Chapter 06

Estimation

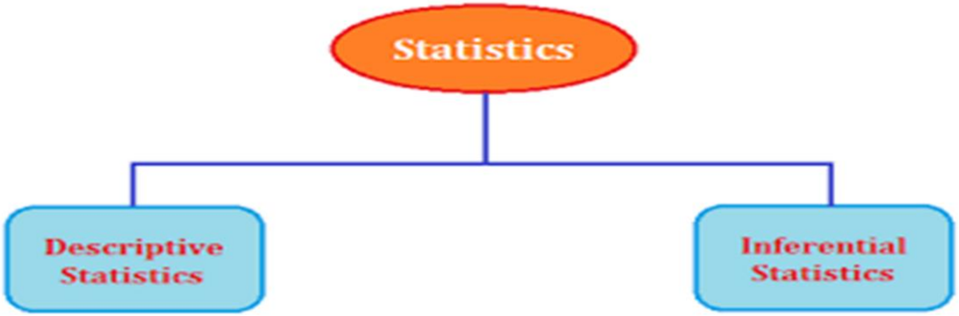
Fundamentals of Biostatistics

Prof. Dr. Moustafa Omar Ahmed Abu-Shawiesh

Professor of Statistics

6.1 Introduction

Chapters 3 through 5 explored the properties of different probability models (probability distributions). In doing so, we always assumed the specific probability distributions were known. In the rest of this text, the problem is that, we have a data set and we want to **infer** the properties of the underlying probability distribution from this data set. In principle, a variety of different probability distributions must at least be explored to see which distribution best **"fits"** the data.

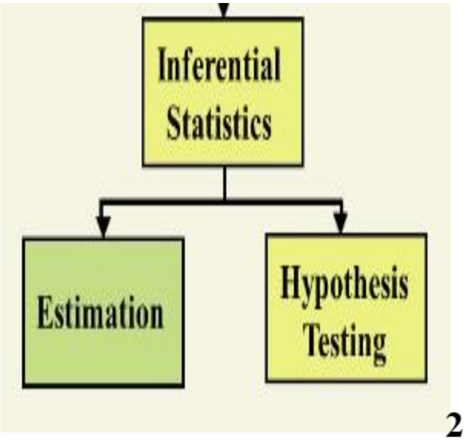


Statistical inference or Inferential statistics

Consists of methods that use sample results to draw conclusions (inferences) in order to help in make decisions or predictions about a population and can be further subdivided into the two main areas (two statistical methods):

(1) Estimation (Point and Interval).

(2) Hypothesis Testing.



Estimation

Is concerned with estimating the values of specific population parameters, like for example: population mean (μ), population standard deviation (σ) and population proportion (p). In this case, we are interested in obtaining specific values or points as estimates of our parameters. These values are often referred to as point estimates. Sometimes we want to specify a range within which the parameter values are likely to fall. If this range is narrow, then we may feel our point estimate is good. This type of problem involves interval estimation.

Hypothesis Testing

Is concerned with testing whether the value of a population parameter is equal to some specific value, like for example:

- $\mu = \mu_0$

  Example

  $\mu = 25$

- $\sigma = \sigma_0$

  Example

  $\sigma = 10$

- $p = p_0$

  Example

  $p = 0.5$

6.2 The Relationship Between Population and Sample

DEFINITION 6.1 **A random sample** is a selection of some members of the population such that each member is independently chosen and has a known nonzero probability of being selected.

EXAMPLE 6.8

Obstetrics Suppose we want to characterize the distribution of birthweights of all liveborn infants born in the United States in 2013. Assume the underlying distribution of birthweight has an expected value (or mean) μ and variance σ^2. Ideally, we wish to estimate μ and σ^2 exactly, based on the entire population of U.S. liveborn infants in 2013. But this task is difficult with such a large group. Instead, we decide to select a random sample of n infants who are *representative* of this large group and use the birthweights $x_1, \ldots, x_n$ from this sample to help us estimate μ and σ^2.

DEFINITION 6.2 **A simple random sample** is a random sample in which each group member has the same probability of being selected.

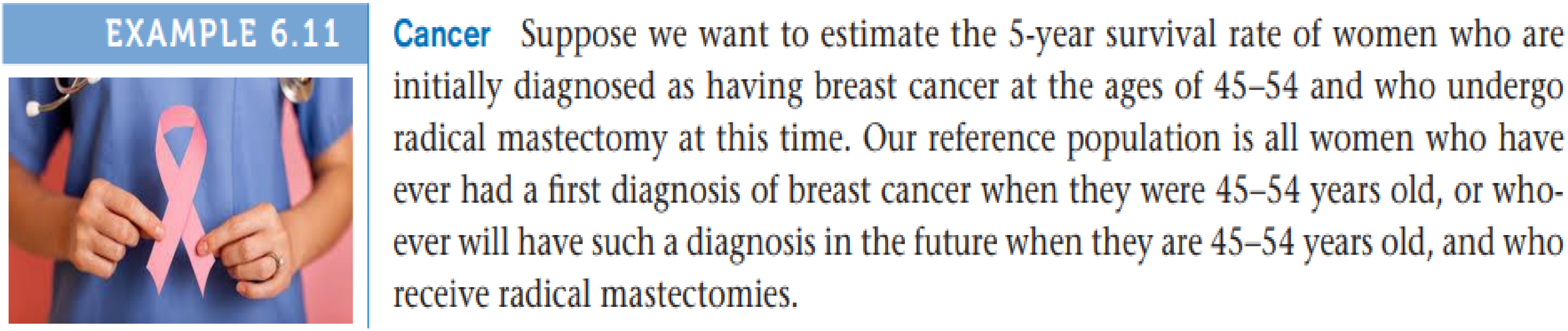
DEFINITION 6.3 The **reference**, **target**, or **study population** is the group we want to study. The random sample is selected from the study population.

Notation: For ease of discussion, we use the abbreviated term "random sample" to denote a simple random sample. Although many samples in practice are random samples, this is not the only type of sample used in practice. A popular alternative design is cluster sampling. In this course, we assume that all samples are random samples from a reference population.

Notation: In general, the reference population can be divided into two types as follows:

a. Finite, well defined and can be enumerated.
b. Effectively infinite, not well defined and can not be enumerated.

Important: In this course, we assume all reference populations discussed are effectively infinite, although, many are actually very large but finite (*see examples 6.8 and 6.10*).



EXAMPLE 6.11

Cancer Suppose we want to estimate the 5-year survival rate of women who are initially diagnosed as having breast cancer at the ages of 45–54 and who undergo radical mastectomy at this time. Our reference population is all women who have ever had a first diagnosis of breast cancer when they were 45–54 years old, or whoever will have such a diagnosis in the future when they are 45–54 years old, and who receive radical mastectomies.

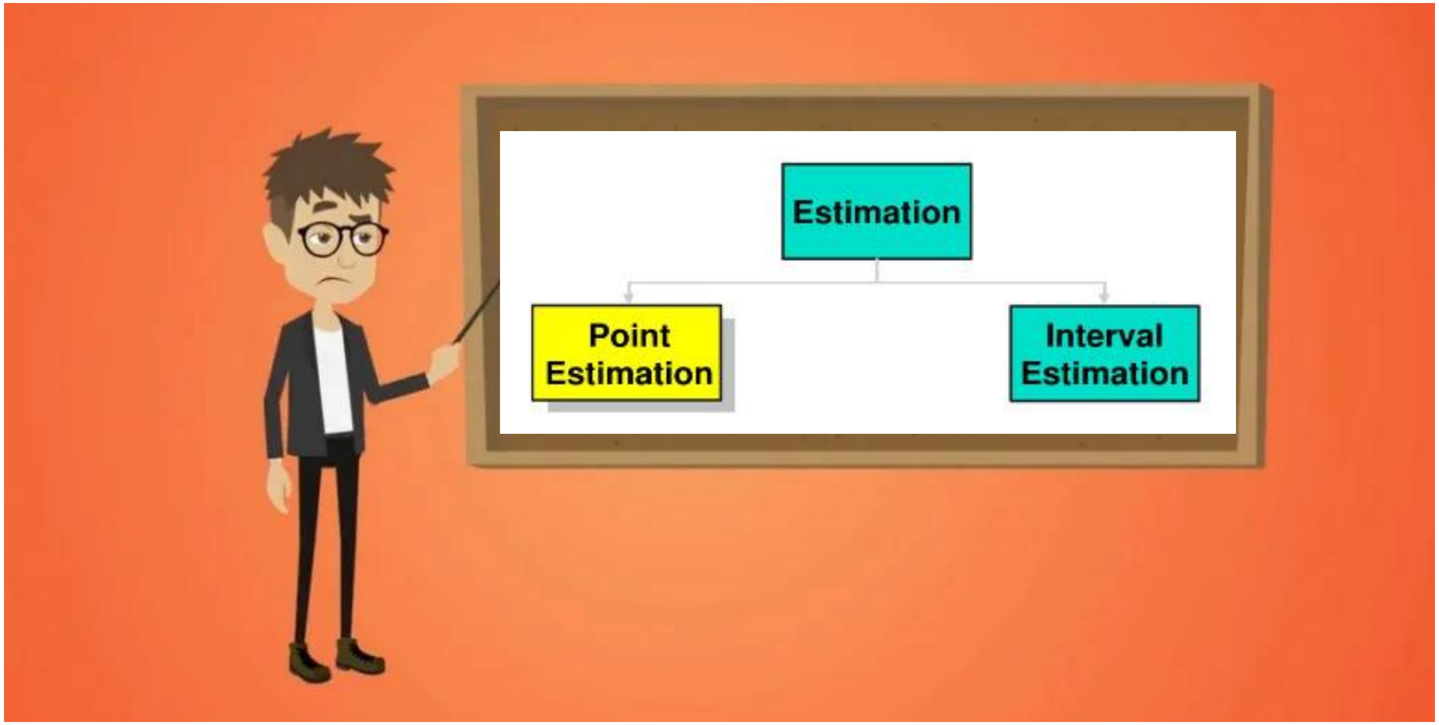
Conclusion: This population is effectively infinite. It cannot be formally enumerated, so a truly random sample cannot be selected from it. However, we again assume the sample we have selected behaves as if it were a random sample.

6.5 Estimation of the Mean of a Distribution

In this section, we will discuss and focus on the estimation method to answer the following question:

How is a specific random sample $X_1, X_2, \ldots, X_n$ of a random variable X can be used to estimate the population mean (μ) of the underlying distribution?

Two types of the estimation method will be used to do that, namely, the point estimation and the interval estimation.



6.5.1 Point Estimation

Each sample has its own sample mean ($\overline{X}$), and the distribution of the sample mean ($\overline{X}$) is known as the sampling distribution. Therefore the sample mean ($\overline{X}$) given by $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ is a natural estimator used for estimating the population mean (μ).

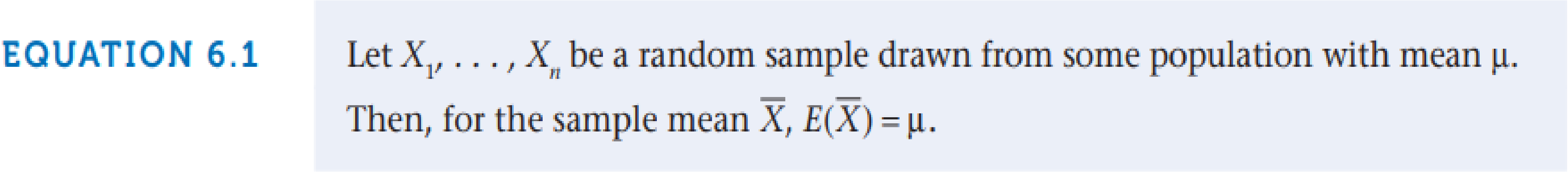
Definition: Suppose all samples of size n are selected from a population with mean μ and standard deviation σ. For each sample, the sample mean $\overline{X}$ is recorded. The probability distribution of these sample means is called **the sampling distribution of the sample means**.

Question

What properties of $\overline{X}$ make it a desirable estimator of population mean (μ)?

Answer

(1) The sample mean ($\overline{X}$) is an unbiased estimator of the population mean (μ).



