Introduction to Probability and Statistics Eleventh 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 quantitative population, the location and shape are described by μ and σ.

  – For a binomial population, the location and shape are determined by $\boldsymbol{p}$.

Introduction

- Chapter 8 covers 4 parameters.
  - The population mean, μ
  - The population proportion, p
  - A difference between 2 population means, $\mu_1 - \mu_2$
  - A difference between 2 population proportions, $p_1 - p_2$
- If the values of parameters are *unknown*, we make inferences about them using sample information.

Statistical Inference

What is a statistical inference?

A process (the theory, methods, and practice) of forming judgments about the parameters of a population, usually on the basis of random sampling.

Statistical Inference

Methods for making statistical inferences are

1. Estimation
2. Hypothesis testing

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 greater 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 Estimation

- 2 Types of Estimation
  - **Point estimation:** A single number is calculated to estimate the population parameter.
  - **Interval estimation:** Two numbers a and b are calculated to create an interval (a, b) within which the parameter is expected to lie.

Point Estimation

- A formula that describes a calculation of an unknown parameter is called a **point estimator**, and the resulting number is called a **point estimate.**

  For example, μ is unknown and perhaps one may use the sample mean $\bar{x} = \sum_{i=1}^{n} x_i / n$ as a point estimator.

  If a sample yields an average of 3.4, then 3.4 is called a point estimate of μ.

Point Estimators

– The sample mean $\bar{x}$ is a point estimator of μ

– The sample variance s^2 is a point estimator of population variance σ^2

– The sample proportion $\hat{p}$ is a point estimator of population proportion p.

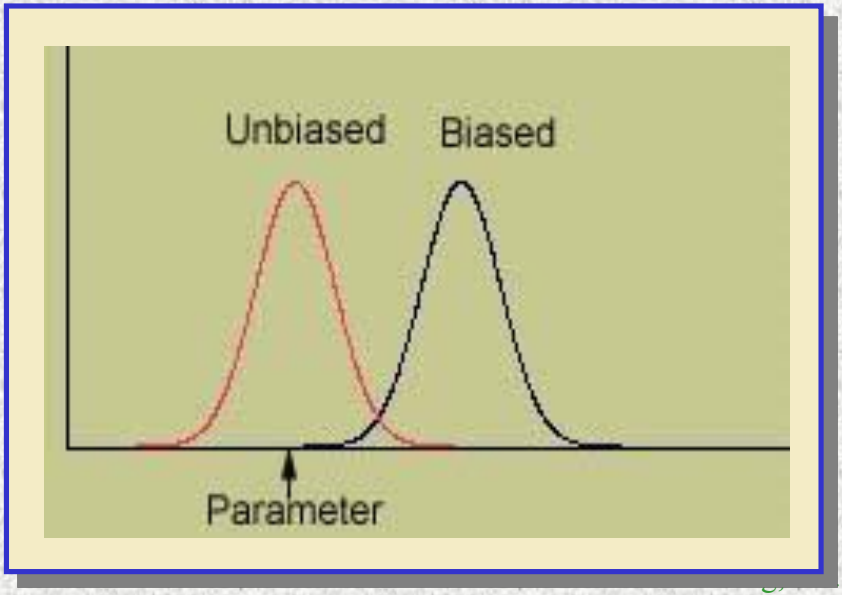
  – *(Also, p is the probability of success from the binomial distribution.)*

Point Estimators

- Two properties of a good point estimator
  - Unbiased
  - Minimum variance (**smallest spread** or **variability**.

