Topic 3

Sampling Distribution and Estimation

Section 1 – Estimation

Section 1.1 Point Estimate

Statistical inference enables us to make judgments about a population on the basis of sample information.

The mean, standard deviation, and proportions of a population are called population parameters; in other words, they serve to define the population.

Estimating a population's parameters is essential to statistical analysis, and sometimes sampling is the best (fastest and most economical) way to approach the study.

Section 1 – Estimation

Section 1.1 Point Estimate

Definitions:

A **parameter** or **population parameter** is a characteristic of an entire population.

A **statistic** is a summary measure that is computed to describe a characteristic for only a sample of the population.

An **estimate** is a specific observed value of a statistic.

Section 1 – Estimation

Section 1.2 Estimator

The rule that specifies how a sample statistic can be obtained for estimating the population parameter is called an **estimator**. It is the random variable, defined by a formula, from which we obtain all possible estimates.

The **point estimate** is the single number that is obtained from the estimator. It is a single value calculated from only one sample, used to estimate a population parameter.

Point estimation is a process that generates specific numbers, each of which is a point estimate.

Section 1 – Estimation
Section 1.2 Estimator

The symbols we use to represent several important population parameters and their sample counterparts:

	Population Parameter	Sample Statistic
Mean	μ	$\overline{X}$
Standard deviation	σ	s
Variance	σ^2	s^2
Proportion	p	$\overline{p}$

Section 1 – Estimation
Section 1.2 Estimator

Example:

If a professor wants information on central tendency in a list of test scores, she can calculate a sample mean.

The number for the sample mean is called the **estimate**, and the sample mean is the **estimator** for the population mean.

Section 1 – Estimation
Section 1.2 Estimator

Example:

Suppose that a professor, whose course has an enrollment of 50 students, wants information on the performance of his class.

He takes a sample of 10 scores:

95, 67, 89, 70, 56, 97, 68, 78, 50, 79

Section 1 – Estimation
Section 1.2 Estimator

The **estimator** for the population mean is the sample mean, $\bar{X}$.

The **estimate** for the population mean, on the basis of the 10 sample scores, is

$$\bar{X} = \frac{95 + 67 + \cdots + 79}{10} = 74.9$$

Section 1 – Estimation

Section 1.2 Estimator

The **estimator** for the population variance is the sample variance, s^2.

The **estimate** of the population variance is

$$s^2 = \frac{\left(95^2 + 67^2 + \cdots + 79^2\right) - 10(74.9)^2}{10 - 1} = 247.65$$

The professor can use $\bar{X} = 74.9$ and $s^2 = 247.65$ to do his or her class performance analysis. The formula for combinations reveals that there are ${}_{50}C_{10} = 10{,}272{,}278{,}000$ possible estimates each for the population mean and the population variance.

Section 1 – Estimation

Section 1.2 Estimator

Definition:

An Interval Estimate is constructed around the point estimate, and it is stated that this interval is likely to contain the corresponding population parameter. Interval estimates indicate the precision, or accuracy, of an estimate and are therefore preferable.

In order to have an in-depth study of the interval estimate, we have to study the sampling distribution for the estimated parameters (i.e. $\bar{X}$, s^2, and $\bar{p}$).

Section 2 – Sampling Distribution
Section 2.1 Sampling Distribution of the Sample Mean

Definition:

The **population distribution** is the probability distribution of the population data.

The probability distribution of $\overline{X}$ is called the **sampling distribution** of $\overline{X}$. It lists the various values that $\overline{X}$ can assume and the probability of each value of $\overline{X}$.

Section 2 – Sampling Distribution
Section 2.1 Sampling Distribution of the Sample Mean

Example:

Suppose there are only five students in an advanced statistics class and the midterm scores of these five students are:

70   78   80   80   95

Section 2 – Sampling Distribution
Section 2.1 Sampling Distribution of the Sample Mean

Let X denote the score of a student. Using single-valued classes, the frequency distribution of scores is depicted as follows:

X	f	$f(x)$
70	1	0.2
78	1	0.2
80	2	0.4
95	1	0.2

The values of the mean and standard deviation calculated for the probability distribution give the values of the population parameters $\mu = 80.6$ and $\sigma = 8.0895$.

Section 2 – Sampling Distribution
Section 2.1 Sampling Distribution of the Sample Mean

Example:

Reconsider the population of midterm scores of five students given in the previous example.

Consider all possible samples of three scores each that can be selected, without replacement, from that population.

→ The total number of possible samples is ${}_5C_3 = 10$.

Section 2 – Sampling Distribution

Section 2.1 Sampling Distribution of the Sample Mean

Suppose we assign letters A, B, C, D, and E to the scores of five students so that

A = 70, B = 78, C = 80, D = 80, E = 95.

Then the 10 possible samples of three scores each are ABC, ABD, ABE, ACD, ACE, ADE, BCD, BCE, BDE, CDE.

Section 2 – Sampling Distribution

Section 2.1 Sampling Distribution of the Sample Mean

These 10 samples and their respective means are listed in the following table:

Sample	Scores in the Sample			$\bar{X}$
ABC	70	78	80	76.00
ABD	70	78	80	76.00
ABE	70	78	95	81.00
ACD	70	80	80	76.67
ACE	70	80	95	81.67
ADE	70	80	95	81.67
BCD	78	80	80	79.33
BCE	78	80	95	84.33
BDE	78	80	95	84.33
CDE	80	80	95	85.00

Section 2 – Sampling Distribution

Section 2.1 Sampling Distribution of the Sample Mean

By using the value of $\bar{X}$, we record the frequency distribution of $\bar{X}$ as follows:

$\bar{X}$	f	$f(\bar{X})$
76.00	2	0.2
76.67	1	0.1
79.33	1	0.1
81.00	1	0.1
81.67	2	0.2
84.33	2	0.2
85.00	1	0.1

Section 2 – Sampling Distribution

Section 2.1 Sampling Distribution of the Sample Mean

Sampling Error is the difference between the value of a sample statistic and the value of the corresponding population parameter.
In the case of the mean,

$$\text{Sample Error} = \bar{X} - \mu$$

assuming that the sample is random and no non-sampling error has been made. A sampling error occurs because of chance.

Non-sampling Errors are errors that occur in the collection, recording, and tabulation of data. Such errors occur because of human mistakes and not chance.

Section 2 – Sampling Distribution
Section 2.1 Sampling Distribution of the Sample Mean

Comparison between Sampling and Non-Sampling Errors

Sampling errors	Non-sampling errors
•occurs only when a sample survey is conducted •impossible to avoid sampling error	•occur both in a sample survey and in a census •can be minimized by preparing the survey questionnaire carefully and handling the data cautiously

Section 2 – Sampling Distribution
Section 2.1 Sampling Distribution of the Sample Mean

Example:

Reconsider the population of midterm scores of five students given in the previous example.