EQUATION 6.1

Let $X_1, \ldots, X_n$ be a random sample drawn from some population with mean μ. Then, for the sample mean $\overline{X}$, $E(\overline{X}) = \mu$.

Note that Equation 6.1 holds for any population regardless of its underlying distribution.

(2) The sample mean ($\overline{\mathrm{X}}$) is called the minimum variance unbiased estimator (MVUE) of μ, that is, if the underlying distribution of the population is normal, then it can be shown that the unbiased estimator with the smallest variance is given by the sample mean ($\overline{\mathrm{X}}$).

Standard Error of the Mean

From Equation 6.1 we see that $\overline{\mathrm{X}}$ is an unbiased estimator of μ for any sample size n. Why then is it preferable to estimate parameters from large samples rather than from small ones? The intuitive reason is that the larger the sample size, the more precise an estimator $\overline{\mathrm{X}}$ is.

Important Results: Let $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_\mathrm{n}$ be a random sample of size (n) from a population with underlying mean μ and variance σ^2, then we have the following two results:

(a) The variance of the sample mean ($\overline{\mathrm{X}}$) is given by Var($\overline{\mathrm{X}}$)$= \sigma^2/n$.

(b) The standard deviation of the sample mean ($\overline{\mathrm{X}}$) is given by SD($\overline{\mathrm{X}}$)$= \sigma/\sqrt{n}$ and is referred to as the standard error of the sample mean or the standard error.

In practice, the population variance σ^2 is rarely known, then a reasonable estimator for the population variance σ^2 is the sample variance S^2, which leads to the following definition:

DEFINITION 6.12 The **standard error of the mean (sem)**, or the **standard error (se)**, is given by $\sigma/\sqrt{n}$ and is estimated by $s/\sqrt{n}$. The standard error represents the estimated standard deviation obtained from a set of sample means from repeated samples of size n from a population with underlying variance σ^2.

Note that the standard error is not the standard deviation of an individual observations $X_1, X_2, \ldots, X_n$ but rather of the sample mean ($\overline{X}$).

EXAMPLE 6.23

Obstetrics Compute the standard error of the mean for the third sample of birthweights in Table 6.3 (p. 161).

Solution: The standard error of the mean is given by

$$s/\sqrt{n} = 22.44/\sqrt{10} = 7.09$$

The observations for the third sample of infants birthweights (oz) given in Table 6.3 Page 161 as follows:

97, 125, 62, 120, 132, 135, 118, 137, 126, 118

$n = 10$, $\overline{X} = 117$, $S = 22.44$

Notations

- The larger sample should provide a more precise estimate of μ.
- The precision of an estimate is also affected by the underlying variance σ^2 from the population of individual observations. However, σ^2 can sometimes be affected by experimental technique.
- The standard error is a quantitative measure of the variability of sample means obtained from repeated random samples of size n drawn from the same population.
- The standard error is directly proportional to both $1/n$ and to the population standard deviation σ of individual observations $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_\mathrm{n}$.

Summary

Let $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_\mathrm{n}$ be a random sample of size (n) from a population with mean (μ) and variance (σ^2) then the following important results will be needed to derive and identify the sampling distribution for the sample mean ($\bar{\mathrm{X}}$):

Result (1): The mean for the sample mean ($\bar{\mathrm{X}}$) is given by $\mu_{\bar{\mathrm{X}}}$ = E($\bar{\mathrm{X}}$) = μ.

Result (2): The variance for the sample mean ($\bar{\mathrm{X}}$) is given by $\sigma^2_{\bar{\mathrm{X}}}$ = Var($\bar{\mathrm{X}}$) $= \sigma^2/n$.

Result (3): The standard deviation (standard error) for the sample mean ($\bar{\mathrm{X}}$) is given by $\sigma_{\bar{\mathrm{X}}}$ = SD($\bar{\mathrm{X}}$)$= \sigma/\sqrt{n}$.

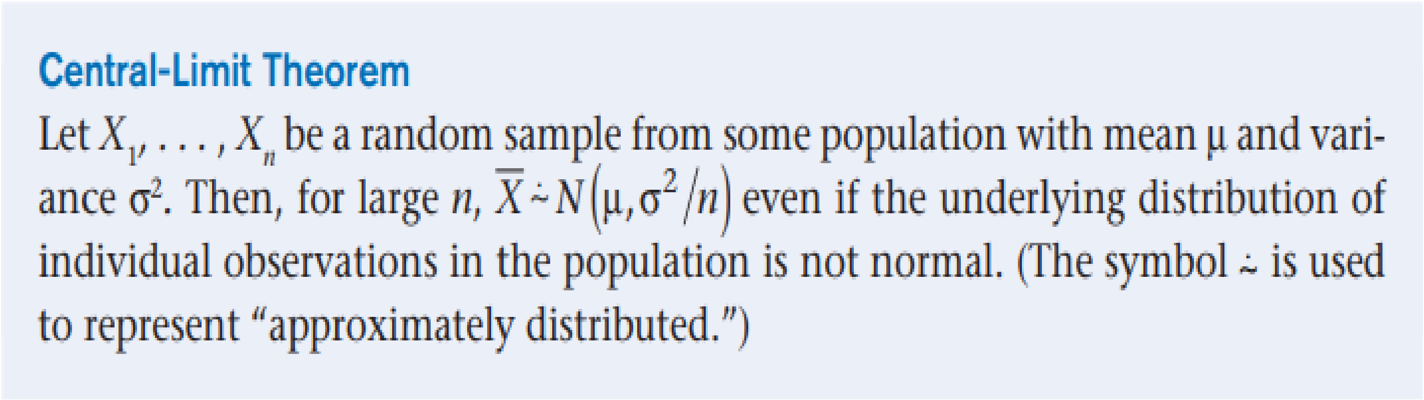
Self-Reading: Study Example 6.24 page 170 by your-self.

The Central-Limit Theorem (CLT)

This theorem is very important because many of the distributions encountered in practice are not normal. In such cases the central limit theorem (CLT) can often be applied.

- If the underlying distribution is normal, then it can be shown that the sampling distribution of the sample mean ($\bar{X}$) is normal distribution with mean μ and variance $\frac{\sigma^2}{n}$. In other words, $\bar{X} \sim N(\mu\,, \frac{\sigma^2}{n})$.
- If the underlying distribution is not normal, we would still like to make some statement about the sampling distribution of the sample mean ($\bar{X}$). This statement is given by the central limit theorem (CLT):

EQUATION 6.3



Central-Limit Theorem

Let $X_1, \ldots, X_n$ be a random sample from some population with mean μ and variance σ^2. Then, for large n, $\bar{X} \dot{\sim} N\left(\mu, \sigma^2/n\right)$ even if the underlying distribution of individual observations in the population is not normal. (The symbol $\dot{\sim}$ is used to represent "approximately distributed.")

Notation

- The sample size (n) is small if $n < 30$.
- The sample size (n) is large if $n \geq 30$.

Note that the central limit theorem (CLT) lets us perform statistical inference based on the approximate normality of the sample mean ($\bar{X}$) despite the non-normality of the distribution of individual observations.

Important Rule

The sample mean ($\bar{X}$) can be converted to a Z-Score by using the following formula:

$$\boxed{Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)}$$

where

$\bar{X}$ = the sample mean
μ = the population mean
σ = the population standard deviation
n = the sample size

Example

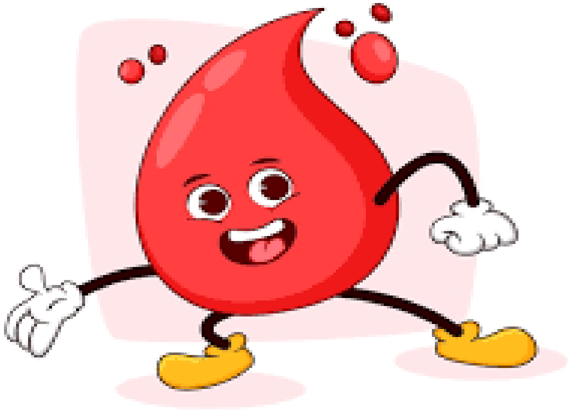
Suppose that for a certain large group of individuals in Jordan, the mean hemoglobin level in the blood is 21 grams per milliliter (g/ml) and the standard deviation is 2 g/ml. If a random sample of size *n* = 25 individuals is selected, what is the probability that the sample mean will be greater than 21.3 g/ml assuming that the hemoglobin level in the blood is normally distributed?

Solution

We have X = hemoglobin level in the blood$\sim N(21, 4)$, then the sampling distribution for the sample mean $(\overline{\mathrm{X}})$ is given as follows:

$\bar{X} \sim N(\mu_{\bar{X}} = 21, \sigma^2_{\overline{\mathrm{X}}} = \frac{(2)^2}{25})$ ➡ $N(21\ , 0.16)$ (Exactly) then $\sigma_{\overline{\mathrm{X}}} = \sqrt{0.16} = 0.4$

Then we need to find $\mathrm{P}(\overline{\mathrm{X}} > 21.3)$ as follows:



$$\begin{aligned} \mathrm{P}(\overline{\mathrm{X}} > 21.3) &= 1 - \mathrm{P}(\overline{\mathrm{X}} \leq 21.3) \\ &= 1 - \Phi\left(\frac{21.3\ - 21}{0.4}\right) \\ &= 1 - \Phi(0.75) \end{aligned}$$

Refer to Table 3 in the Appendix and obtain:

$$\begin{aligned} &= 1 - 0.7734 \\ &= \ 0.2266 \end{aligned}$$

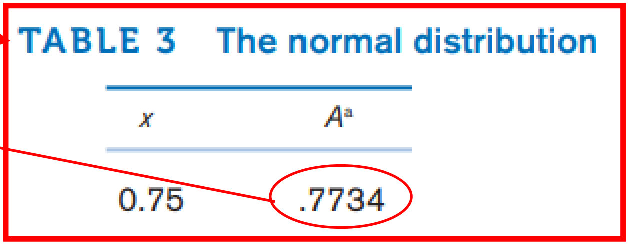


TABLE 3 The normal distribution

x	A^a
0.75	.7734

Conclusion

Thus 22.66% of the samples of size 25 would be expected to have mean of hemoglobin levels in the blood greater than 21.3 g/ml .

EXAMPLE 6.27

Obstetrics Compute the probability that the mean birthweight from a sample of 10 infants from the Boston City Hospital population [redacted] will fall between 98.0 and 126.0 oz (i.e., $98 \le \bar{X} < 126$) if the mean birthweight for the 1000 birthweights from the Boston City Hospital population is 112.0 oz with a standard deviation of 20.6 oz.

Assume that the infants birthweights (X) from the Boston Hospital is normally distributed?

Solution: Since the distribution is normal then the sample mean $(\bar{X})$ follows a normal distribution with mean $\mu_{\bar{X}} = \mu = 112.0$ and standard deviation $\sigma_{\bar{X}} = \sigma/\sqrt{n} = 20.6/\sqrt{10} = 6.51$ that is $\bar{X} \sim N(112\,,(6.51)^2)$ Exactly

$$Pr\left(98.0 \le \bar{X} < 126.0\right) = \Phi\left(\frac{126.0-112.0}{6.51}\right) - \Phi\left(\frac{98.0-112.0}{6.51}\right)$$

$$= \Phi(2.15) - \Phi(-2.15)$$

$$= \Phi(2.15) - [1-\Phi(2.15)] = 2\Phi(2.15) - 1$$

Refer to Table 3 in the Appendix and obtain:

$$Pr\left(98.0 \le \bar{X} < 126.0\right) = 2(.9842) - 1.0 = .968$$

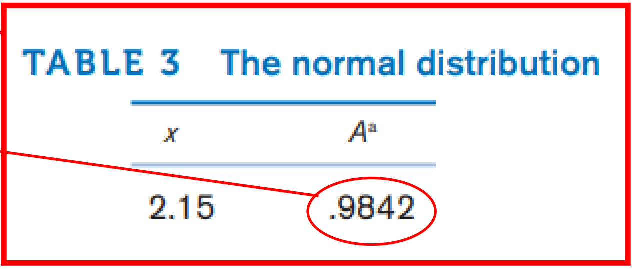


TABLE 3 The normal distribution

x	A^a
2.15	.9842

Conclusion

Thus 96.8% of the samples of size 10 would be expected to have mean birthweights between 98 and 126 oz.