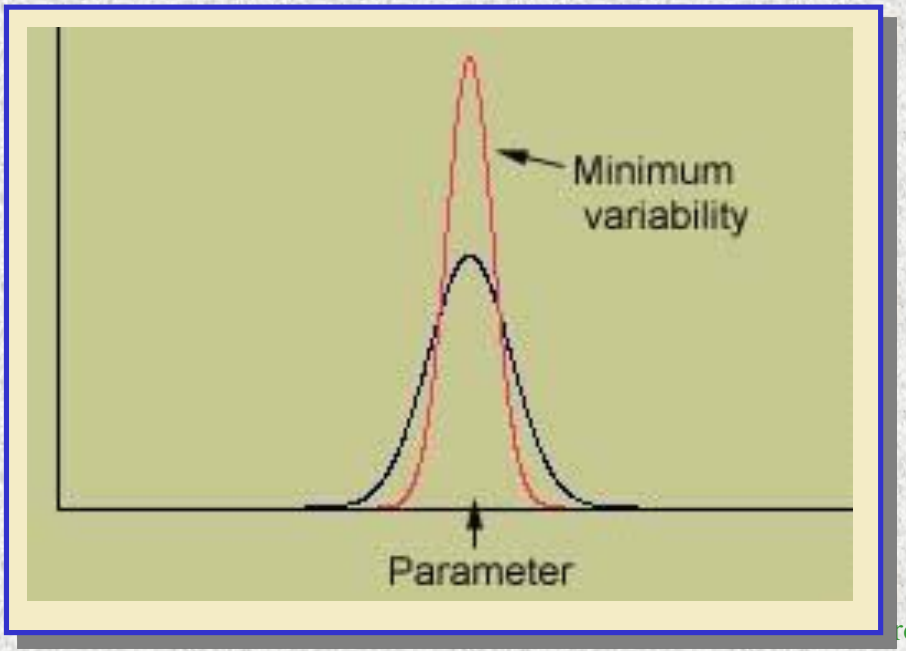
Point Estimators

- A point **estimator** is **unbiased** if the mean of its sampling distribution equals the parameter of interest.
  - **Unbiased**: The point estimator 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**.



Point Estimators

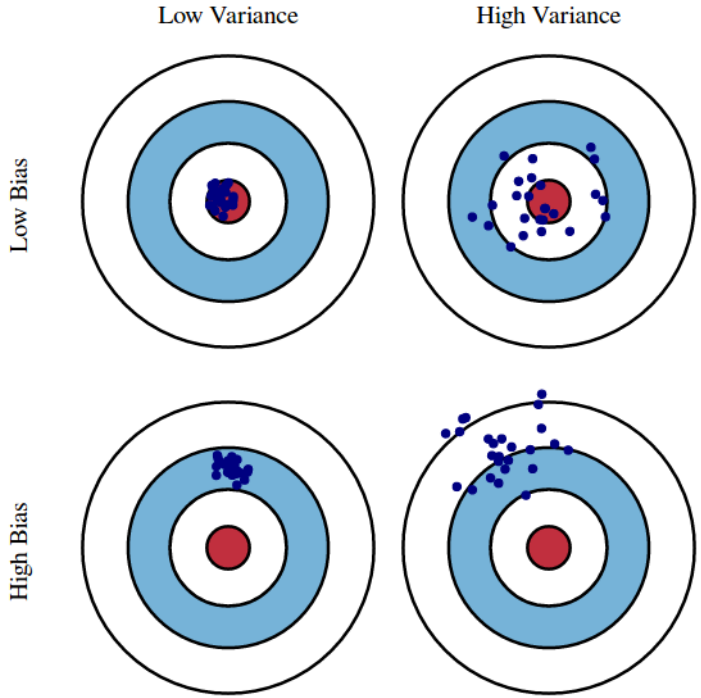


Fig. 1 Graphical illustration of bias and variance.

Interval Estimation

- **Interval estimation:** Two numbers a and b are calculated to create an interval (a, b) that contains the parameter with high probability denoted by $1-\alpha$.
  - The formula that describes this calculation is called the **interval estimator**, and the resulting pair of numbers is called an **interval estimate** or **confidence interval.**
  - $1-\alpha$ is called **confidence coefficient**
  - Often , $1-\alpha = 0.90, 0.95, 0.98$ and 0.99

Confidence Interval for a Population Mean μ

Recall Chapter 7,

If n is large, the sampling distribution of the sample mean is **approximately normally distributed**.

Then, $z = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ has a standard normal distribution.

Confidence Interval for μ

Let $z_{\alpha/2}$ be a value of the standard normal random variable such that

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

Then,

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

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

$\bar{X} \pm 2\alpha/2 \frac{6}{\sqrt{n}}$

Confidence Interval for a Population Mean μ

Assume that the sample size n is large ($n >= 30$).