The population mean is

$$\mu = \frac{70+78+80+80+95}{5} = 80.60$$

Section 2 – Sampling Distribution

Section 2.1 Sampling Distribution of the Sample Mean

Now suppose we take a random sample of three scores from this population. Assume that this sample includes the scores 70, 80, and 95. The mean for this sample is

$$\bar{X} = \frac{70+80+95}{3} = 81.67$$

Consequently, Sample Error = $\bar{X} - \mu = 81.67 - 80.60 = 1.07$

That is, the mean score estimated from the sample is 1.07 higher than the mean score of the population. Note that this difference occurred due to chance, that is, because we used a sample instead of the population.

Section 2 – Sampling Distribution

Section 2.1 Sampling Distribution of the Sample Mean

Now suppose, when we select the above mentioned sample, we mistakenly record the second score as 82 instead of 80. As a result, we calculate the sample mean as

$$\bar{X} = \frac{70+82+95}{3} = 82.33$$

Consequently, this difference between the sample mean and the population mean is

$$\bar{X} - \mu = 82.33 - 80.60 = 1.73$$

Section 2 – Sampling Distribution

Section 2.1 Sampling Distribution of the Sample Mean

However, this difference between the sample mean and the population mean does not represent the sampling error.

As we calculated earlier, only 1.07 of this difference is due to the sampling error.

The remaining portion, which is equal to $1.73 - 1.07 = 0.66$ represents the non-sampling error because it occurred due to the error we made in recording the second score in the sample.

→ Sampling error $= 1.07$, Non-sampling error $= 0.66$

Section 2 – Sampling Distribution

Section 2.1 Sampling Distribution of the Sample Mean

The mean and standard deviation calculated for the sampling distribution of $\bar{X}$ are called the mean $\mu_{\bar{X}}$ and standard deviation $\sigma_{\bar{X}}$ of $\bar{X}$.

Actually, the mean and standard deviation of $\bar{X}$ are, respectively, the mean and standard deviation of the means of all samples of the same size selected from a population.

The standard deviation of $\sigma_{\bar{X}}$ is also called the standard error of $\bar{X}$.

Section 2 – Sampling Distribution

Section 2.1 Sampling Distribution of the Sample Mean

Mean of the Sampling Distribution of $\bar{X}$

The mean of the sampling distribution of $\bar{X}$ is equal to the mean of the population. Thus,

$$\mu_{\bar{X}} = \mu$$

Section 2 – Sampling Distribution

Section 2.1 Sampling Distribution of the Sample Mean

Standard Deviation of the Sampling Distribution of $\bar{X}$

The standard deviation of the sampling distribution of $\bar{X}$ is

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

where σ is the standard deviation of the population and n is the sample size. This formula is used when $n / N \leq 0.05$, where N is the population size.

Section 2 – Sampling Distribution

Section 2.1 Sampling Distribution of the Sample Mean

If this condition is not satisfied, we use the following formula to calculate $\sigma_{\bar{X}}$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \cdot \underset{\uparrow}{\sqrt{\frac{N-n}{N-1}}}$$

finite population
correction factor

Section 2 – Sampling Distribution

Section 2.1 Sampling Distribution of the Sample Mean

The Shape of the Sampling Distribution of $\bar{X}$

Case I:
Sampling from a Normally Distributed Population

Case II:
Sampling from a population that is not Normally Distributed

Section 2 – Sampling Distribution

Section 2.1 Sampling Distribution of the Sample Mean

Case I: Sampling from a Normally Distributed Population

When the population from which samples are drawn is normally distributed with its mean equals to μ and standard deviation equal to σ , then

1. The shape of the sampling distribution of $\overline{X}$ is normal, whatever the value of *n*.
2. The mean of $\overline{X}$, $\mu_{\overline{X}}$, is equal to μ .
3. The standard deviation of $\overline{X}$, $\sigma_{\overline{X}}$, is equal to $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$.

(assume $n / N \leq 0.05$)

Section 2 – Sampling Distribution

Section 2.1 Sampling Distribution of the Sample Mean

Case II: Sampling from a population that is not Normally Distributed

Most of the time the population from which the samples are selected is not normally distributed.

However, if the sample size is at least 30, the shape of the sampling distribution of $\overline{X}$ is inferred from a very important theorem called the Central Limit Theorem (CLT).

Section 2 – Sampling Distribution
Section 2.1 Sampling Distribution of the Sample Mean

Central Limit Theorem (CLT)

For a large sample size (usually considered large if $n \geq 30$)

1. The sampling distribution of the sample mean $\overline{X}$ is approximately normal, irrespective of the shape of the population distribution.
2. The mean of $\overline{X}$, $\mu_{\overline{X}}$, is equal to μ .
3. The standard deviation of $\overline{X}$, $\sigma_{\overline{X}}$, is equal to $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$.

If the population distribution is fairly symmetrical, the sampling distribution of the sample mean $\overline{X}$ is approximately normal if sample size $n \geq 15$.

Section 2 – Sampling Distribution
Section 2.1 Sampling Distribution of the Sample Mean

	Sampling Distribution of $\overline{X}$	
	Normal Population	Non-normal Population
Mean	$\mu_{\overline{X}} = \mu$	$\mu_{\overline{X}} = \mu$
Standard error	$\sigma_{\overline{X}} = \sigma / \sqrt{n}$	$\sigma_{\overline{X}} = \sigma / \sqrt{n}$
Shape	Normal	Approximate Normal if $n \geq 30$
Notation	$\overline{X} \sim N\left(\mu, \left(\dfrac{\sigma}{\sqrt{n}}\right)^2\right)$	$\overline{X} \sim N\left(\mu, \left(\dfrac{\sigma}{\sqrt{n}}\right)^2\right)$

Section 2 – Sampling Distribution

Section 2.1 Sampling Distribution of the Sample Mean

Example:

A company which manufactures drink dispensing machines sets the fill level at 198cc. The standard deviation is 4cc. Assume that the fill levels have a normal distribution.

(a) A drink is randomly selected, what is the probability that the drink will have less than 195cc?

(b) What is the probability that a random sample of 50 drinks has a mean value greater than 199cc?

Section 2 – Sampling Distribution

Section 2.1 Sampling Distribution of the Sample Mean

Solution:

(a) A drink is randomly selected, what is the probability that the drink will have less than 195cc?

Let X be the fill level and μ be the mean fill level.
Given $X \sim N(198, 4^2)$,

$$P(X < 195) = P\left(\frac{X-\mu}{\sigma} < \frac{195-198}{4}\right)$$

$$= P\left(Z < -0.75\right) = 0.2266$$

Section 2 – Sampling Distribution

Section 2.1 Sampling Distribution of the Sample Mean

(b) What is the probability that a random sample of 50 drinks has a mean value greater than 199cc?

Let $\bar{X}$ be the sample mean. Since the population is normally distributed, thus the shape of the sampling distribution of $\bar{X}$ is normal. We have,

$$\mu_{\bar{X}} = \mu = 198; \ \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{50}} \quad \Rightarrow \quad \bar{X} \sim N\left(198, \left(\frac{4}{\sqrt{50}}\right)^2\right)$$

$$P(\bar{X} > 199) = P\left(\frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} > \frac{199 - 198}{4/\sqrt{50}}\right) = P\left(Z < 1.77\right) = 0.0384$$

Section 2 – Sampling Distribution

Section 2.2 Sampling Distribution of the Sample Proportion

Definition:

The population proportion, denoted by p, is obtained by taking the ratio of the number of elements in a population with a specific characteristic to the total number of elements in the population.

