$$= 0.25 \pm 0.19$$

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

$$\text{or} \quad 0.06 < p_1 - p_2 < 0.44$$

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 $p_1 - p_2 = 0$. Therefore, it is not likely that $p_1 = p_2$. 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.

Class Activity/ Homework 10

6. Independent random samples were selected from populations 1 and 2. The sample's sizes, means, and variances are as follows:

Population	1	2
Sample size	35	49
Sample mean	12.7	7.4
Sample variance	1.38	4.14

a) Find a 95% confidence interval for estimating the difference in the population means (μ_1-μ_2).

b) Based on the confidence interval in part a, can you conclude that there is a difference in the means for the two populations? Explain.

Class Activity/ Homework 10

7. Independent random samples of $n_1 = 800$ and $n_2 = 640$ observations were selected from binomial populations 1 and 2, and $x_1 = 337$ and $x_2 = 374$ successes were observed. Find a 90% confidence interval for the difference $(p_1 - p_2)$ in the two population proportions. Interpret the interval.

Class Activity/ Homework 10

8. **M&M'S.** Does Mars, Incorporated use the same proportion of red candies in its plain and peanut varieties? A random sample of 56 plain M&M'S contained 12 red candies, and another random sample of 32 peanut M&M'S contained 8 red candies.

   a) Construct a 95% confidence interval for the difference in the proportions of red candies for the plain and peanut varieties.

   b) Based on the confidence interval in part a, can you conclude that there is a difference in the proportions of red candies for the plain and peanut varieties? Explain.

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.

For 95% confidence

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.

Video Links

- The sampling distribution of sample means https://www.youtube.com/watch?v=q50GpTdFYyI
- Introduction to confidence intervals https://www.youtube.com/watch?v=27iSnzss2wM
- Intro to confidence intervals for one mean (sigma known) https://www.youtube.com/watch?v=KG921rfbTDw
- The sampling distribution of the difference in sample means https://www.youtube.com/watch?v=4HB-FL529ag