A (1-α)100% Confidence Interval for μ is approximated to be:

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

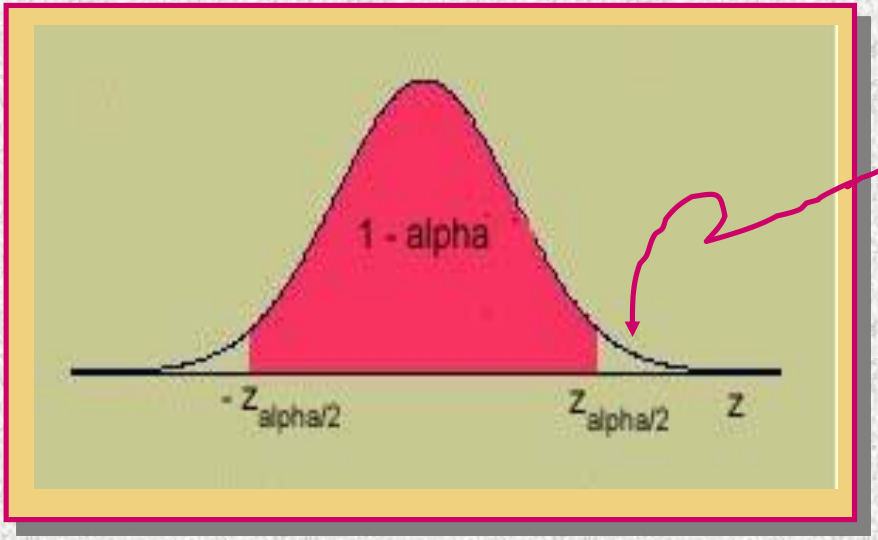
The interval is based on the following probability result:

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

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

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.



1-α	Tail area	$z_{\alpha/2}$
.90	0.05	1.645
.95	0.025	1.96
.98	0.01	2.33
.99	0.005	2.58

Example 1

- 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.

$0.05 \rightarrow \alpha/2 = 0.025$

$z_{\alpha/2} = 1.96$

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

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

Thus, μ is estimated to be between 746.3 and 765.5 grams.

Example 2 (Based on Ex 1)

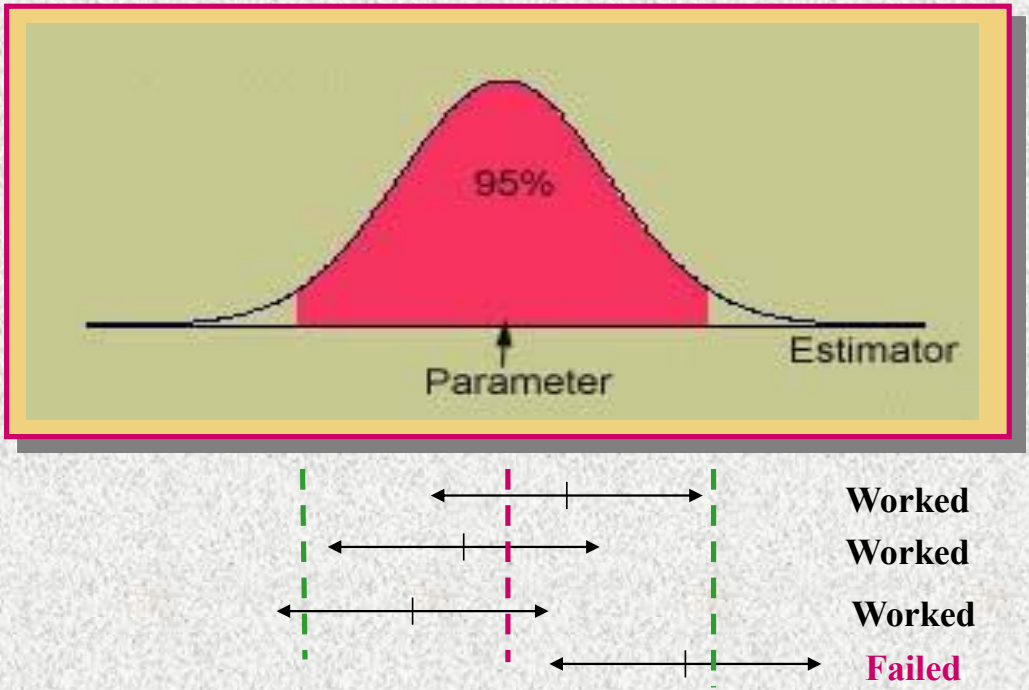
- Find a 99% confidence interval for μ, the population average daily intake of dairy products for men.