The sample proportion, denoted by $\bar{p}$, gives a similar ratio for a sample.

Section 2 – Sampling Distribution
Section 2.2 Sampling Distribution of the Sample Proportion

The population and sample proportions, denoted by p and $\bar{p}$, respectively, are calculated as

$$p = \frac{x}{N} \qquad \text{and} \qquad \bar{p} = \frac{x}{n}$$

where

N – Total number of elements in the population

n – Total number of elements in the sample

x – Number of elements in the population or sample that possess a specific characteristic

Section 2 – Sampling Distribution
Section 2.2 Sampling Distribution of the Sample Proportion

Example:

Suppose a total of 789,654 families live in a city and 563,282 of them own homes.

Then, $N = 789{,}654$ and $x = 563{,}282$

The proportion of all families in this city who own homes is

$$p = \frac{x}{N} = \frac{563{,}282}{789{,}654} = 0.7133$$

Section 2 – Sampling Distribution

Section 2.2 Sampling Distribution of the Sample Proportion

Now, suppose a sample of 240 families is taken from this city and 158 of them are homeowner.

Then, $n = 240$ and $x = 158$.

The sample proportion is $\overline{p} = \dfrac{x}{n} = \dfrac{158}{240} = 0.6583$

The difference between the sample proportion and the corresponding population proportion gives the sampling error, assuming that the sample is random and no non-sampling error has been made. That is, in case of the proportion,

$$\text{Sample Error} = \overline{p} - p = 0.6583 - 0.7133 = -0.055$$

Section 2 – Sampling Distribution

Section 2.2 Sampling Distribution of the Sample Proportion

The probability distribution of the sample proportion $\overline{p}$ is called the sampling distribution of $\overline{p}$.

It gives the various values that $\overline{p}$ can assume and their probabilities.

Example:

Boe Consultant Associates has five employees. The following table gives the name of these five employees and information concerning their knowledge of statistics.

Name	Ally	John	Susan	Peter	Tom
Knows Statistics	Yes	No	No	Yes	Yes

Section 2 – Sampling Distribution
Section 2.2 Sampling Distribution of the Sample Proportion

If we define the population proportion *p* as the proportion of employees who know statistics, then, $p = 3/5 = 0.6$.

Now, suppose we draw all possible samples of three employees each and compute the proportion of employees, for each sample, who know statistics. The total number of samples of size three that can be drawn from the population of five employees is ${}_5C_3 = 10$.

Section 2 – Sampling Distribution
Section 2.2 Sampling Distribution of the Sample Proportion

The table lists the 10 possible samples and the proportion of employees who know for each of those samples

Sample	Prop. ($\bar{p}$)	Sample	Prop. ($\bar{p}$)
Ally, John, Susan	1/3 = 0.33	Ally, Peter, Tom	3/3 = 1.00
Ally, John, Peter	2/3 = 0.67	John, Susan, Peter	1/3 = 0.33
Ally, John, Tom	2/3 = 0.67	John, Susan, Tom	1/3 = 0.33
Ally, Susan, Peter	2/3 = 0.67	John, Peter, Tom	2/3 = 0.67
Ally, Susan, Tom	2/3 = 0.67	Susan, Peter, Tom	2/3 = 0.67

Section 2 – Sampling Distribution

Section 2.2 Sampling Distribution of the Sample Proportion

The sampling distribution of $\bar{p}$ as

$\bar{p}$	$f(\bar{p})$
0.33	0.30
0.67	0.60
1.00	0.10

Section 2 – Sampling Distribution

Section 2.2 Sampling Distribution of the Sample Proportion

Mean of the Sampling Distribution of $\bar{p}$

The mean of the sample proportion $\bar{p}$ is denoted by $\mu_{\bar{p}}$ and is equal to the population proportion p. Thus, $\mu_{\bar{p}} = p$.

Standard Deviation of the Sampling Distribution of $\bar{p}$

The standard deviation of the sampling distribution of $\bar{p}$ is

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \qquad \sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

$$n/N \le 0.05 \qquad n/N > 0.05$$

Section 2 – Sampling Distribution

Section 2.2 Sampling Distribution of the Sample Proportion

The Shape of the Sampling Distribution of $\overline{p}$

Central Limit Theorem – The sampling distribution of $\overline{p}$ is approximately normal for a sufficiently large sample size.

In the case of proportion, the sample size n is considered to be sufficiently large if $np \geq 5$ and $n(1-p) \geq 5$.

Section 2 – Sampling Distribution

Section 2.2 Sampling Distribution of the Sample Proportion

Sampling Distribution of $\overline{p}$

Mean	$\mu_{\overline{p}} = p$
Standard error	$\sigma_{\overline{p}} = \sqrt{\frac{p(1-p)}{n}}$
Shape	Normal if $np \geq 5$ and $n(1-p) \geq 5$
Notation	$\overline{p} \sim N\left(p, \left(\sqrt{\frac{p(1-p)}{n}}\right)^2\right)$

Section 2 – Sampling Distribution

Section 2.2 Sampling Distribution of the Sample Proportion

Example:

The election returns showed that a certain candidate received 46% of the votes.

(a) Determine the probability that a poll of 200 people selected at random from the voting population would have shown a majority (over 50%) of votes in favor of the candidate.

(b) 95% of the sample proportions will be greater than what value?

Section 2 – Sampling Distribution

Section 2.2 Sampling Distribution of the Sample Proportion

Solution:

(a) Determine the probability that a poll of 200 people selected at random from the voting population would have shown a majority (over 50%) of votes in favor of the candidate.

From the given information: $p = 0.46$

This gives:

$$\mu_{\bar{p}} = p = 0.46, \quad \sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(0.46)(0.54)}{200}} = 0.0352$$

Section 2 – Sampling Distribution

Section 2.2 Sampling Distribution of the Sample Proportion

Since $np = 200(0.46) = 92 > 5$ and $n(1-p) = 200(0.54) = 108 > 5$, we can infer from the Central Limit Theorem that the sampling distribution of $\overline{p}$ is approximately normal. Thus,

$$\overline{p} \sim N(0.46, (0.0352)^2)$$

Required probability:

$$P(\overline{p} > 0.50) = P\left(\frac{\overline{p}-\mu_{\overline{p}}}{\sigma_{\overline{p}}} > \frac{0.50-0.46}{0.0352}\right) = P(Z > 1.14) = 0.1271$$

Section 2 – Sampling Distribution

Section 2.2 Sampling Distribution of the Sample Proportion

(b) 95% of the sample proportions will be greater than what value?

Let A be the required value. We want $P(\overline{p} > A) = 0.95$ and from the standard normal table, $P(Z > -1.645) = 0.95$

$$\Rightarrow \quad \frac{A-0.46}{0.0352} = -1.645 \quad \Rightarrow \quad A = 0.4021$$

Section 2 – Sampling Distribution

Section 2.3 Sampling Distribution of the Sample Variance

Considering a random sample of n observations drawn from a population with unknown mean μ and unknown variance σ^2.

