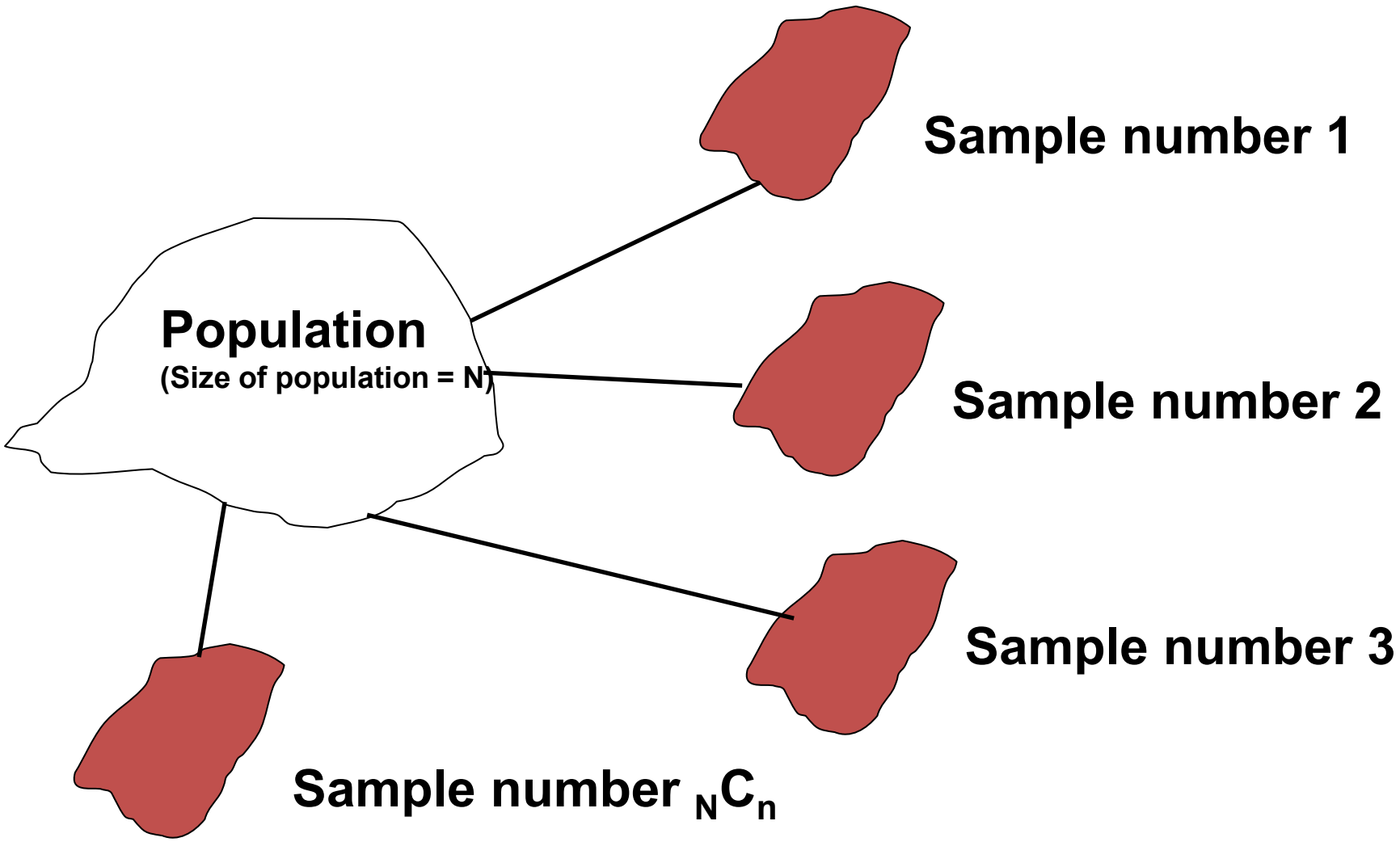
# Lecture 7

# Large and Small-Sample Estimation

# Large sample estimation


Population
(Size of population = N)
Sample number 1
Sample number 2
Sample number 3
Sample number $_{N}C_{n}$
Each sample size = n

# INTRODUCTION

- Populations are described by their probability distributions and parameters.
  - For quantitative populations, the location and shape are described by $\mu$ and $\sigma$.
  - 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 m 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 $p = 0.2$?"