Exercise

Find the value of the following probabilities:

(1) $P(\bar{X} > 98) = 1 - P(\bar{X} \le 98) = 1 - \Phi\left(\frac{98 - 112}{6.51}\right) = 1 - \Phi(-2.15) = 1 - (1 - \Phi(2.15)) = 0.9842$

(2) $P(\bar{X} \le 98) = \Phi\left(\frac{98 - 112}{6.51}\right) = \Phi(2.15) = 0.9842$ AND (3) $P(\bar{X} = 98) = 0$

Exercise

Suppose that at a large Dairy Production Company in Jordan, the mean age of employees is 36.2 year, and the standard deviation is 3.7 year. Assume that the variable is normally distributed, then answer the following:

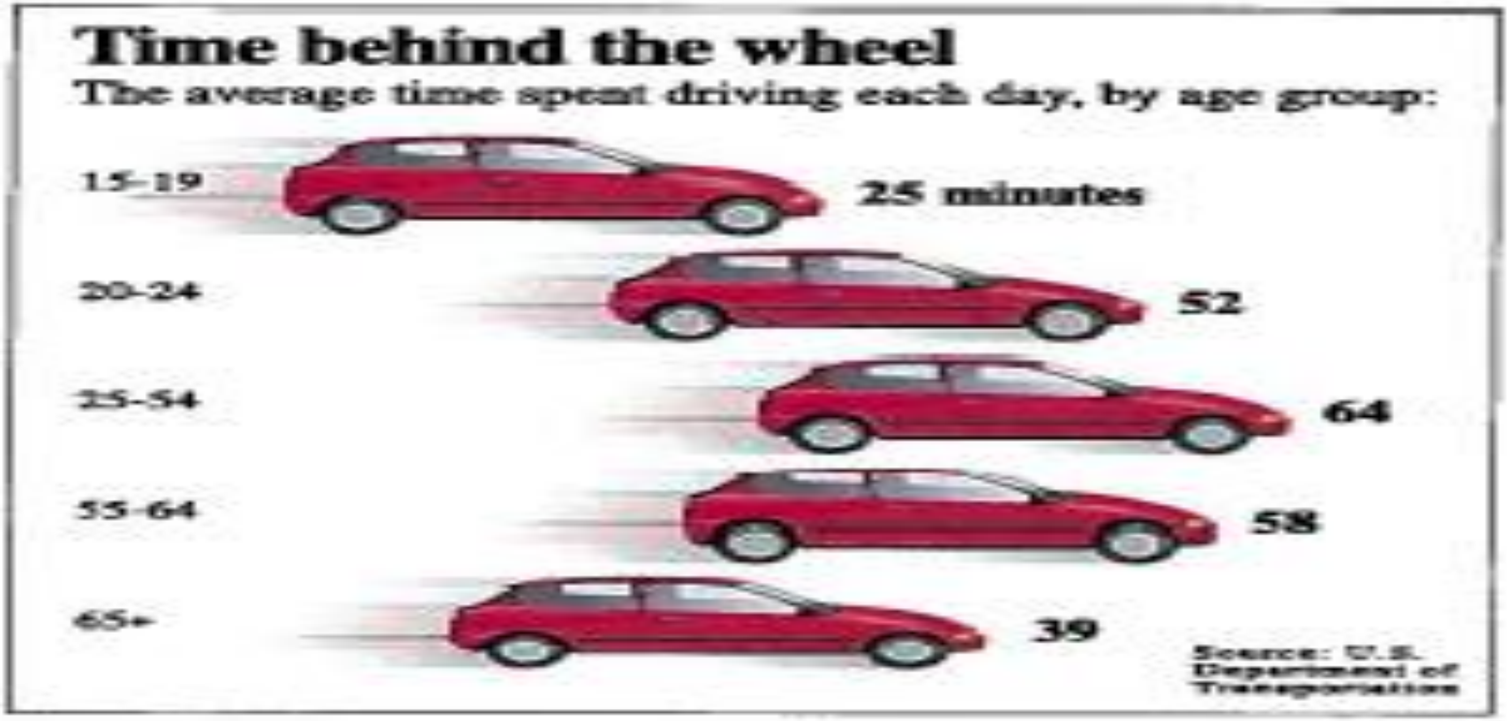
(a) If an employee from the company is randomly selected, find the probability that his/her age will be between 36 and 37.5 year?

Answer: 0.1567

(b) If a random sample of 15 employees is selected, find the probability that the mean age of the employees in the sample will be between 36 and 37.5 year? Answer: 0.4963

Example

The graph below shows the length of time people spend driving each day. You randomly select 50 drivers from age group 15 to 19. What is the probability that the sample mean ($\overline{\mathrm{X}}$) of time they spend driving each day is between 24.7 and 25.5 minutes? Assuming that μ = 25 minutes and σ = 1.5 minutes?



Solution

From the Central Limit Theorem (sample size $n = 50$ is greater than 30), then the sampling distribution of the sample mean ($\overline{\mathrm{X}}$) is approximately normal with mean and standard deviation given as follows:

$$\mu_{\bar{x}} = \mu = 25 \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1.5}{\sqrt{50}} \approx 0.21213$$

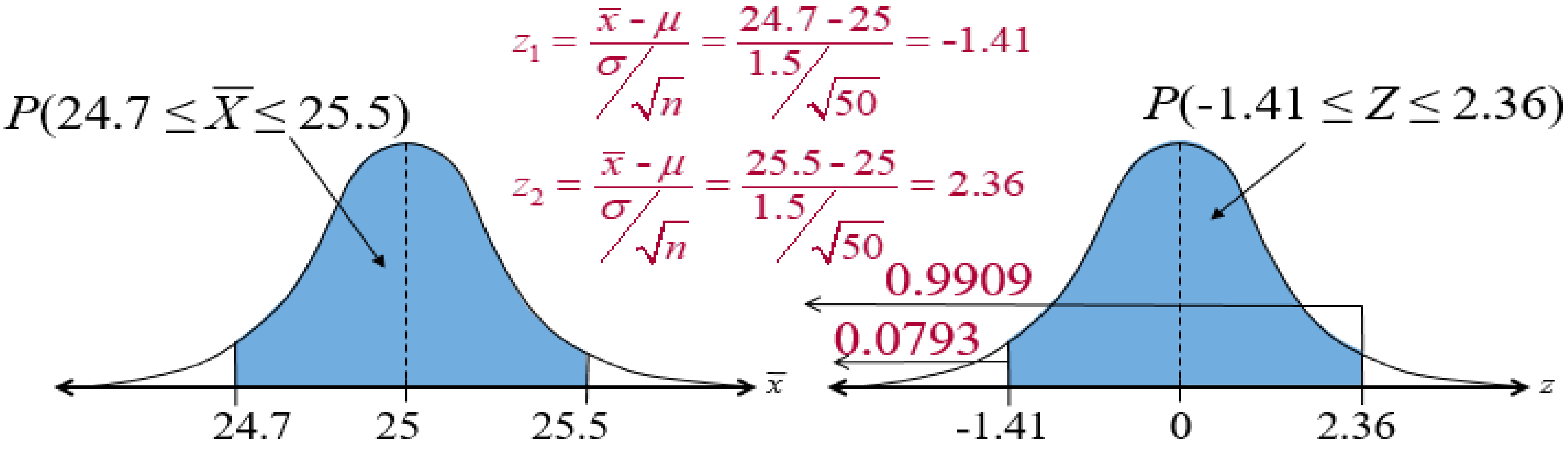
Then the probability that the mean time they spend driving each day is between 24.7 and 25.5 minutes can be calculated as follows:

Normal Distribution

μ(xbar) = 25 ; σ(xbar) = 0.21213

Standard Normal Distribution

μ = 0 ; σ = 1



$$P(24.7 \le \overline{X} \le 25.5)$$

$$z_1 = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{24.7 - 25}{1.5 / \sqrt{50}} = -1.41$$

$$z_2 = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{25.5 - 25}{1.5 / \sqrt{50}} = 2.36$$

$$P(-1.41 \le Z \le 2.36)$$

$P(24.7 \le \overline{X} \le 25.5)$

$= P(-1.41 \le Z \le 2.36)$

$= \Phi(2.36) - \Phi(-1.41)$

$= \Phi(2.36) - [1 - \Phi(1.41)] = \Phi(2.36) - 1 + \Phi(1.41)] = 0.9909 - 1 + 0.9207 = 0.9116$

Refer to Table 3 in the Appendix

Conclusion

Thus, if the central-limit theorem (CLT) holds, then 91.16% of the all samples of size $n = 50$ for drivers from age group 15 to 19 would be expected to have mean time spend driving each day between 24.7 and 25.5 minutes. This confirming that the central-limit theorem holds approximately for averages from samples of size $n = 50$ drawn from this population.

Exercise

In a recent study reported October 29, 2023 on the Jordan University Hospital (JUH) , the mean age of physical therapy sessions patients is 34 years with a standard deviation of 15 years. If a random sample of size $n = 100$ patient is taken from those who are physical therapy sessions users in JUH, then answer the following:

(a) What is the sampling distribution of the sample mean ($\overline{\mathrm{X}}$)?

Answer: $\overline{\mathrm{X}} \dot{\sim} N(34\,, \frac{(15)^2}{100})$ that is $\overline{\mathrm{X}} \dot{\sim} N(34\,, 2.25)$.

(b) Find the probability that the sample mean ($\overline{\mathrm{X}}$) is:

(i) more than 30 years?

Answer: 0.9962

(ii) exactly 30 years?

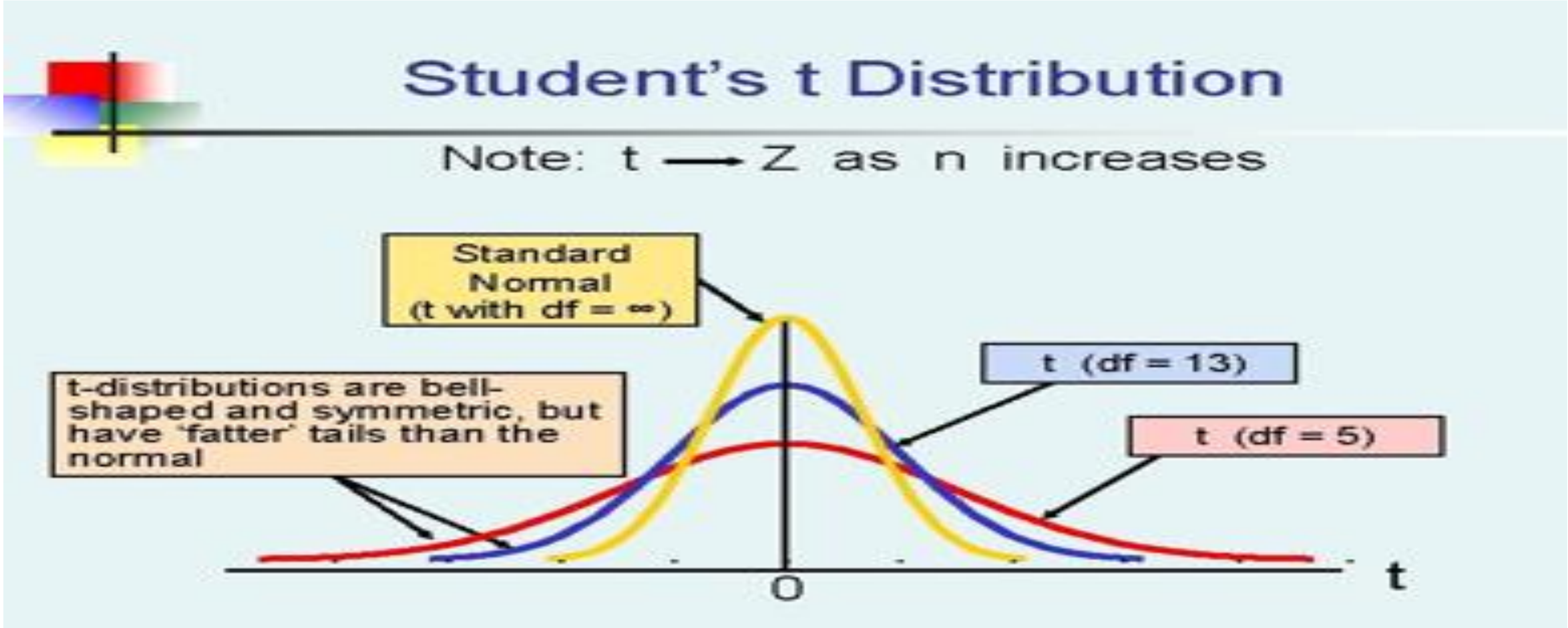
Answer: 0

(iii) Between 30 and 38 years?