Denote the sample observations as $x_1, x_2, \ldots, x_n$.

The population variance is the expectation $\sigma^2 = E\left[(X-\mu)^2\right]$ which suggests that the mean of $(x_i-\mu)^2$ over n observations. Since μ is unknown, the sample mean $\overline{x}$ is used to compute a sample variance.

Section 2 – Sampling Distribution

Section 2.3 Sampling Distribution of the Sample Variance

Definition:

Let $x_1, x_2, \ldots, x_n$ be a random sample of observations from a population.

The quantity

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

is called the sample variance, and its square root, s, is called the sample standard deviation.

Section 2 – Sampling Distribution

Section 2.3 Sampling Distribution of the Sample Variance

Suppose a random sample of n observations with sample variance s^2 is taken from a normally distributed population with population variance σ^2.

Then,

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

has a chi-square (χ^2) distribution with $n-1$ degrees of freedom

Section 2 – Sampling Distribution

Section 2.3 Sampling Distribution of the Sample Variance

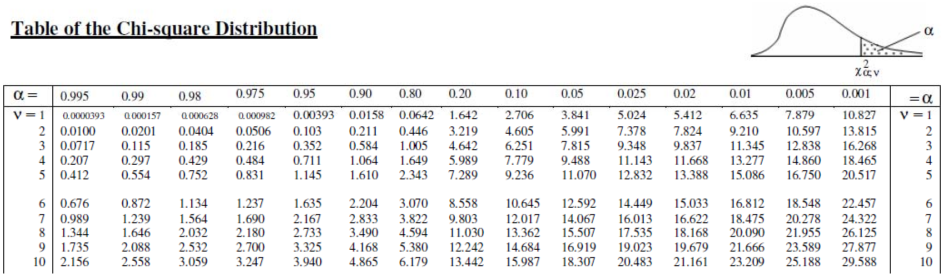

α
$\chi^2_{\alpha,\nu}$

Table of the Chi-square Distribution

α =	0.995	0.99	0.98	0.975	0.95	0.90	0.80	0.20	0.10	0.05	0.025	0.02	0.01	0.005	0.001	= α
ν = 1	0.0000393	0.000157	0.000628	0.000982	0.00393	0.0158	0.0642	1.642	2.706	3.841	5.024	5.412	6.635	7.879	10.827	ν = 1
2	0.0100	0.0201	0.0404	0.0506	0.103	0.211	0.446	3.219	4.605	5.991	7.378	7.824	9.210	10.597	13.815	2
3	0.0717	0.115	0.185	0.216	0.352	0.584	1.005	4.642	6.251	7.815	9.348	9.837	11.345	12.838	16.268	3
4	0.207	0.297	0.429	0.484	0.711	1.064	1.649	5.989	7.779	9.488	11.143	11.668	13.277	14.860	18.465	4
5	0.412	0.554	0.752	0.831	1.145	1.610	2.343	7.289	9.236	11.070	12.832	13.388	15.086	16.750	20.517	5
6	0.676	0.872	1.134	1.237	1.635	2.204	3.070	8.558	10.645	12.592	14.449	15.033	16.812	18.548	22.457	6
7	0.989	1.239	1.564	1.690	2.167	2.833	3.822	9.803	12.017	14.067	16.013	16.622	18.475	20.278	24.322	7
8	1.344	1.646	2.032	2.180	2.733	3.490	4.594	11.030	13.362	15.507	17.535	18.168	20.090	21.955	26.125	8
9	1.735	2.088	2.532	2.700	3.325	4.168	5.380	12.242	14.684	16.919	19.023	19.679	21.666	23.589	27.877	9
10	2.156	2.558	3.059	3.247	3.940	4.865	6.179	13.442	15.987	18.307	20.483	21.161	23.209	25.188	29.588	10

Verify: a. $\chi^2_{0.005,\,5} = 16.750$ b. $\chi^2_{0.9,\,9} = 4.168$

c. $P(\mathrm{X}^2 > \chi^2_\alpha) = 0.05$ with $\nu = 10$ → $\chi^2_\alpha = \chi^2_{0.05,10} = 18.307$

d. $P(\mathrm{X}^2 < \chi^2_\alpha) = 0.05$ with $\nu = 10$ → $\chi^2_\alpha = \chi^2_{0.95,10} = 3.940$

e. Given that $\mathrm{X}^2 \sim \chi^2_{22}$, $P(10.982 < \mathrm{X}^2 < 36.781) = 0.95$

Section 2 – Sampling Distribution

Section 2.3 Sampling Distribution of the Sample Variance

Mean of the Sampling Distribution of s^2

The mean of the sample variance s^2 is equal to the population variance σ^2 .

Variance of the Sampling Distribution of s^2

The variance of the sample variance s^2 is given by the formula

$$Var(s^2) = \frac{2\sigma^4}{n-1}$$

Section 2 – Sampling Distribution

Section 2.3 Sampling Distribution of the Sample Variance

Example:

The variability of the electrical resistance is critical for manufacturing a control device. Manufacturing standards specify a standard deviation of 3.6, and the population distribution of resistance measures is normal.

The monitoring process requires that a random sample for $n = 6$ observations be obtained from the population of devices and the sample variance be computed.

Determine an upper limit for the sample variance such that the probability of exceeding this limit, given a population standard deviation of 3.6, is less than 0.05.

Section 2 – Sampling Distribution
Section 2.3 Sampling Distribution of the Sample Variance

Solution:

From the given information, $n = 6$ and $\sigma^2 = 3.6^2 = 12.96$

Let K be the required upper bound.

We have,

$$P(s^2 > K) = P\left(\frac{(n-1)s^2}{12.96} > \chi_5^2\right) = 0.05$$

$\chi_5^2 = 11.07$ is the upper 0.05 critical value of the chi-square distribution with 5 d.f.

Section 2 – Sampling Distribution
Section 2.3 Sampling Distribution of the Sample Variance

The required upper limit for s^2 – labelled as K – can be obtained by

$$\frac{(n-1)s^2}{12.96} = \frac{(6-1)K}{12.96} = 11.07 \quad \Rightarrow \quad K = 28.69$$

If the sample variance, s^2, from a random sample of size $n = 6$ exceeds 28.69, there is strong evidence to suspect that the population variance exceeds 12.96 and that the manufacturing process should be halted and appropriate adjustments should be performed.

Section 2 – Sampling Distribution

Section 2.3 Sampling Distribution of the Sample Variance

Example:

A manager of a quality assurance food company wants to ensure the variation of package weights is small so that the company does not produce a large proportion of packages that are under the stated package weight. The manager wants to obtain upper and lower limits for the ratio of the sample variance divided by the population variance for a random sample of $n = 20$ observations.

The limits are such that the probability that the ratio is below the lower limit is 0.025 and the probability that the ratio is above the upper limit is 0.025. Thus, 95% of the ratios will be between these limits. The population distribution can be assumed to be normal.