# 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 $\mu$, 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, $\mu_N$ equal to the old average resistance, $\mu_O$?

# 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

# WHAT DO WE FREQUENTLY NEED TO ESTIMATE?

- An unknown population proportion p
- An unknown population mean $\mu$

# DEFINITIONS

- 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.

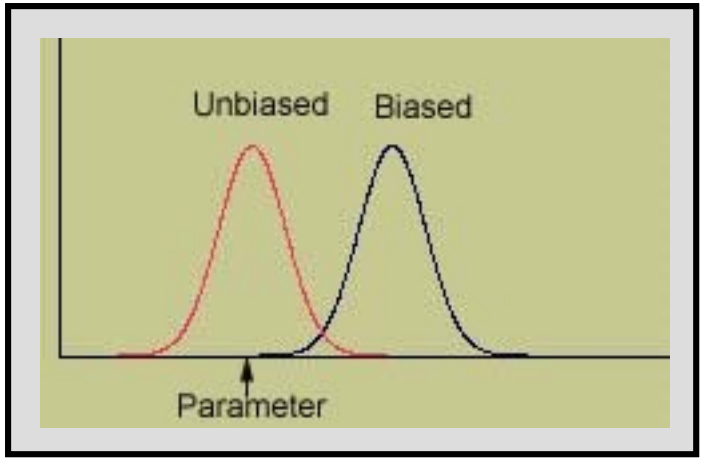
# A. Point Estimators

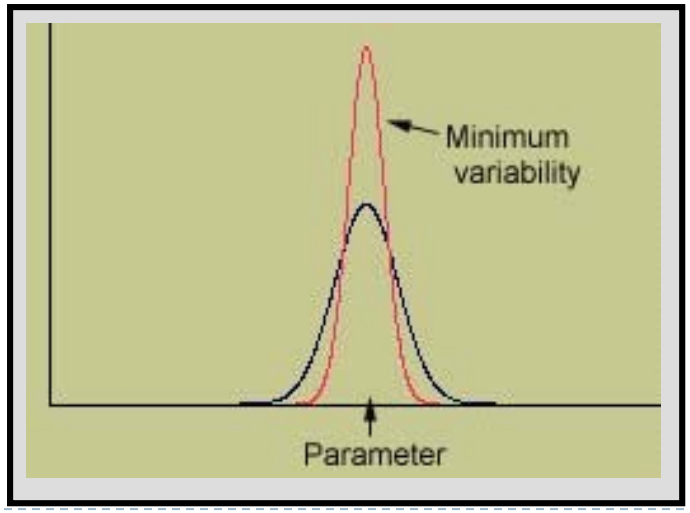
**Properties**

- 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.

## Properties

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



# Measuring the Goodness of an Estimator

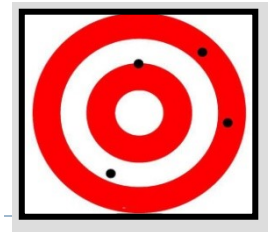
- The distance between an estimate and the true value of the parameter is the **error of estimation.**

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

- When the sample sizes are large, our ***unbiased*** estimators will have **normal** distributions.

Because of the Central Limit Theorem.

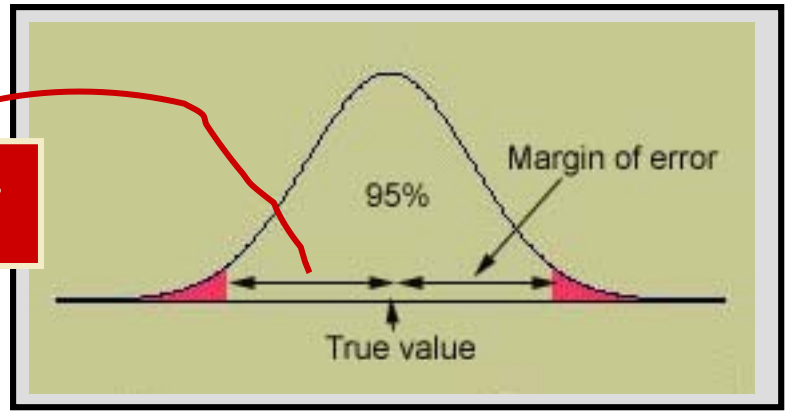
# The Margin of Error



- **Margin of error:** The maximum error of estimation, is the maximum likely difference observed between sample mean $x$ and true population mean $\mu$, calculated as :