Answer: 0.9924

The t-Distribution (Student's t-Distribution)

The t-distribution (Student's t-distribution) is a theoretical probability distribution. The t-distribution is continuous, unimodal (bell-shaped), symmetrical about 0 and similar to the standard normal distribution N(0, 1) but flatter and shorter than a normal distribution. It differs from the standard normal curve, however, in that it has an additional parameter called the degrees of freedom which changes its shape as shown in the figure below:



Notation

The derivation of the t-distribution was published in 1908 by William Sealy Gosset. His work was published under the pseudonym "Student".

Mean and Variance

The mean and the variance for the t-distribution are as follows:

- The mean of the distribution is equal to 0 .
- The variance is equal to v / (v - 2), where v is the degrees of freedom and $v \geq 2$.
- The variance is always greater than 1, although it is close to 1 when there are many degrees of freedom. With infinite degrees of freedom, the t-distribution is the same as the standard normal distribution.

Note that (*Important*)

The conditions for using the t-distribution are:

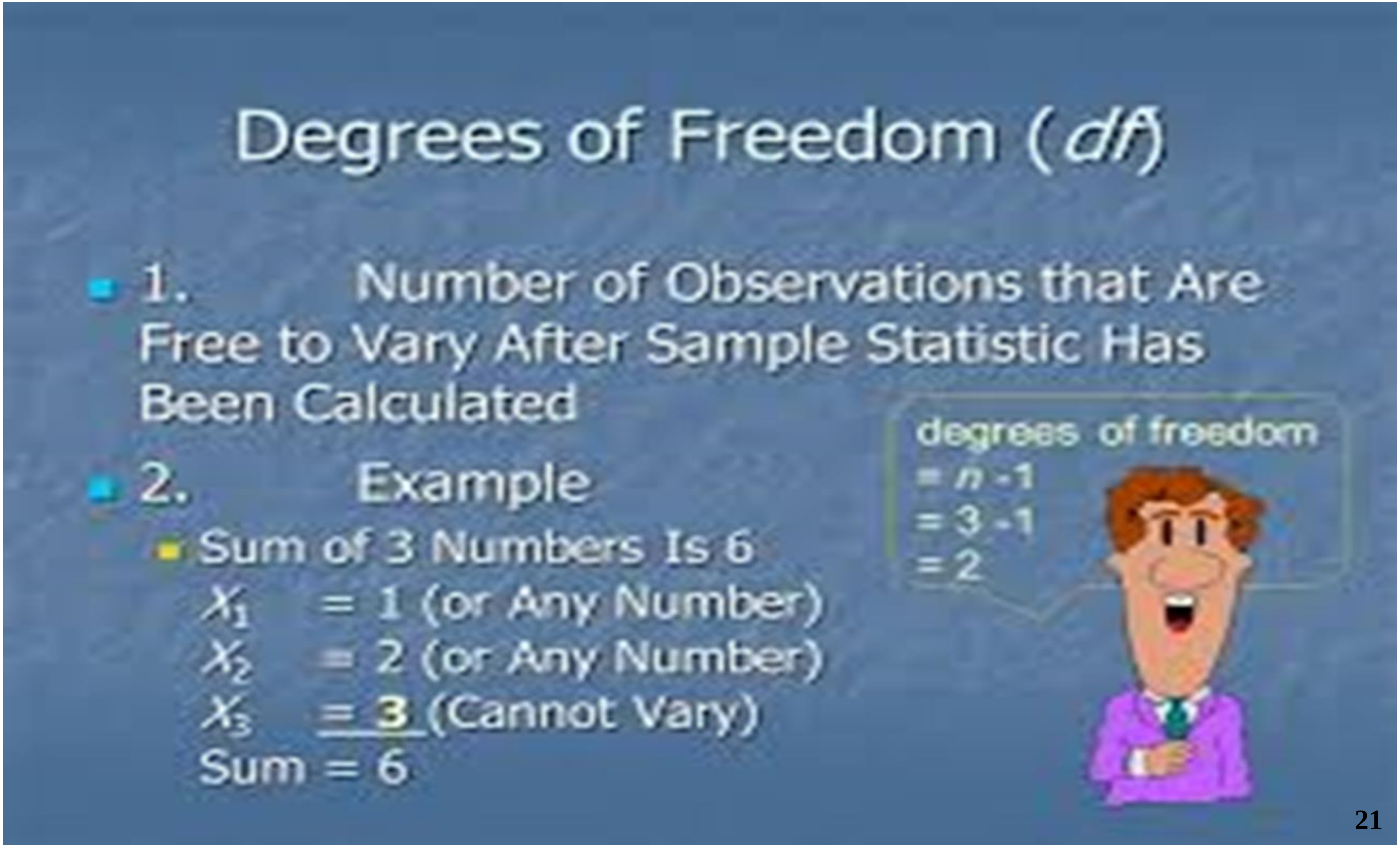
(1) The random sample $X_1, X_2, \dots, X_n$ is taken from $N(\mu, \sigma^2)$.

(2) The population standard deviation $(\sigma = \sqrt{\sigma^2})$ is unknown

(3) The sample size n is small $(n < 30)$.

Notation

Because σ (population standard deviation) is unknown, it is reasonable to estimate σ by using the sample standard deviation $S = \sqrt{S^2}$.



Degrees of Freedom (*df*)

- 1. Number of Observations that Are Free to Vary After Sample Statistic Has Been Calculated
- 2. Example
  - Sum of 3 Numbers Is 6

X_1 = 1 (or Any Number)

X_2 = 2 (or Any Number)

X_3 $\underline{= 3}$ (Cannot Vary)

Sum = 6

degrees of freedom
$= n - 1$
$= 3 - 1$
$= 2$

The *t*-Distribution Properties

1. The total area under the t-distribution curve is equal to 1.

2. The variable t ranges from $-\infty$ to $+\infty$; $-\infty < t < +\infty$.

3. It has a mean of 0 $(\mu = 0)$.

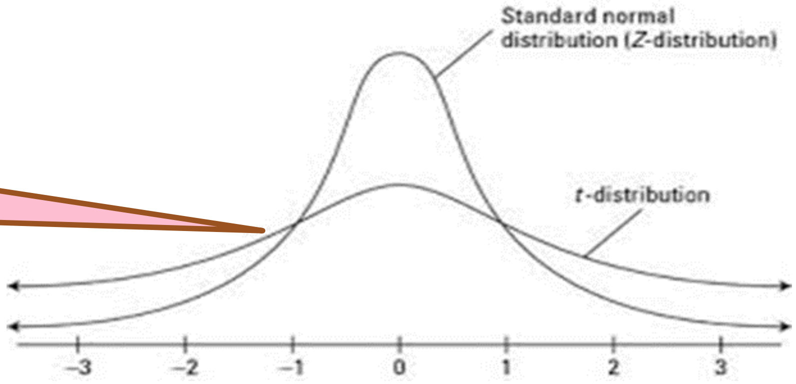
4. It is symmetrical about the mean (t = 0).

5. In general it has a variance greater than 1, but the variance approaches 1 as the sample size n becomes larger.

6. Compared to the standard normal distribution, the t distribution is less peaked in the center and has higher tails.

7. The t distribution approaches the standard normal distribution as n increases.

8. Degree of Freedom = n - 1.



A comparison between N(0, 1) and t-distribution Curves

Theorem (1)

If $X_1, X_2, \dots, X_n \sim N(\mu, \sigma^2)$

then

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{for } n < 30 \ \text{ or } \ n \geq 30 .$$

Sampling Distribution
of the Sample Mean ($\bar{X}$)

Note that

In Theorem (1), the population variance σ^2 is known, or the population standard deviation $\sigma = \sqrt{\sigma^2}$ is known.

Remark

Theorem (1) gives that if the population is normally distributed then we have:

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,\ 1)$$

EQUATION 6.5

If $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$ where the size of the random sample is small ($n < 30$) and the population standard deviation ($\sigma = \sqrt{\sigma^2}$) is unknown, then:

(1) Replace the population standard deviation (σ) by the sample standard deviation (S) where $S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$.

(2) $t = \frac{(\bar{X} - \mu)}{S/\sqrt{n}}$ is distributed as a t distribution with (n – 1) df.

Sampling Distribution of the Sample Mean ($\bar{X}$)

Notation

Once again, Student's t distribution is not a unique distribution but is a family of distributions indexed by the degrees of freedom d. The t distribution with d degrees of freedom is sometimes referred to as the t_d distribution.

Remark

In EQUATION 6.5, if the sample size (n) is large, that is, $n \geq 30$ then

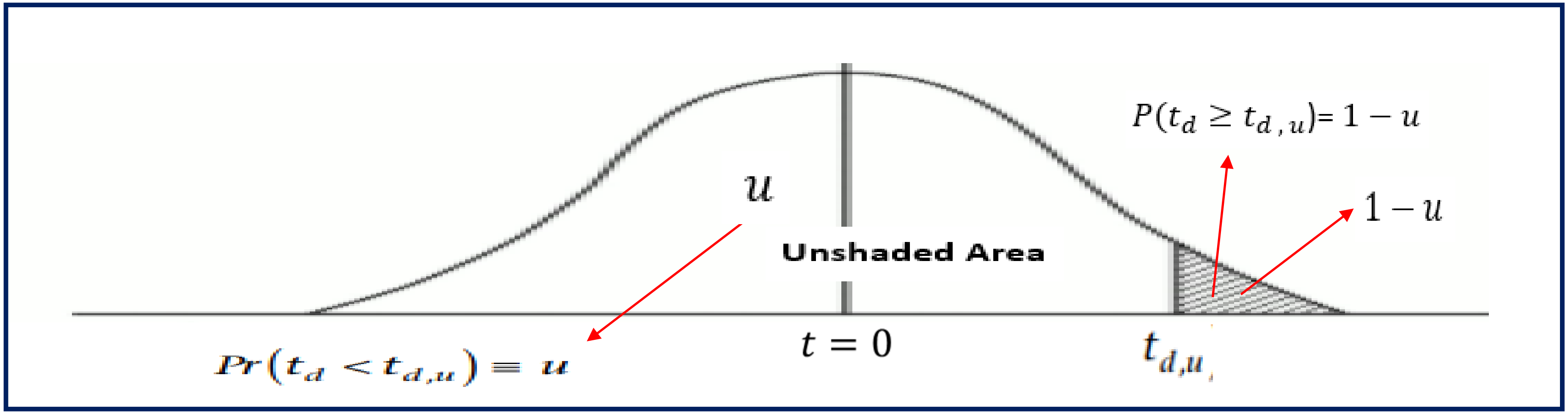
$$Z = \frac{\bar{X} - \mu}{S/\sqrt{n}} \text{ is approximately N(0, 1)}$$

DEFINITION 6.13 The 100 × *u*th percentile of a *t* distribution with *d* degrees of freedom is denoted by $t_{d,u}$, that is,

$$Pr\left(t_d < t_{d,u}\right) \equiv u$$

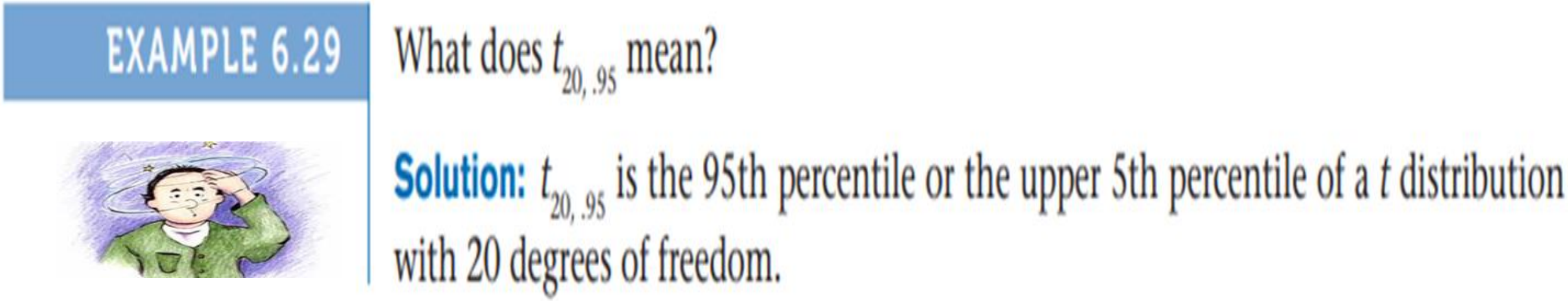
Notation

The value $t_{d\,,u} = t(d,u) = t(n-1,u)$ can be defined as the value of t having area u to its left side $P(t_d < t_{d\,,u})= u$ (Unshaded Area) or the value of t having area $1-u$ to its right side $P(t_d \geq t_{d\,,u})= 1-u$ (Shaded Area) as shown below:


$P(t_d \geq t_{d\,,u})= 1-u$
u
$1-u$
Unshaded Area
$t = 0$
$t_{d,u}$
$Pr\left(t_d < t_{d,u}\right) = u$

Rule

$$P(t_d < t_{d\,,u}) + P(t_d \geq t_{d\,,u}) = u + (1-u) = 1$$



EXAMPLE 6.29 What does $t_{20,.95}$ mean?