Section 2 – Sampling Distribution

Section 2.3 Sampling Distribution of the Sample Variance

Solution:

To obtain values K_L and K_U such that

$$P\left(\frac{s^2}{\sigma^2} < K_L\right) = 0.025 \quad \text{and} \quad P\left(\frac{s^2}{\sigma^2} > K_U\right) = 0.025$$

given that $n = 20$ is used to compute the sample variance.

Section 2 – Sampling Distribution
Section 2.3 Sampling Distribution of the Sample Variance

For the lower limit: $$0.025 = P\left(\frac{(n-1)s^2}{\sigma^2} < (n-1)K_L\right) = P(\mathrm{X}^2 < (n-1)K_L)$$

$$8.91 = 19K_L \quad \Rightarrow \quad K_L = 0.4689$$

For the upper limit: $$0.975 = P\left(\frac{(n-1)s^2}{\sigma^2} \overset{<}{\cancel{>}} (n-1)K_U\right) = P(\mathrm{X}^2 < (n-1)K_U)$$

$$32.85 = 19K_U \quad \Rightarrow \quad K_U = 1.7289$$

→ The 95% acceptance interval for the ratio (s^2/σ^2) is

$$0.4689 \le s^2/\sigma^2 \le 1.7289$$

Section 2 – Sampling Distribution
Section 2.4 Properties of Estimators

A number of different estimators are possible for the same population parameter, but some estimators are better than others.

To understand how, we need to look at three important properties of estimators.

I. Unbiasedness

II. Efficiency

III. Consistency

Section 2 – Sampling Distribution
Section 2.4 Properties of Estimators

Unbiasedness

An estimator exhibits unbiasedness when the mean of the sampling estimator $\hat{\theta}$ is equal to the population parameter θ. That is, $E(\hat{\theta}) = \theta$.

The sample mean is an unbiased estimator of the population mean because the mean of the sampling distribution of $\bar{X}$, $E(\bar{X})$, is equal to the population mean μ.

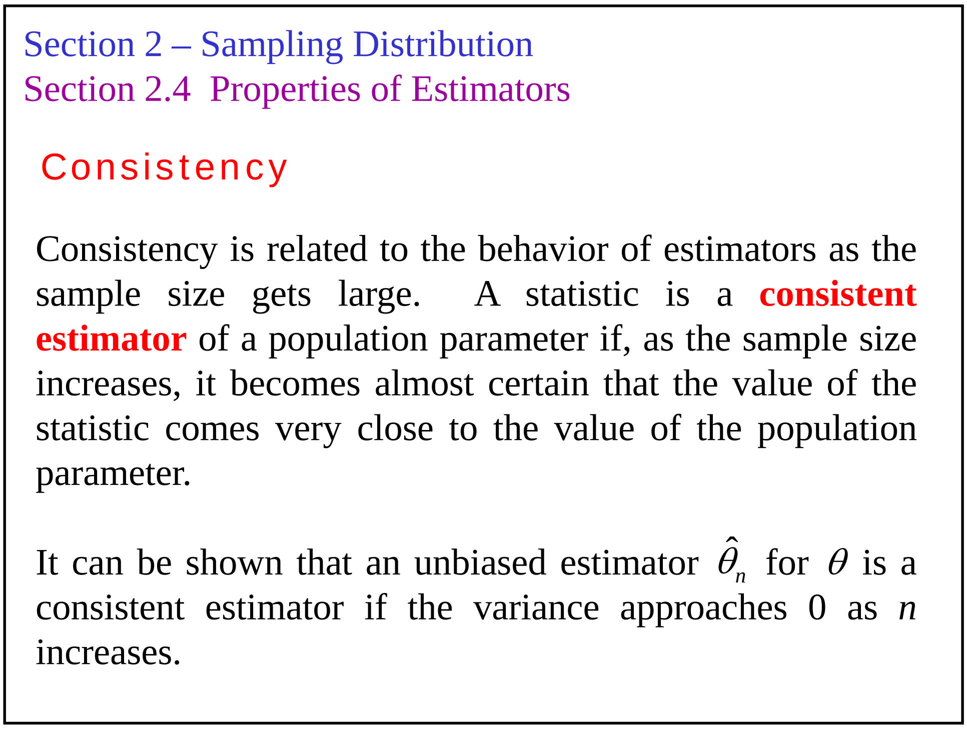
The sample proportion is an unbiased estimator of the population proportion, $E(\bar{p}) = p$.

Section 2 – Sampling Distribution
Section 2.4 Properties of Estimators

Efficiency

Efficiency refers to the size of the standard error of the statistics. The most efficient estimator is the one with the smallest variance.

Thus, if there are two estimators for θ with variances $Var(\hat{\theta}_1)$ and $Var(\hat{\theta}_2)$, then the first estimator $\hat{\theta}_1$ is said to be more efficient than the second estimator $\hat{\theta}_2$, if $Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$ although $E(\hat{\theta}_1) = E(\hat{\theta}_2) = \theta$.



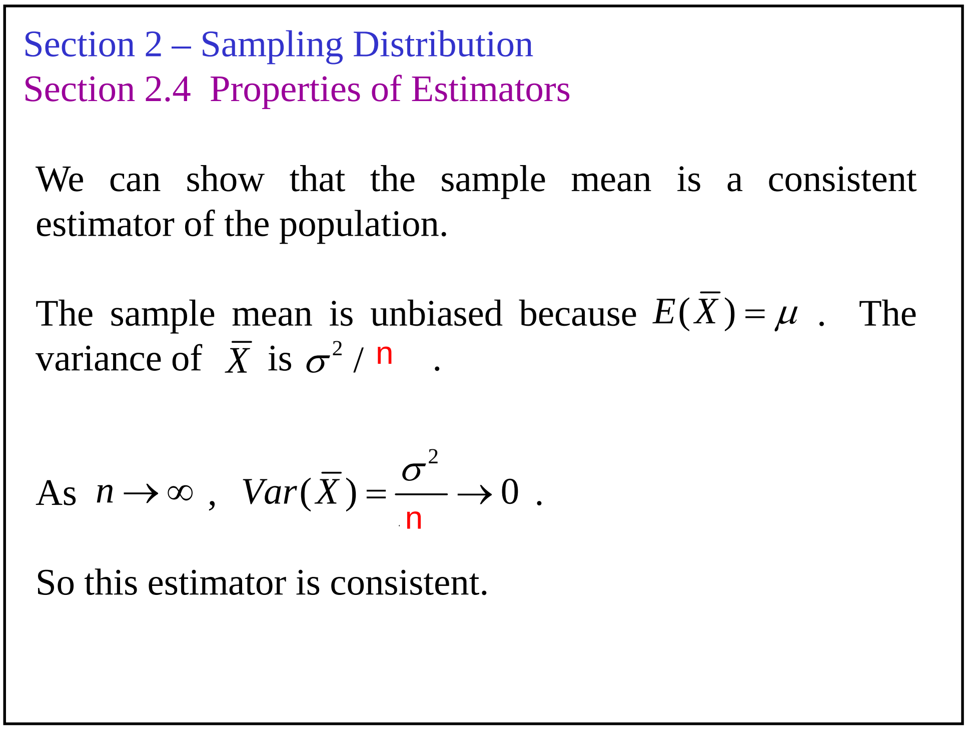
Section 2 – Sampling Distribution

Section 2.4 Properties of Estimators

Consistency

Consistency is related to the behavior of estimators as the sample size gets large. A statistic is a **consistent estimator** of a population parameter if, as the sample size increases, it becomes almost certain that the value of the statistic comes very close to the value of the population parameter.