1.645 **1.96** 2.33 2.575

$z_{\alpha/2} \times$ std error of the estimator

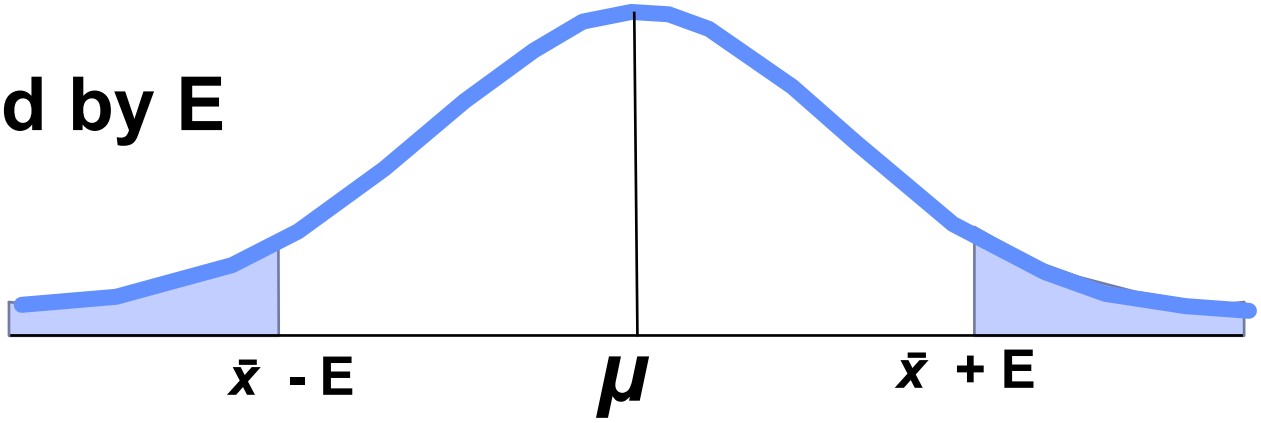


# Definition

## Margin of Error

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

**denoted by E**




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

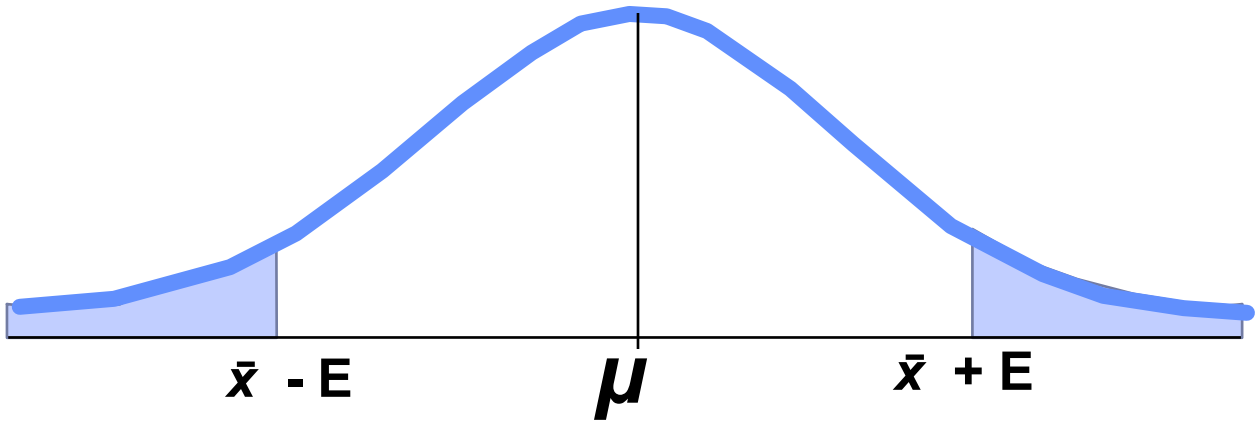
lower limit

upper limit

# Definition

## Margin of Error

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




**also called the maximum error of the estimate**

# Estimating Means and Proportions

- For a quantitative population,

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

Margin of error $(n > 30)$ : $\pm z_{\alpha/2} \frac{s}{\sqrt{n}}$

- For a binomial population,

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

Margin of error $(n > 30)$ : $\pm z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$