Solution: $t_{20,.95}$ is the 95th percentile or the upper 5th percentile of a t distribution with 20 degrees of freedom.

Important Notation

The difference between the t-distribution and the standard normal distribution is greatest for small values of n $(n < 30)$.

Question: How to find the percentage points $(t_{d\,,\,u})$, (area or probability), using the Student's t-distribution?

Answer

Table 5 (Page 879) in the Appendix gives the percentage points $(t_{d\,,\,u})$, *that is, the uth percentile of a t-distribution with d degrees of freedom*, of the t-distribution for various degrees of freedom (d) as follows:

- The degrees of freedom (d) are given in the first column of the table.
- The percentiles $(t_{d\,,\,u})$ are given across the first row.
- The uth percentile of a t-distribution with d degrees of freedom is found by reading across the row marked d and reading down the column marked u.

$P(t_d < t_{d\,,u})$

the percentile (u)

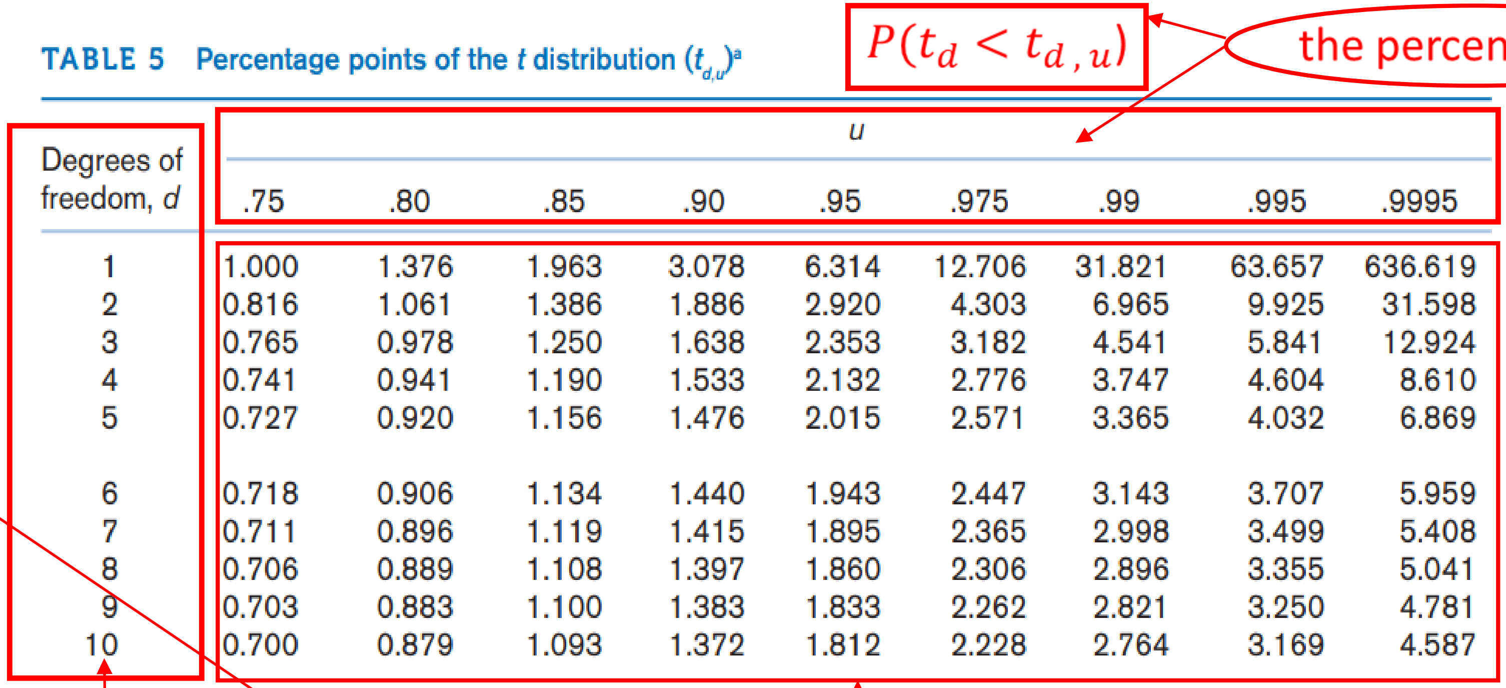


TABLE 5 Percentage points of the *t* distribution ($t_{d,u}$)[a]

Degrees of freedom, *d*	*u* .75	.80	.85	.90	.95	.975	.99	.995	.9995
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.598
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587

$(n-1)$

degrees of freedom (d)

uth percentile

$(t_{d\,,u})$

EXAMPLE 6.30

Find the upper $5th$ percentile (lower $95th$ percentile) of a t-distribution with 23 df?

Solution

We want to find $t_{23\,,\,0.95}$ which is given by:

$$t_{23\,,\,0.95} = t(23\,,0.95) = P(t_{23} < t_{23\,,\,0.95}) = 0.95 \quad \text{or} \quad P(t_{23} \geq t_{23\,,\,0.95}) = 0.05$$

The value $t_{23\,,\,0.95}$ is given in row 23 and column 0.95 of Appendix Table 5 and is equal to $t_{23\,,\,0.95} = 1.714$ implies that $P(t_{23} < 1.714) = 0.95$.

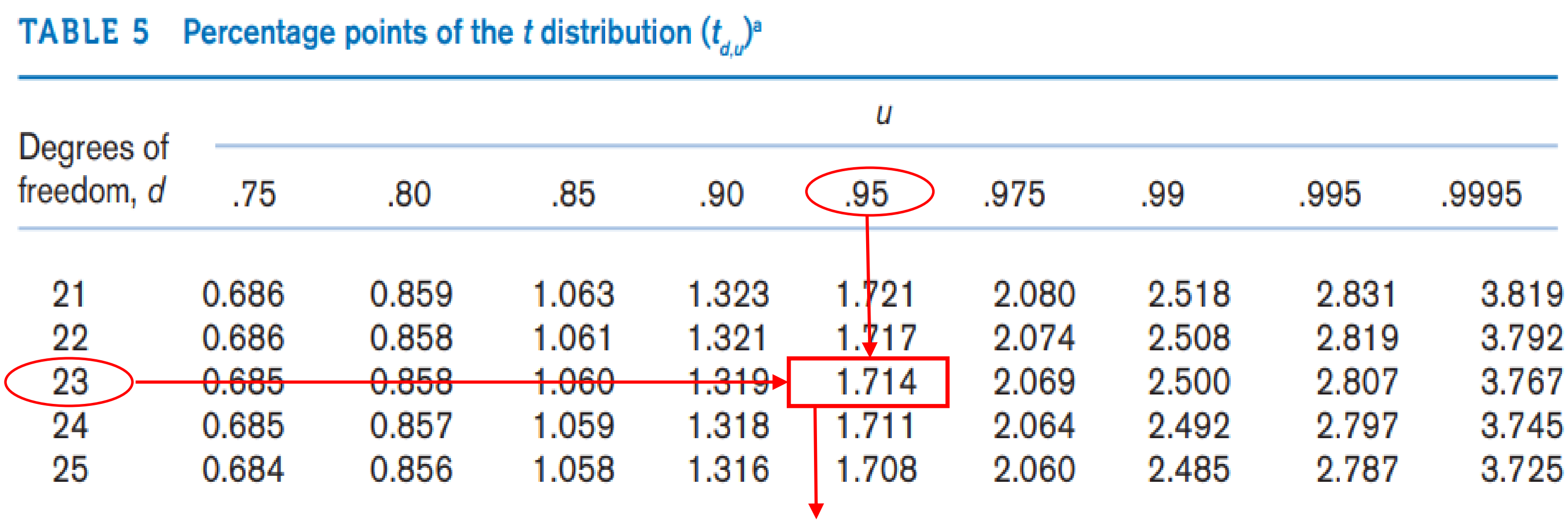


TABLE 5 Percentage points of the t distribution $(t_{d,u})^a$

Degrees of freedom, d	u								
	.75	.80	.85	.90	.95	.975	.99	.995	.9995
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.767
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725

$$\boxed{t_{23\,,\,0.95} = 1.714}$$

Notation

Statistical packages such as MINITAB, Excel, SAS, Stata, or R, will also compute exact probabilities associated with the t-distribution. This is particularly useful for values of the degrees of freedom (d) that are not given in Table 5.

Rule

If σ is unknown, we can replace σ by S and correspondingly replace the z statistic by a t statistic which should follow a t-distribution with degrees of freedom $(d) = n - 1$. The t statistic is given by:

$$t = \frac{(\overline{X} - \mu)}{S/\sqrt{n}} \sim t_{(n-1)}$$

Example

Suppose that the weights in kg's of newborn babies in Jordan are normally distributed with $\mu = 3$ kg. A random sample of size $n = 10$ babies is taken showed that the sample standard deviation is $S = 2$ kg. Find the probability that the sample mean $(\bar{X})$ is less than [redacted] 4.16 kg?