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

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

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

Interval Estimation



- If many confidence intervals are constructed with 1-α, about (1-α)100% of intervals will cover the unknown parameter.

Interval Estimation

A 95% CI for μ : $\bar{x} \pm 1.96 \frac{s}{\sqrt{n}}$

- For example, in repeated sampling if 200 confidence intervals for μ are constructed and all with a 95%, then there will be about 190 intervals that contain the unknown value of μ.

Margin of Error

- The distance between an ***estimate*** and the ***true*** value of the parameter is the **error of estimation.**
- The **margin of error, *E*,** is the maximum distance (error) between the estimator and the true value of the parameter.

Margin of Error

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

Rewrite $\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ as

$$|\bar{x} - \mu| < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Definition

Margin of Error

is the maximum likely difference observed between sample mean $\bar{x}$ and true population mean μ.

denoted by E

$\bar{x} - E < \mu < \bar{x} + E$

lower limit upper limit

Estimating Population Mean, μ

- A point estimator of μ is $\bar{x}$.
- The standard error of $\bar{x}$ is $\frac{\sigma}{\sqrt{n}}$ but is estimated as

$$\text{SE} \approx \frac{s}{\sqrt{n}}$$

- A (1-α)100% margin of error when $n \geq 30$ is estimated as

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

Example 1 (continued)

- Based on the statistical results of the sample size $n = 50$ males, the sample average daily intake = 756 grams, and sample standard deviation = 35 grams. For the future study if the margin of error in estimating the population average μ using a 95% CI must be 7, then what is the required sample size?

Let E be a margin of error, then $1.96\frac{s}{\sqrt{n}} = E$

and $n = \left(\frac{1.96s}{E}\right)^2$. If $E = 7$, then $n = 96.04$.

The least sample size needed is 97.

Class Activity

1. 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) Find a 95% CI. for the average selling price.

a) $\bar{x} = 250{,}000$

b) $250{,}000 \pm 1.96\left(\frac{15{,}000}{\sqrt{64}}\right)$

Class Activity

1. (continue)

   c) Find the sample size n that is needed so that a 95% CI. for μ has a length of 6000.

(That is given the margin of error to be 6000/2 = 3000 if you want to make a new study of constructing a new 95% CI using the data from the previous slide, then what is n?)

$E = 3000$

$3000 = 1.96\left(\frac{15000}{\sqrt{n}}\right)$ ✓

Class Activity

2. Calculate the 90% margin of error in estimating a population mean μ for these values:
   a) $n = 36, s^2 = 4$
   b) $n = 3600, s^2 = 4$
   c) Interpret results from a) and b).

$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Estimating Population Proportion, *p*

- Recall that *p* is the probability of success in the binomial distribution.
- Here, *p* is called the ***population proportion***.
- The point estimator of *p* is the ***sample proportion*** denoted as

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

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 :

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

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

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

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

Example 3

- Of a random sample of $n = 150$ college students, 104 of the students are obese. Estimate the proportion of college students who are obese with a 98% confidence interval.