SE

1.645
1.96
2.33
2.575

# 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 $\mu$: $\bar{x} = 252,000$

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

# Example

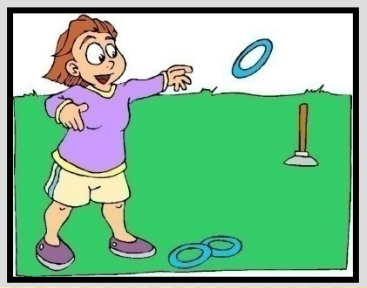

A quality control technician wants to estimate the proportion of soda cans that are underfilled. He randomly samples 200 cans of soda and finds 10 underfilled cans.

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

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

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

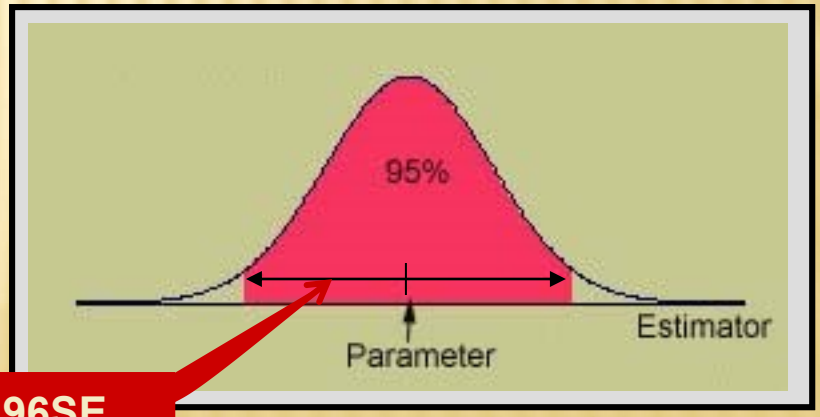
# B. INTERVAL ESTIMATION



- 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

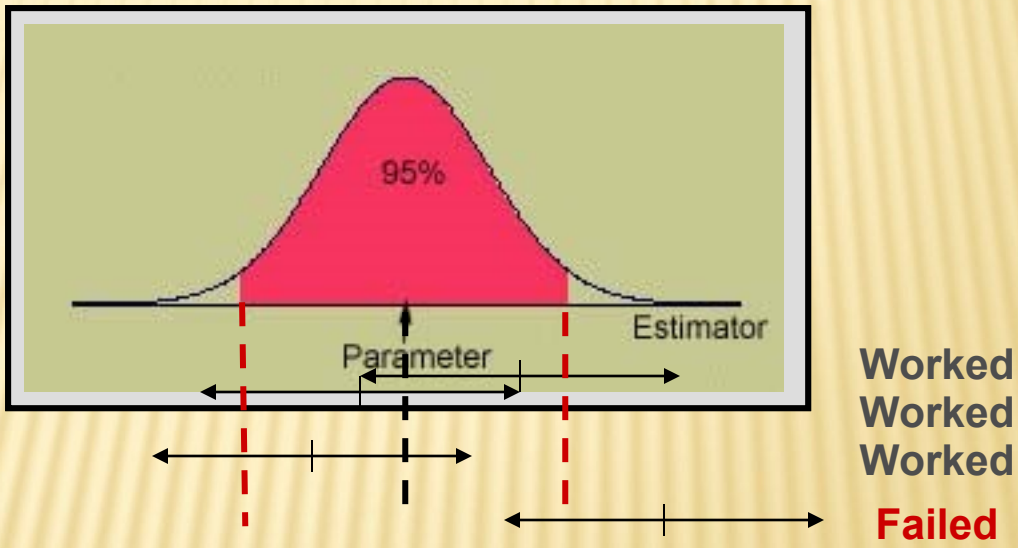
- Suppose 1-α = .95 and that the estimator has a normal distribution.