Solution

Since the population is normally distributed with unknown population standard deviation (σ), and small sample size $(n = 10 < 30)$, the sampling distribution of the sample mean $(\bar{X})$ is a student's t-distribution with degrees of freedom [redacted] $n - 1$ $= \quad 10 - 1 = 9$. To find the probability that the sample mean $(\bar{X})$ is less than [redacted] 4.16 kg, then we need to find $P(\bar{X} < 4.16)$ can be calculated as follows:

$$n = 10, \quad d = df = n - 1 = 10 - 1 = 9, \quad \mu = 3, \quad S = 2$$

$$P(\bar{X} < 4.16) = P\left(\frac{\bar{X}-\mu}{s/\sqrt{n}} < \frac{4.16-3}{2/\sqrt{10}}\right) = \boxed{P(t < 1.834)} \sim t_{(9)}$$

the percentile (u)

The value will be obtained from row 9 of Appendix Table 5

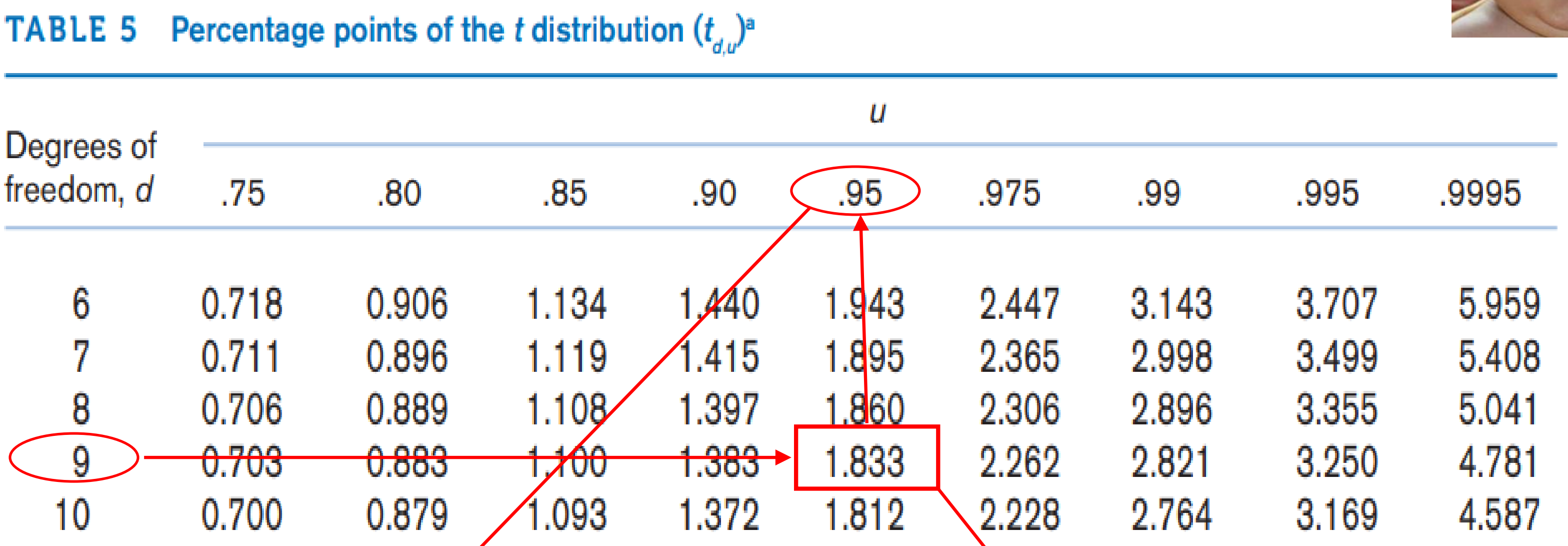


TABLE 5 Percentage points of the *t* distribution ($t_{d,u}$)[a]

Degrees of freedom, *d*	*u*								
	.75	.80	.85	.90	.95	.975	.99	.995	.9995
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587

$t_{9\,,\,0.95} = 1.833$

Then $\boxed{P(t < 1.834) = 0.95}$

Conclusion

Thus 95% of the samples of size 10 would be expected to have mean of newborn babies weights less than 4.16 kg.

Example

Suppose it is known that in a certain large human population cranial length is normally distributed with a mean of 185.6 mm. A random sample of size 10 is taken from this population showed that the standard deviation is 12.7 mm. What is the probability that the sample mean ($\bar{X}$) will be greater than or equal to 180 mm?

Solution

We have the following:

- X = cranial length$\sim N(185.6, \sigma^2)$.
- The population standard deviation $\sigma = \sqrt{\sigma^2}$ is unknown and therefore σ is replaced by the sample standard deviation $S = \sqrt{S^2}$= 12.7 mm.
- The sample size $n = 10 < 30$ is small.

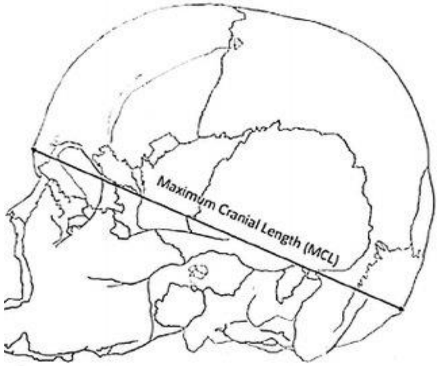
Thus, the sampling distribution of the sample mean ($\bar{X}$) is a student's t-distribution with degrees of freedom $d = n - 1 = 10 - 1 = 9$.

We need to find $P(\bar{X} \geq 180)$ as follows:

$$P(\bar{X} \geq 180) = P(\bar{X} \geq 180) = P\left(\frac{\bar{X}-\mu}{S/\sqrt{n}} \geq \frac{180-185.6}{4.016}\right)$$

$$= P(t_d \geq -1.394)$$

$$= P(t_d < 1.394) \sim t(9)$$



Because the t-distribution is symmetric about t = 0, we have area to the right of the value (-1.394) equals to the area to the left of the value (1.394), and this value will be obtained from row 9 of Table 5 in the Appendix.

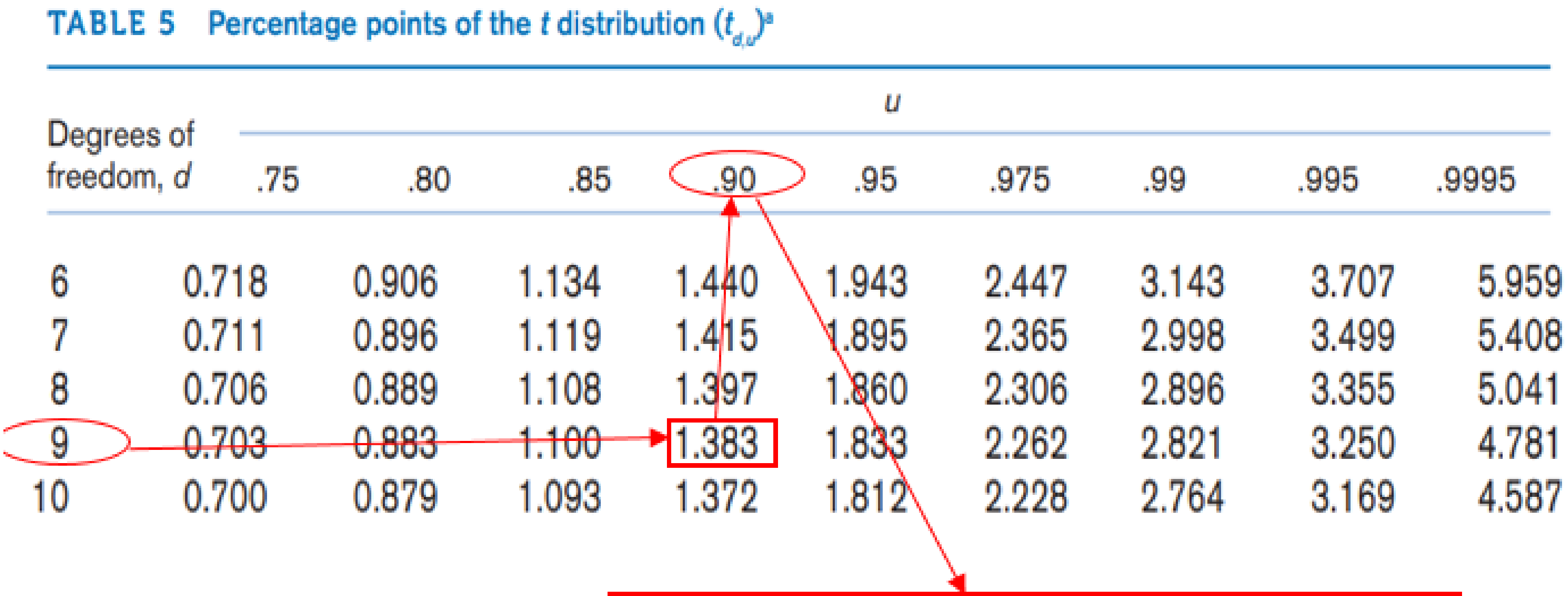


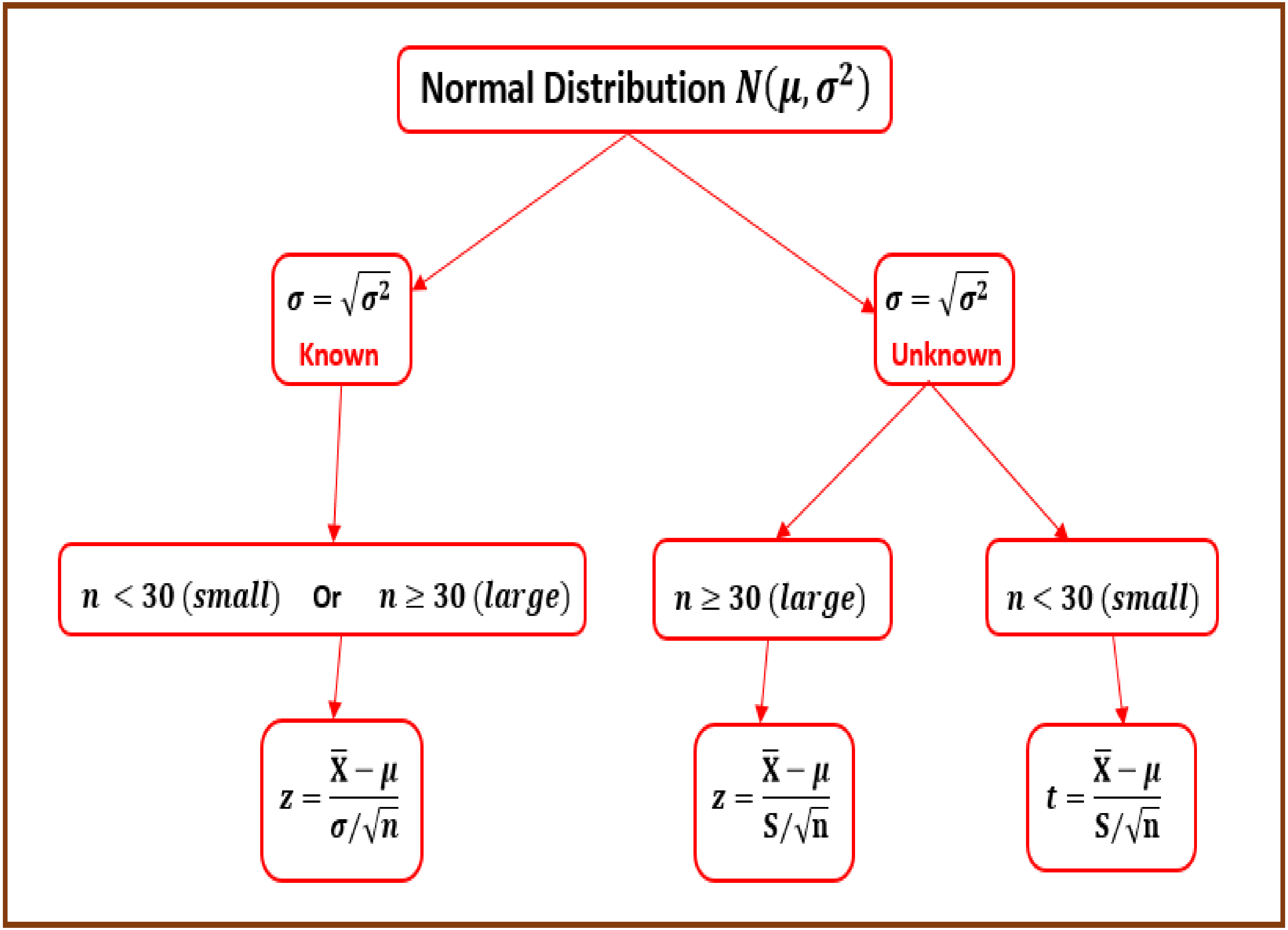
TABLE 5 Percentage points of the *t* distribution ($t_{d,u}$)[a]

Degrees of freedom, *d*	*u* .75	.80	.85	.90	.95	.975	.99	.995	.9995
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587

Then $\boxed{P(t_d \geq -1.394) = P(t_d < 1.394) = 0.90}$

Conclusion

Thus 90% of the samples of size 10 would be expected to have a mean of cranial length greater than or equal to 180 mm.