It can be shown that an unbiased estimator $\hat{\theta}_n$ for θ is a consistent estimator if the variance approaches 0 as n increases.



Section 2 – Sampling Distribution

Section 2.4 Properties of Estimators

We can show that the sample mean is a consistent estimator of the population.

The sample mean is unbiased because $E(\bar{X}) = \mu$. The variance of $\bar{X}$ is σ^2 / n .

As $n \to \infty$, $Var(\bar{X}) = \dfrac{\sigma^2}{n} \to 0$.

So this estimator is consistent.

Section 3 – Confidence Interval

Definitions:

Each interval is constructed with regard to a given confidence level and is called a confidence interval. The confidence level associated with a confidence interval states how much confidence we have that this interval contains the true population parameter.

The confidence level is denoted by $(1-\alpha)100\%$. When expressed as a probability, it is called the confidence coefficient and is denoted by $1-\alpha$.

Section 3 – Confidence Interval

Although any value of the confidence level can be chosen to construct a confidence interval, the more common values are 90%, 95% and 99%. The corresponding confidence coefficients are 0.90, 0.95 and 0.99.

Section 3 – Confidence Interval

Interval Estimation of a Population Mean: Known Variances

Recall that in the case of $\overline{X}$, the sample size is considered to be large when $n \geq 30$. According to the central limit theorem, for a large sample the sampling distribution of the sample mean $\overline{X}$ is (approximately) normal irrespective of the shape of the population from which the sample is drawn.

Therefore, when $n \geq 30$, use the normal distribution to construct a confidence interval for μ.

Section 3 – Confidence Interval

Confidence Interval for population mean μ

The $(1-\alpha)100\%$ confidence interval for μ is

$$\boxed{\overline{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}$$

where

$\overline{X}$ is sample mean; σ is population standard deviation; n is the sample size; and $Z_{\alpha/2}$ is read from the standard normal distribution table for the given confidence level.

Conditions: Normal population with known variance
OR Non-normal population, large sample with known variance

Section 3 – Confidence Interval

Maximum Error of Estimate for μ

The maximum error of estimate for μ, denoted by E, is the quantity that is subtracted from and added to the value of $\bar{X}$ to obtain a confidence interval for μ.

Thus, given the $(1-\alpha)100\%$ confidence interval,

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

Section 3 – Confidence Interval

Example:

A publishing company has just published a new college textbook. Before the company decides the price at which to sell this textbook, it wants to know the average price of all such textbooks in the market.

The research department at the company took a sample of 36 such textbooks and collected information on their prices. This information produced a mean price of \$48.4 for this sample. It is known that the standard deviation of the prices of all such textbooks is \$4.50.

Section 3 – Confidence Interval

Assume that the prices of all such textbooks are normally distributed.

(a) What is the point estimate of the mean price of all such college textbooks?

(b) Construct a 95% confidence interval for the mean price of all such college textbooks.

Section 3 – Confidence Interval

Solution:

From the given information,

$$n = 36, \quad \overline{X} = 48.40, \quad \sigma = 4.50$$

(a) What is the point estimate of the mean price of all such college textbooks?

The point estimate of the mean price of all such college textbooks is \$48.40, that is,

$$\text{Point estimate of } \mu = \overline{X} = \$48.40$$

Section 3 – Confidence Interval

(b) Construct a 95% confidence interval for the mean price of all such college textbooks.

The confidence level is 95% or 0.95 ⇒ $\alpha = 0.05$

The 95% confidence interval for μ is

$$\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 48.40 \pm 1.96 \frac{4.5}{\sqrt{36}} = (46.93, 49.87)$$

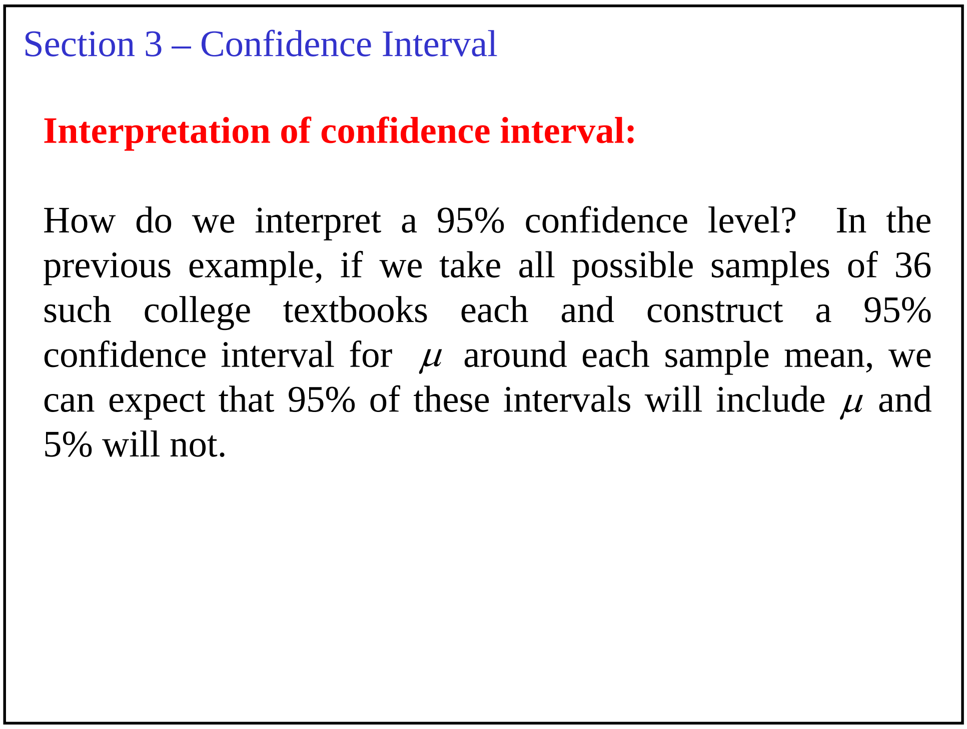
Thus, we are 95% confident that the mean price of all such college textbooks is between $46.93 and $49.87.

Section 3 – Confidence Interval

Note: We cannot say for sure whether the interval $46.93 to $49.87 contains the true population mean or not.

Since μ is a constant, we cannot say that the probability is 0.95 that this interval contains μ because either it contains μ or it does not. Consequently, the probability is either 1 or 0 that this interval contains μ.