**Parameter ± 1.96SE**

# INTERVAL ESTIMATION *(CONT'D)*

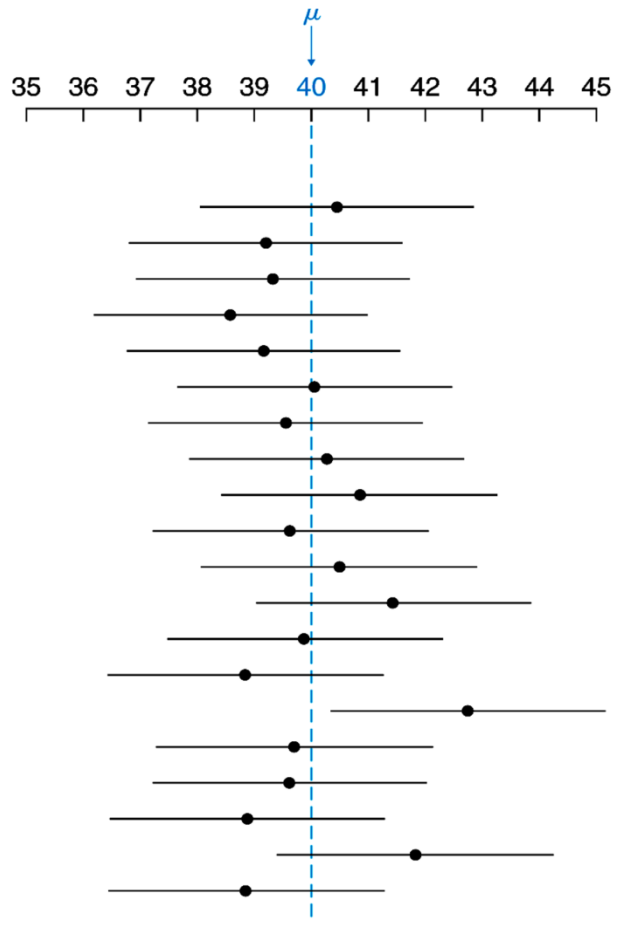
- Since we don't know the value of the parameter, consider **Estimator ± 1.96 SE** which has a variable center.




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

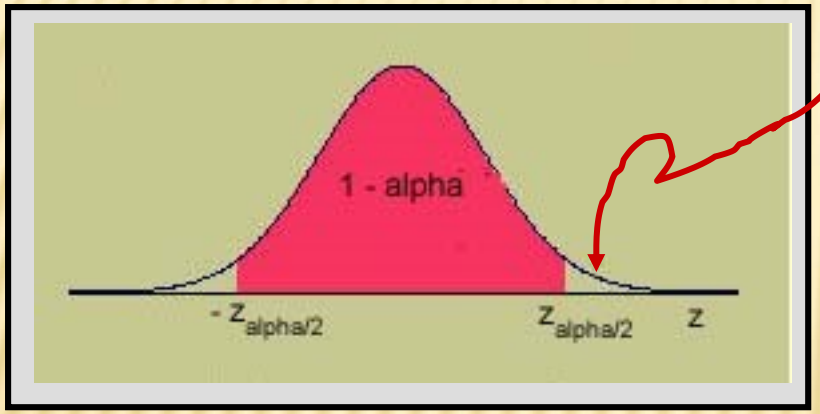
# INTERVAL ESTIMATION/*CONFIDENCE INTERVALS*

| Sample | $\bar{x}$ | 95.44% CI | $\mu$ in CI? |
|---|---|---|---|
| 1 | 40.45 | 38.06 to 42.85 | yes |
| 2 | 39.21 | 36.81 to 41.61 | yes |
| 3 | 39.33 | 36.93 to 41.73 | yes |
| 4 | 38.59 | 36.19 to 40.99 | yes |
| 5 | 39.17 | 36.77 to 41.57 | yes |
| 6 | 40.07 | 37.67 to 42.47 | yes |
| 7 | 39.56 | 37.16 to 41.96 | yes |
| 8 | 40.28 | 37.88 to 42.68 | yes |
| 9 | 40.87 | 38.48 to 43.27 | yes |
| 10 | 39.61 | 37.22 to 42.01 | yes |
| 11 | 40.51 | 38.11 to 42.91 | yes |
| 12 | 41.45 | 39.05 to 43.85 | yes |
| 13 | 39.88 | 37.48 to 42.28 | yes |
| 14 | 38.85 | 36.45 to 41.25 | yes |
| 15 | 42.73 | 40.33 to 45.13 | no |
| 16 | 39.70 | 37.30 to 42.10 | yes |
| 17 | 39.60 | 37.20 to 42.00 | yes |
| 18 | 38.88 | 36.48 to 41.28 | yes |
| 19 | 41.82 | 39.42 to 44.22 | yes |
| 20 | 38.84 | 36.45 to 41.24 | yes |