Normal Distribution $N(\mu, \sigma^2)$

$\sigma = \sqrt{\sigma^2}$ Known

$\sigma = \sqrt{\sigma^2}$ Unknown

$n < 30\ (small)$ Or $n \geq 30\ (large)$

$n \geq 30\ (large)$

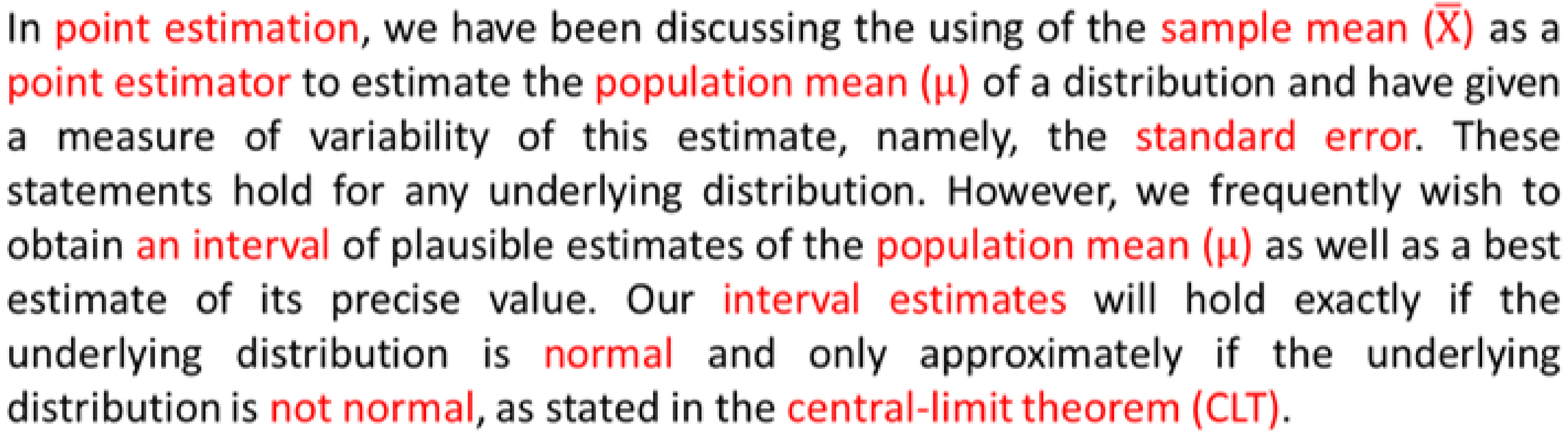
$n < 30\ (small)$

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

$$z = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

6.5.2 Interval Estimation



In point estimation, we have been discussing the using of the sample mean ($\overline{X}$) as a point estimator to estimate the population mean (μ) of a distribution and have given a measure of variability of this estimate, namely, the standard error. These statements hold for any underlying distribution. However, we frequently wish to obtain an interval of plausible estimates of the population mean (μ) as well as a best estimate of its precise value. Our interval estimates will hold exactly if the underlying distribution is normal and only approximately if the underlying distribution is not normal, as stated in the central-limit theorem (CLT).

Notation: Although the point estimator is a good estimator of the population parameter (θ), it is more meaningful to estimate θ by an interval that communicates information regarding the probable magnitude of θ, this interval is known as the confidence interval.

Definition: Interval Estimation

An interval estimator of a population parameter θ is an interval of the form:

$$\hat{\theta}_L < \theta < \hat{\theta}_U$$

where $\hat{\theta}_L$ and $\hat{\theta}_U$ depends on the value of the statistic $\hat{\theta}$ for a particular random sample $X_1, X_2, \dots, X_n$ and also on the sampling distribution of $\hat{\theta}$.

Notation: From the sampling distribution of $\hat{\theta}$ we shall be able to determine $\hat{\theta}_L$ and $\hat{\theta}_U$ such that the $P(\hat{\theta}_L < \theta < \hat{\theta}_U)$ is equal to any positive fractional value we care to specify. If for instance we find $\hat{\theta}_L$ and $\hat{\theta}_U$ such that:

$$P(\hat{\theta}_L < \theta < \hat{\theta}_U) = 1 - \alpha \quad \text{for } 0 < \alpha < 1$$

then we have a probability of $(1 - \alpha)$ of selecting a random sample that will produce an interval containing θ.

Definition: Confidence Interval

The interval $\hat{\theta}_L < \theta < \hat{\theta}_U$ computed from the selected random sample, is then called a $(1 - \alpha)100\%$ confidence interval, the fraction $(1 - \alpha)$ is called confidence coefficient (confidence level) or the degree of confidence and the end points $\hat{\theta}_L$ and $\hat{\theta}_U$ are called the lower and upper confidence limits. The general formula for constructing a confidence interval (CI) is given as follows:

CI = Point Estimator ± [(Critical Value)(Standard Error)]

where

- Point Estimator: is the sample statistic estimating the population parameter of interest.
- Critical Value: is a table value based on the sampling distribution of the point estimator and the desired confidence level.
- Standard Error: is the standard deviation of the point estimator.

In this section, we will learn how to:

Construct and interpret the (1 − α)100% confidence interval for the population mean (μ).

In the case that the underlying distribution is Normal Distribution, three cases for the Confidence Interval of the Population Mean (μ) are discussed:

- when Population Standard Deviation σ is Known (One Case).
- when Population Standard Deviation σ is Unknown (Two Cases).

Case (1): Confidence Interval for the Population Mean (μ) of a Normal Distribution (σ is Known)

Let $X_1, X_2, \dots, X_n$ be a random sample of size n taken from a normal distribution, $N(\mu, \sigma^2)$. If the population variance σ^2 (or population standard deviation $\sigma = \sqrt{\sigma^2}$) is known, then for ($n < 30$ (small) or $n \geq 30$ (large)), the (1 − α) × 100% confidence interval (CI) for the population mean μ of a normal distribution can be constructed as follows:

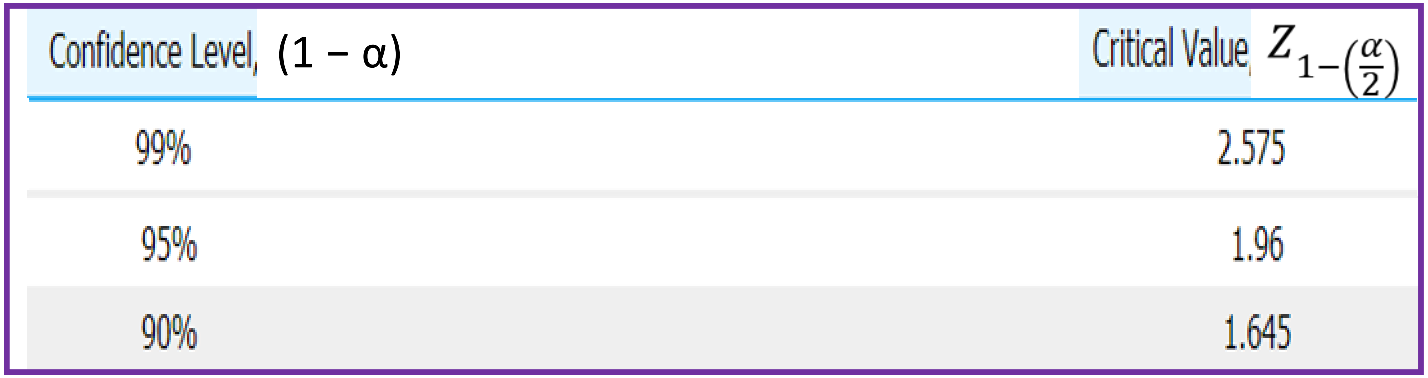
$$\boxed{\mathrm{CI} = \left(\overline{\mathrm{X}} - Z_{1-\left(\frac{\alpha}{2}\right)} \times \frac{\sigma}{\sqrt{n}} \quad , \quad \overline{\mathrm{X}} + Z_{1-\left(\frac{\alpha}{2}\right)} \times \frac{\sigma}{\sqrt{n}}\right)}$$

Notation: The conditions to use this confidence interval (CI) are:

(1) Normal Distribution.

(2) Population Standard Deviation $\sigma = \sqrt{\sigma^2}$ is Known ($n < 30$ or $n \geq 30$).

Normal Critical Values for Confidence Levels



Confidence Level, $(1-\alpha)$	Critical Value, $Z_{1-\left(\frac{\alpha}{2}\right)}$
99%	2.575
95%	1.96
90%	1.645

Example

Suppose that a physical therapist wished to estimate the mean maximal strength of a particular muscle in a certain group of individuals. Assume that the strength scores are normally distributed with a variance of 144. A random sample of 15 subjects who participated in the study is taken showed that the mean is 84.3. Construct the 99% confidence interval for the mean (μ) of maximal strength scores?

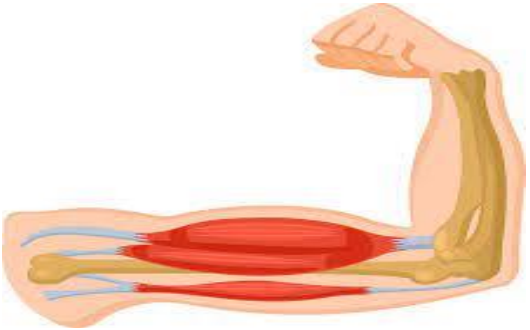
Solution *We have:*

1- Normal distribution.

2- The standard deviation σ is known.

$$n = 15$$

$$\bar{X} = 84.3 \quad , \sigma = \sqrt{144} = 12 \,, (1-\alpha)\times 100\,\% = 99\,\%$$



The (1 - α) × 100 % = 99% confidence interval for the population mean (μ) of maximal strength scores can be constructed as follows:

Step(1)

$(1-\alpha) \times 100\,\% = 99\%$ implies that $Z_{1-\left(\frac{\alpha}{2}\right)} = 2.575$

Step(2)

$$\text{CI} = \bar{X} \pm Z_{1-\left(\frac{\alpha}{2}\right)} \times \frac{\sigma}{\sqrt{n}}$$

$$\text{Lower Limit} = \bar{X} - Z_{1-\left(\frac{\alpha}{2}\right)} \times \frac{\sigma}{\sqrt{n}}$$

$$= 84.3 - (2.575)\left(\frac{12}{\sqrt{15}}\right)$$

$$= 84.3 - 7.978$$

$$= 76.322 \approx 76.3$$

$$\text{Upper Limit} = \bar{X} + Z_{1-\left(\frac{\alpha}{2}\right)} \times \frac{\sigma}{\sqrt{n}}$$

$$= 84.3 + (2.575)\left(\frac{12}{\sqrt{15}}\right)$$

$$= 84.3 + 7.978$$

$$= 92.278 \approx 92.3$$

Then CI =(L, U) = (76.3, 92.3)

Conclusion

We are 99% confident that the population mean (μ) of maximal strength scores is between 76.3 and 92.3 since, in repeated sampling, 99% of all intervals that could be constructed would include the population mean (μ) and 1% of these intervals will not include the population mean (μ).

Case (2): Confidence Interval for the Population Mean (μ) of a Normal Distribution (σ is Unknown and n is Large)

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n taken from a normal distribution, $N(\mu, \sigma^2)$. If the population variance σ^2 (or population standard deviation $\sigma = \sqrt{\sigma^2}$) is unknown and sample size n is large ($n \geq 30$), then replace $\sigma = \sqrt{\sigma^2}$ by the sample standard deviation $S = \sqrt{S^2}$ and the (1 – α) × 100% confidence interval (CI) for the population mean μ of a normal distribution can be constructed as follows:

$$\boxed{\mathrm{CI} = \left(\overline{\mathrm{X}} - Z_{1-\left(\frac{\alpha}{2}\right)} \times \frac{S}{\sqrt{n}} \quad , \quad \overline{\mathrm{X}} + Z_{1-\left(\frac{\alpha}{2}\right)} \times \frac{S}{\sqrt{n}}\right)}$$

Notation: The conditions to use this confidence interval (CI) are:

(1) Normal Distribution.

(2) Population Standard Deviation $\sigma = \sqrt{\sigma^2}$ is Unknown.

(3) The sample size (n) is large ($n \geq 30$).

Example

Suppose that we wish to estimate the mean number of heartbeats per minute for a certain population (μ). Assume that the number of heartbeats per minute is normally distributed. The average number of heartbeats per minute for a random sample of size 49 subjects was found to be 90 with a standard deviation of 10. Find the 95% confidence interval for the population mean (μ)?

Solution

We have:

1- Normal distribution.

2- The standard deviation σ is unknown ($S = 10$).