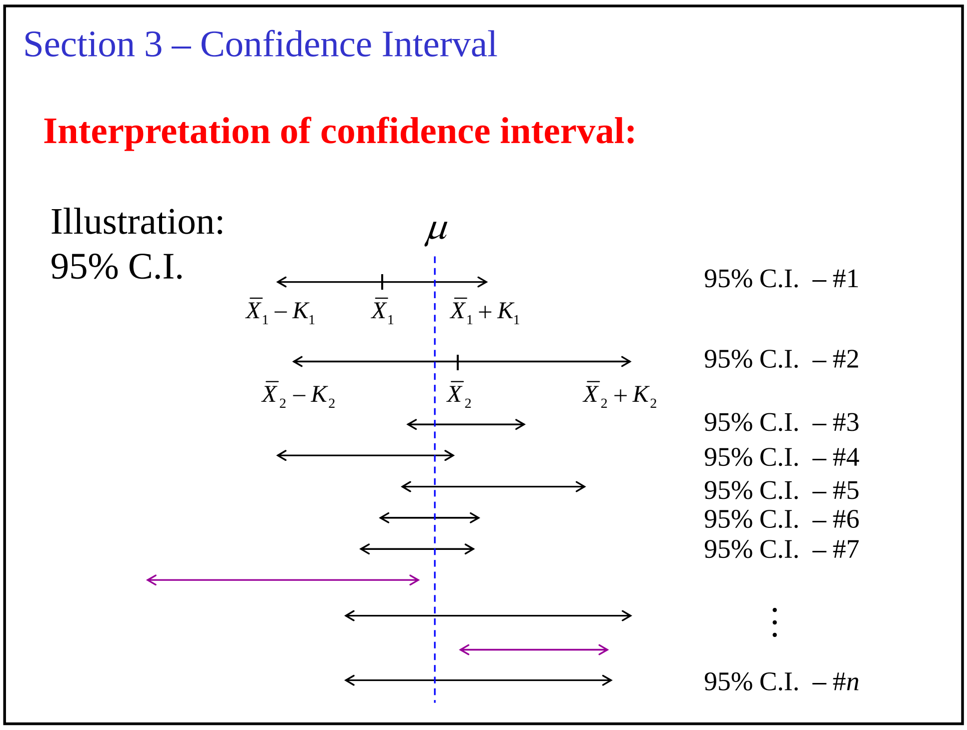
All we can say is that we are 95% confident that the mean price of all such college textbooks between $46.93 and $49.87.



Section 3 – Confidence Interval

Interpretation of confidence interval:

How do we interpret a 95% confidence level? In the previous example, if we take all possible samples of 36 such college textbooks each and construct a 95% confidence interval for μ around each sample mean, we can expect that 95% of these intervals will include μ and 5% will not.



Section 3 – Confidence Interval

Interpretation of confidence interval:

Illustration:
95% C.I.

Section 3 – Confidence Interval

The Width of a Confidence Interval

The width of a confidence interval depends on the size of the maximum error $Z \cdot \sigma_{\overline{X}}$, which depends on the values of Z, σ, and n because $\sigma_{\overline{X}} = \sigma / \sqrt{n}$.

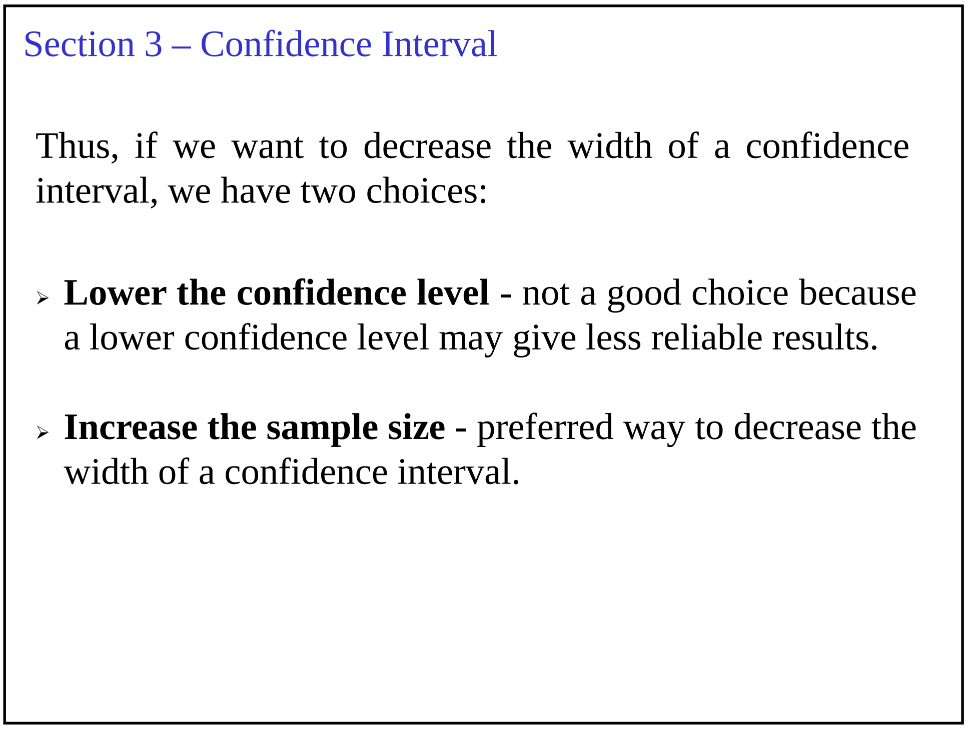
However, the value of σ is not within the control of the investigator. Hence, the width of a confidence interval depends on

(*i*) The value of *Z*
(*ii*) The sample size *n*

Section 3 – Confidence Interval

The value of *Z* which depends on the confidence level
The value of *Z* increases as the confidence level increases, and it decrease as the confidence level decreases. Therefore, the width of a confidence interval increases or decreases with the confidence level.

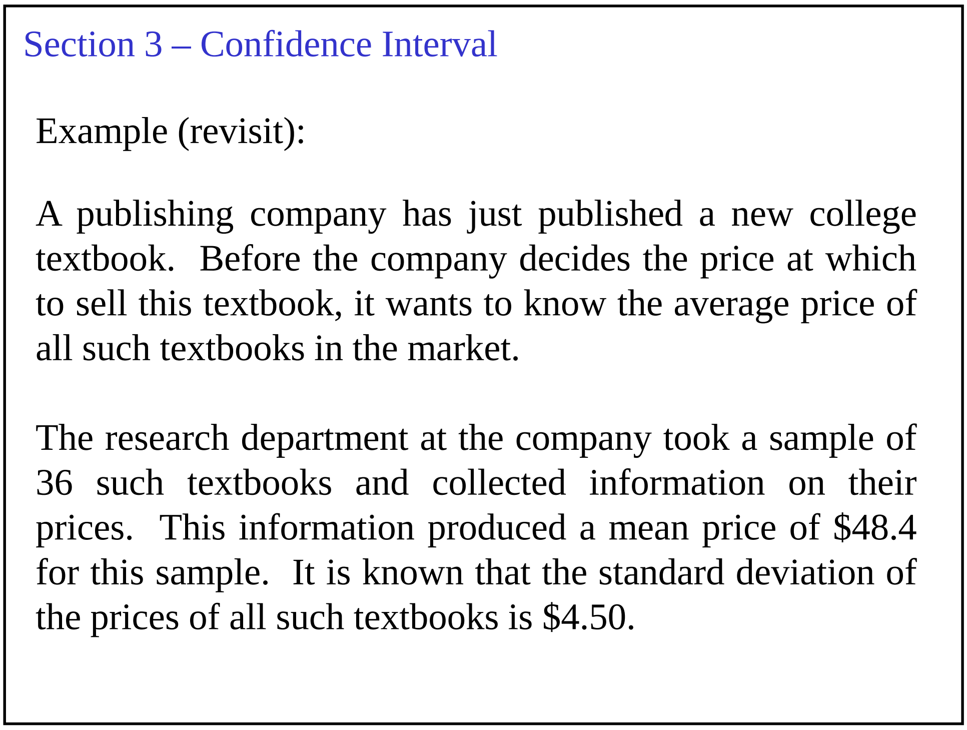
The sample size *n*
For the same value of σ, an increase in *n* decreases the value of $\sigma_{\overline{X}}$, which in turn decreases the size of the maximum error when the confidence level remains unchanged. Therefore, an increase in the sample size decreases the width of the confidence interval.