# TO CHANGE THE CONFIDENCE LEVEL

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




| Tail area | $z_{\alpha/2}$ |
|---|---|
| .05 | 1.645 |
| .025 | 1.96 |
| .01 | 2.33 |
| .005 | 2.575 |

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

# 1. CONFIDENCE INTERVALS FOR MEANS AND PROPORTIONS

- For a quantitative population,

Confidence interval for a population mean $\mu$:

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

- For a binomial population,

Confidence interval for a population proportion $p$:

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

1.96

# 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.

1.96

$$\bar{x} \pm z_{0.05/2}\frac{s}{\sqrt{n}} \Rightarrow 756 \pm 1.96\frac{35}{\sqrt{50}} \Rightarrow 756 \pm 9.70$$

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

# EXAMPLE


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

2.575

$$\bar{x} \pm z_{0.01/2}\frac{s}{\sqrt{n}} \Rightarrow 756 \pm 2.58\frac{35}{\sqrt{50}} \Rightarrow 756 \pm 12.75$$

or $743.25 < \mu < 768.75$ grams.

or $746.30 < \mu < 765.70$ 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.

2.33

$$\hat{p} \pm z_{0.02/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} \Rightarrow \frac{104}{150} \pm 2.33\sqrt{\frac{.69(.31)}{150}}$$

$$\Rightarrow \quad .69 \pm .09$$

or $.60 < p < .78.$

# 2. 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.
- To make this comparison,

A random sample of size $n_1$ drawn from population 1 w

A random sample of size $n_2$ drawn from population 2 with mean $\mu_2$ and variance $\sigma_2^2$.

# Estimating the Difference between Two Means *(Cont'd)*

- We compare the two averages by making inferences about $\mu_1$-$\mu_2$, the difference in the two population averages.
  - If the two population averages are the same, then $\mu_1$-$\mu_2$ = 0.
  - The best estimate of $\mu_1$-$\mu_2$ is the difference in the two sample means,

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

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

- Properties of the Sampling Distribution of $\bar{x}_1 - \bar{x}_2$
  - Expected Value

$$E(\bar{x}_1 - \bar{x}_2) = \mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$$

  - Standard Deviation/Standard Error

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where: $\sigma_1$ = standard deviation of population 1

$\sigma_2$ = standard deviation of population 2

$n_1$ = sample size from population 1

$n_2$ = sample size from population 2

# Interval Estimate of $\mu_1 - \mu_2$: Large-Sample Case ($n_1 > 30$ and $n_2 > 30$)

**Interval Estimate with $\sigma_1$ and $\sigma_2$ Known**

$$\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2}\sigma_{\bar{x}_1-\bar{x}_2}$$

SE

where:

1 - $\alpha$ is the confidence coefficient

- **Interval Estimate with $\sigma_1$ and $\sigma_2$ Unknown**

$$\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2}s_{\bar{x}_1-\bar{x}_2}$$

where:

$$s_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad \sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

# EXAMPLE


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

1.96

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

$$(\bar{x}_1 - \bar{x}_2) \pm z_{0.05/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

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

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

# EXAMPLE *(CONT'D)*

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

- 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.

# 3. 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 populat

A random sample of size $n_2$ drawn from binomial population 2 with parameter $p_2$.

# Estimating the Difference between Two Proportions *(cont'd)*

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

- Expected Value/mean

$$E(\hat{p}_1 - \hat{p}_2) = \mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$$

- Standard Deviation/Standard Error

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

- <u>Distribution Form</u>

  If the sample sizes are large ($n_1p_1$, $n_1q_1$, $n_2p_2$, $n_2q_2$) are all greater than to 5), the sampling distribution of $\hat{p}_1 - \hat{p}_2$ can be approximated by a normal probability distribution.