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

$$\frac{104}{150} \pm 2.33\sqrt{\frac{\frac{104}{150}\left(\frac{46}{150}\right)}{150}}$$

$$0.69 \pm 0.09$$

$$\text{or} \quad 0.60 < p < 0.78$$

Class Activity

3. Some people claimed there are health benefits to eating less meat. Based on a random sample of 500 people taken by a health club committee revealed that 70 were vegetarians.

   a) Find a point estimate of p, the true proportion of all vegetarian eaters in this particular city.

   b) Construct a 92% confidence interval for p.

a) $70/500$

b) $70/500 \pm z_{\alpha/2} \sqrt{\frac{\frac{70}{500}\left(\frac{430}{500}\right)}{500}}$

Class Activity

3. (Continue).

   c) Based on your result of Part b, is it reasonable to conclude that the proportion of vegetarian eaters in this city is 0.1?

   d) Suppose that 1000 samples of the same size of 500 people are taken, and a 90% CI is constructed for each, then on the average how many intervals do you expect to contain p?

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 is drawn from population 1 with μ_1 and variance σ_1^2.

A random sample of size n_2 drawn from population 2 with 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-$\mu_2 = 0$.
- The best point estimator of μ_1-μ_2 is the difference in the two sample means,

$$\bar{x}_1 - \bar{x}_2$$

Estimating μ_1-μ_2

For large samples ($n_1 \geq 30$ and $n_2 \geq 30$), point estimator and the margin of error as well as confidence interval are based on the standard normal (z) distribution.

Point estimator for μ_1- μ_2 : $\bar{x}_1 - \bar{x}_2$

100(1-α)% Margin of Error: $\pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$

A (1-α)100% Confidence Interval for μ_1- μ_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 4

Average Daily Intakes	Men (1)	Women (2)
Sample size	50	50
Sample mean	756	762
Sample standard deviation	35	30

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

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

$$= -6 \pm 12.78$$

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

$(756 - 762) \pm z_{\alpha/2}\sqrt{\frac{35^2}{50} + \frac{30^2}{50}}$

Example 4 (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.

Class Activity

4. A study was conducted to compare the mean numbers of police emergency calls per 8-hour shift in two districts of a large city. Samples of 100 8-hour shifts were randomly selected from the police records for each of the two regions, and the number of emergency calls was recorded for each shift. The sample statistics are listed here:

Region	1	2
Sample Size	100	100
Sample Mean	2.4	3.1
Sample Variance	1.44	2.64

b) $\mu_1 - \mu_2 = 2.4 - 3.1$
$= -0.7$

a) $\mu_1 - \mu_2 \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

a) Find a 90% confidence interval for the difference in the mean numbers of police emergency calls per shift between the two districts of the city. Interpret the interval.

b) Find a point estimate for the difference as defined in part (a).

Answer a) -0.7 ± 0.3323 or 0.7 ± 0.3323

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 is drawn from binomial population 1 with parameter p_1.

A random sample of size n_2 is 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 point estimator of p_1-p_2 is the difference in the two sample proportions,

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

Estimating p_1-p_2

For large samples, point estimator and the margin of error as well as confidence interval are based on the standard normal (z) distribution.

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

$$(1-\alpha)100\% \text{ Margin of Error} : \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

A (1-α)100% Confidence Interval for p_1- p_2 is

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

Example 5

College Students	Male (1)	Female (2)
Sample size	80	70
Obese	65	39

- Construct a 99% confidence interval for the difference between the proportions of obesity in male versus female college students.

$\hat{p}_1 = \frac{65}{80}$, $\hat{p}_2 = 39/70$

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\frac{65}{80}\left(\frac{15}{80}\right)}{80} + \frac{\frac{39}{70}\left(\frac{31}{70}\right)}{70}}$$

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

$$= \frac{65}{80} - \frac{39}{70} \pm 2.58\sqrt{\frac{0.81(0.19)}{80} + \frac{0.54(0.44)}{70}}$$

$$= 0.25 \pm 0.19$$

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

Example 5 (continued)

$$0.06 < p_1 - p_2 < 0.44$$

- Could you conclude, based on this confidence interval, that there is a difference in the proportions of obesity between male and female college students?
- The confidence interval does not contain 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 obesity among males than females students.

Class Activity

6. 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.

$$Answer = -0.16 \pm 0.026$$

$\hat{p}_1 = \frac{337}{800}, \hat{p}_2 = \frac{374}{640}$

$(\hat{p}_1 - \hat{p}_2) \pm 2.33\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$

Class Activity

7. 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. *Answer* : -0.04 ± 0.18

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.

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.