Section 3 – Confidence Interval

Thus, if we want to decrease the width of a confidence interval, we have two choices:

- **Lower the confidence level -** not a good choice because a lower confidence level may give less reliable results.
- **Increase the sample size -** preferred way to decrease the width of a confidence interval.



Section 3 – Confidence Interval

Example (revisit):

A publishing company has just published a new college textbook. Before the company decides the price at which to sell this textbook, it wants to know the average price of all such textbooks in the market.

The research department at the company took a sample of 36 such textbooks and collected information on their prices. This information produced a mean price of $48.4 for this sample. It is known that the standard deviation of the prices of all such textbooks is $4.50.

Section 3 – Confidence Interval

Assume that the prices of all such textbooks are normally distributed. Construct a 90% confidence interval for the mean price of all such college textbooks.

Solution:

$$\bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 48.40 \pm 1.65\frac{4.5}{\sqrt{36}} = (47.16, 49.64)$$

Comparing this to the 95% confidence interval obtained previously, $(46.93, 49.87)$, it is observed that the width of the confidence interval for a 95% C.I. is wider than the one for a 90% C.I.

Section 3 – Confidence Interval

Example (revisit):

Consider the previous example again. Now suppose the information given in that example is based on a sample size of 160. Further assume that all other information given in that example, construct the 95% confidence level.

Solution:

$$\bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 48.40 \pm 1.96\frac{4.5}{\sqrt{160}} = (47.70, 49.10)$$

Comparing this to the 95% confidence interval obtained previously, $(46.93, 49.87)$, it is observed that the width of the 95% confidence interval for $n = 160$ is smaller than the one for $n = 36$.

Section 3 – Confidence Interval

Interval Estimation of a Population Mean: Unknown Variances

If the sample size is small, the **normal distribution** can still be used to construct a confidence interval for μ if

1.the population from which the sample is drawn is normally distributed, and
2.the population standard deviation σ is known.

Section 3 – Confidence Interval

The *t* distribution is used to make a confidence interval about μ if

1.the population from which the sample is selected is (approximately) normally distributed, and
2.the population standard deviation σ is not known.

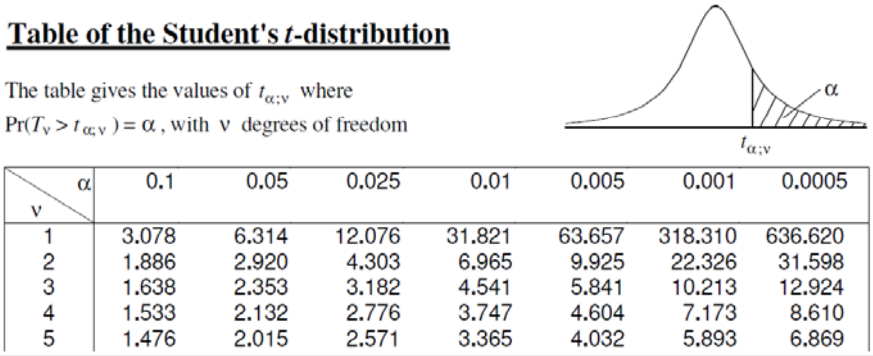


Table of the Student's *t*-distribution

The table gives the values of $t_{\alpha;\nu}$ where
$\Pr(T_\nu > t_{\alpha;\nu}) = \alpha$, with ν degrees of freedom

ν \ α	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.078	6.314	12.076	31.821	63.657	318.310	636.620
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869

Section 3 – Confidence Interval

Verify:

a. $t_{4,0.05} = 2.132$ and $t_{4,0.95} = -2.132$

b. $t_{6,0.005} = 3.707$ and $t_{6,0.995} = -3.707$

c. $P(T > t_{\alpha}) = 0.10$ with $\nu = 22$ → $t_{22,0.1} = 1.321$

d. $P(T < t_{\alpha}) = 0.05$ with $\nu = 16$ → $t_{16,0.95} = -1.746$

e. Given that $T \sim t_5$, $P(T \le 3.365) = 0.99$

f. Given that $T \sim t_8$, $P(-2.306 < T < 2.306) = 0.95$

g. Given that $T \sim t_{26}$, $P(T > -3.435) = 0.999$

Section 3 – Confidence Interval

Confidence Interval for population mean μ using *t* distribution

The $(1-\alpha)100\%$ confidence interval for μ is

$$\boxed{\bar{X} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}}$$

where

$\bar{X}$ is sample mean; s is sample standard deviation; n is the sample size; and $t_{\alpha/2,n-1}$ is obtained from the t distribution table for $n-1$ d.f. and the $(1-\alpha)100\%$ confidence level.

Conditions: Population is approximately normal distributed
σ is not known

Section 3 – Confidence Interval

Example:

Dr. Moore wanted to estimate the mean cholesterol level for all adult males living in London. He took a sample of 25 adult males from London and found that the mean cholesterol level for this sample is 186 with a standard deviation of 12.

Assume that the cholesterol levels for all adult males in London are (approximately) normally distributed. Construct a 95% confidence interval for the population mean.

Section 3 – Confidence Interval

Solution:

From the given information,

$$n = 25, \ \ \bar{X} = 186, \ \ s = 12$$

The confidence level is 95% or 0.95 $\Rightarrow$ $\alpha = 0.05$

Degree of freedom: $25 - 1 = 24$

Area in each tail: $0.05 / 2 = 0.025$

From the *t* distribution table, the value for *t* is $t_{0.025,24} = 2.064$

Section 3 – Confidence Interval

The 95% confidence interval for μ is

$$\bar{X} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} = 186 \pm 2.064\frac{12}{\sqrt{25}} = (181.0464,190.9536)$$

Thus, we can state with 95% confidence that the mean cholesterol level for all adult males living in London lies between 181.05 and 190.95.

Note that $\bar{X} = 186$ is a point estimate of μ in this example.

AMA 1006
Lecture Notes