# Interval Estimate of $p_1$ - $p_2$: Large-Sample Case

- **Interval Estimate with $p_1$ and $p_2$ <u>Known</u>**

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\,\sigma_{\hat{p}_1-\hat{p}_2}$$

where:

1 - $\alpha$ is the confidence coefficient

- **Interval Estimate with $p_1$ and $p_2$ <u>Unknown</u>**

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\,\sigma_{\hat{p}_1-\hat{p}_2}$$

where: $\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\dfrac{\hat{p}_1\hat{q}_1}{n_1} + \dfrac{\hat{p}_2\hat{q}_2}{n_2}}$

$$\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_1q_1}{n_2}}$$

# 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.

2.575

$$(\hat{p}_1 - \hat{p}_2) \pm z_{0.01/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

$$\Rightarrow \left(\frac{65}{80} - \frac{39}{70}\right) \pm 2.575\sqrt{\frac{.81(.19)}{80} + \frac{.56(.44)}{70}} \Rightarrow \quad 0.26 \pm 0.19$$

$$\text{or} \quad 0.07 < p_1 - p_2 < 0.45$$

# Example *(cont'd)*

$$0.07 < p_1 - p_2 < 0.45$$

- 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.**

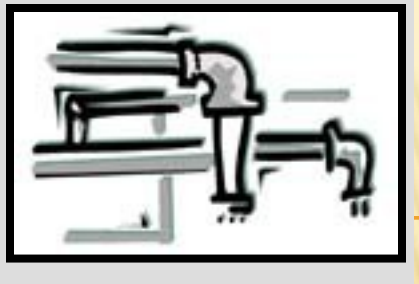
# ONE SIDED CONFIDENCE BOUNDS

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

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

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

# CHOOSING THE SAMPLE SIZE

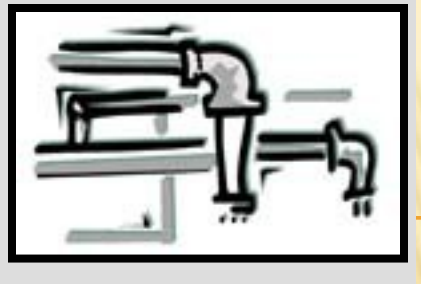


- The total amount of relevant information in a sample is controlled by two factors:
  - The **sampling plan** or **experimental design**: the procedure for collecting the information
  - The **sample size *n***: the amount of information you collect.
- In a statistical estimation problem, the accuracy of the estimation is measured by the **margin of error** or the **width of the confidence interval.**

# CHOOSING THE SAMPLE SIZE *(CONT'D)*

1. Determine the size of the **margin of error**, ***E***, that you are willing to tolerate.
2. Choose the sample size by solving for $n$ or $n = n_1 = n_2$ in the inequality: $1.96\,SE \leq E$, 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 0.04 of the actual proportion with probability equal to 0.95?

$$1.96\sqrt{\frac{pq}{n}} \le .04 \quad \Rightarrow 1.96\sqrt{\frac{0.5(0.5)}{n}} \le 0.04$$

$$\Rightarrow \sqrt{n} \ge \frac{1.96\sqrt{0.5(0.5)}}{0.04} = 24.5 \quad \Rightarrow n \ge 24.5^2 = 600.25$$

He should survey at least 600 wholesalers.

# 4. Estimating the Variance

The ***sample variance*** is defined by

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

$$= \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}$$

$$Var(X) = E(X - E(X))^2$$

$$= E(X-\mu)^2 = \sigma^2$$

$$E\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right) = E\left(\sum_{i=1}^{n}(x_i - \mu)(\mu - \bar{x})\right)^2$$

$$= \sum E\left[(x_i - \mu)^2\right] - nE\left[(\bar{x} - \mu)^2\right]$$

$$= n\sigma^2 - \frac{n\sigma^2}{n}$$

$$E\left(\sum_{i=1}^{n}(x_i - \mu)^2\right) = (n-1)\sigma^2$$

$$E\left(\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1}\right) = \sigma^2 \Rightarrow E(s^2) = \sigma^2$$

# Analysis of Sample Variance

If $s^2$ is the variance of a random sample size *n* from a normal population, a 100(1-$\alpha$)% confidence interval for $\sigma^2$ is

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

Where $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$ are $\chi^2$ values with (*n-1*) degrees of freedom.