3- The sample size (n) is large ($n = 49 \geq 30$).

$$n = 49$$
$$\bar{X} = 90 \quad , S = 10 \, , (1 - \alpha) \times 100 \,\% = 95\%$$

The (1 - α) × 100 % = 95% confidence interval for the population mean (μ) of the number of heartbeats per minute can be constructed as follows:

Step(1)

$(1 - \alpha) \times 100\,\% = 95\%$ implies that $Z_{1-\left(\frac{\alpha}{2}\right)} = 1.96$

Step(2)

$$CI = \bar{X} \pm Z_{1-\left(\frac{\alpha}{2}\right)} \times \frac{S}{\sqrt{n}}$$

$$\text{Lower Limit} = \bar{X} - Z_{1-\left(\frac{\alpha}{2}\right)} \frac{S}{\sqrt{n}}$$
$$= 90 - (1.96)\left(\frac{10}{\sqrt{49}}\right)$$
$$= 90 - 2.8$$
$$= 87.2$$

$$\text{Upper Limit} = \bar{X} + Z_{1-\left(\frac{\alpha}{2}\right)} \frac{S}{\sqrt{n}}$$
$$= 90 + (1.96)\left(\frac{10}{\sqrt{49}}\right)$$
$$= 90 + 2.8$$
$$= 92.8$$

Then CI =(L, U) = (87.2 , 92.8)

Conclusion

With 95% confidence we can say that the mean number of heartbeats per minute for all subjects in the population (μ) is between 87.2 and 92.8.

Case (3): Confidence Interval for the Population Mean () of a Normal Distribution (σ is Unknown and n is Small)

Let $X_1, X_2, \dots, X_n$ be a random sample of size (n) taken from a normal distribution, $N(\mu, \sigma^2)$. If the population variance σ^2 (or the population standard deviation $\sigma = \sqrt{\sigma^2}$) is unknown and sample size n is small (n < 30), then replace $\sigma = \sqrt{\sigma^2}$ by the sample standard deviation $S = \sqrt{S^2}$ and the $(1-\alpha) \times 100\%$ confidence interval (CI) for the population mean (μ) of a normal distribution will be constructed as follows:

$$\boxed{\mathrm{CI} = \left(\bar{\mathrm{X}} - t_{\left(n-1\,,\,1-\left(\frac{\alpha}{2}\right)\right)} \times \frac{\mathrm{S}}{\sqrt{\mathrm{n}}}\,, \bar{\mathrm{X}} + t_{\left(n-1\,,\,1-\left(\frac{\alpha}{2}\right)\right)} \times \frac{\mathrm{S}}{\sqrt{\mathrm{n}}}\right)}$$

Notation: The conditions to use this confidence interval (CI) are:

(1) Normal Distribution.

(2) Population Standard Deviation $\sigma = \sqrt{\sigma^2}$ is Unknown.

(3) The Sample size (n) is small (n < 30).

Important Notation: An important point to understand is that the boundaries of any confidence interval (CI) depends on the sample mean and population or sample variance and vary from sample to sample.

Example

In a random sample of 20 patients at a given clinic in Jordan, the mean waiting time to measure the blood pressure is 95 seconds, and the standard deviation is 21 seconds. Assume that the waiting times are normally distributed, then construct a 99% confidence interval (CI) for the mean of waiting times of all patients (μ)?

Solution

We have:

1- Normal distribution.

2- The standard deviation σ is unknown (S = 21).

3- The sample size (n) is small (n = 20 < 30).

$$n = 20$$
$$\bar{X} = 95 \quad , S = 21$$
$$(1 - \alpha)\ 100\ \% = 99\ \%$$

The (1-α)100 % = 99% confidence interval for the population mean (μ) can be constructed as follows:

Step(1)

$$(1 - \alpha)\ 100\ \% = 99\ \%$$

$$1 - \alpha = 0.99$$

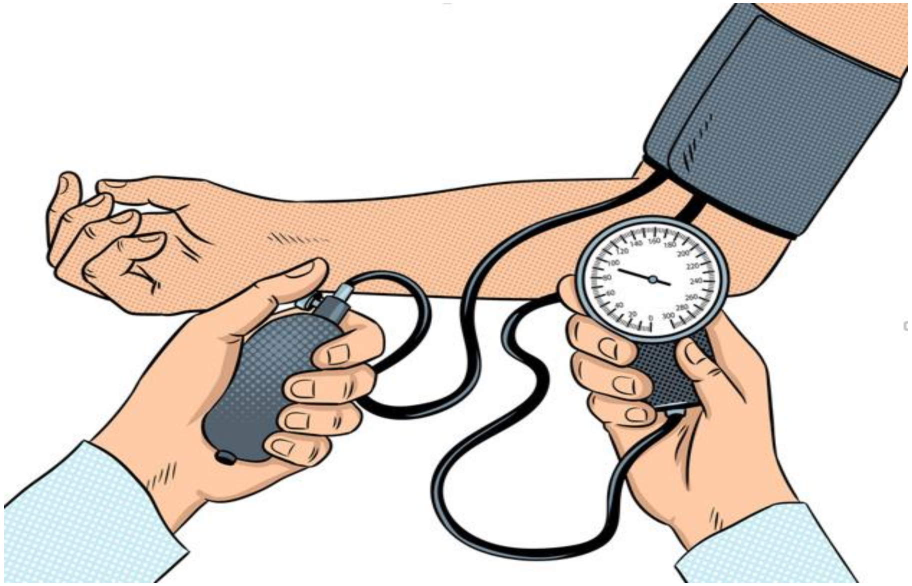
$$\alpha = 0.01$$

$$\frac{\alpha}{2} = \frac{0.01}{2} = 0.005$$

$$1 - \frac{\alpha}{2} = 1 - 0.005 = 0.995$$

$$d = n - 1 = 20 - 1 = 19$$

$$t_{(n-1,\, 1-\frac{\alpha}{2})} = t_{(19\,,\, 0.995)} = 2.861$$



Step(2)

$$CI = \bar{X} \pm t_{(n-1,\,1-\frac{\alpha}{2})} \frac{S}{\sqrt{n}}$$

Lower Limit = $\bar{X} - t_{(n-1,\,1-\frac{\alpha}{2})} \frac{S}{\sqrt{n}}$

$= 95 - (2.861)\left(\frac{21}{\sqrt{20}}\right)$

$= 95 + (2.861)\left(\frac{21}{\sqrt{20}}\right)$

$= 95 + (2.861)\,(4.583)$

$= 95 + 13.112$

$= 108.11$

Upper Limit = $\bar{X} + t_{(n-1,\,1-\frac{\alpha}{2})} \frac{S}{\sqrt{n}}$

$= 95 + (2.861)\left(\frac{21}{\sqrt{20}}\right)$

$= 95 - (2.861)\left(\frac{21}{\sqrt{20}}\right)$

$= 95 - (2.861)\,(4.583)$

$= 95 - 13.112$

$= 81.89$

Then the 99% confidence interval for μ is CI =(L, U) = (81.89 , 108.11) seconds.

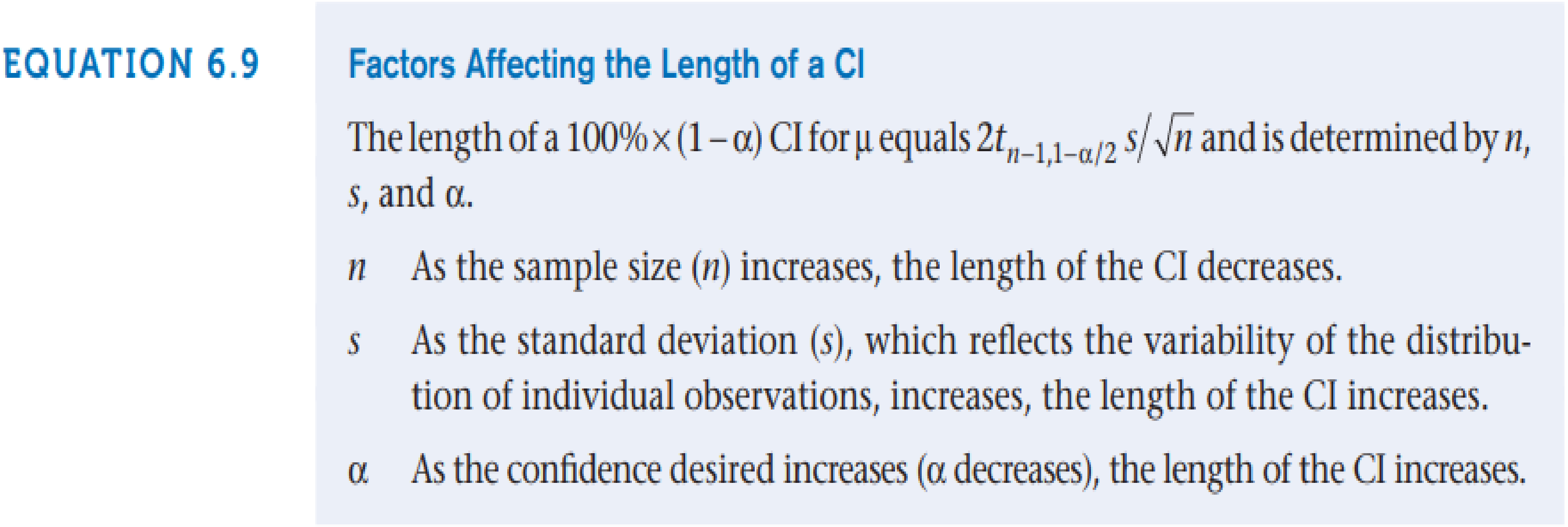
Conclusion

With 99% confidence interval we can say that the mean waiting time of all patients (μ) is between 81.89 and 108.11 seconds.

Result: The length of a confidence interval (CI) is governed by three variables or factors as follows:

- Sample size (n).
- Standard Deviation (S or σ).
- Level of Significance (α).

In general, the length of the (1 − α) × 100% CI is given by $2t_{(n-1,\,1-\frac{\alpha}{2})} \frac{S}{\sqrt{n}}$.



EQUATION 6.9

Factors Affecting the Length of a CI

The length of a 100% × (1 − α) CI for μ equals $2t_{n-1,1-\alpha/2}\, s/\sqrt{n}$ and is determined by n, s, and α.

n As the sample size (n) increases, the length of the CI decreases.

s As the standard deviation (s), which reflects the variability of the distribution of individual observations, increases, the length of the CI increases.

α As the confidence desired increases (α decreases), the length of the CI increases.

Note that: To understand the three points given in Equation (6.9) and to make more practice on the confidence intervals for the population mean (μ) study the following examples given in pages 177 – 180 of the textbook:

EXAMPLE 6.31

EXAMPLE 6.32

EXAMPLE 6.33

EXAMPLE 6.35

EXAMPLE 6.36

EXAMPLE 6.37

Central Limit Theorem (CLT) Case for Confidence Interval

To use the central limit theorem (CLT) to construct the (1-α)100% confidence interval for the population mean (μ) the following two conditions should be satisfied:

1- Unknown distribution (population) or non-normal distribution with a population mean (μ) and a population variance (σ^2) or a population standard deviation ($\sigma = \sqrt{\sigma^2}$). If σ is unknown replace it by the sample standard deviation (S).

2- The sample size (n) is large ($n \geq 30$).

Then we assume the sampling distribution of the sample mean ($\overline{X}$) to be approximately normally distributed by using the central limit theorem (CLT) and therefore the (1-α)100% confidence interval for the population mean (μ) can be constructed using the following formula:

$$\boxed{CI = \left(\overline{X} - Z_{1-\left(\frac{\alpha}{2}\right)} \times \frac{\sigma}{\sqrt{n}} \quad , \quad \overline{X} + Z_{1-\left(\frac{\alpha}{2}\right)} \times \frac{\sigma}{\sqrt{n}}\right)}$$

Example

Suppose that we wish to estimate the mean number of heartbeats per minute for a certain population (μ). The average number of heartbeats per minute for a random sample of size 49 subjects was found to be 90. Previous research has shown that the standard deviation for the population to be about 10. Find the 90% confidence interval (CI) for the population mean (μ)?

Solution

We have:

1- Unknown distribution (population).

2- The standard deviation σ is known (σ = 10).

3- The sample size (n) is large (n = 49 > 30).

Now, since the above conditions are satisfied, we draw on the central limit theorem and assume that the sampling distribution of the sample mean ($\overline{\mathrm{X}}$) to be approximately normally distributed.

The (1 - α) × 100 % = 90% confidence interval for the population mean (μ) of the number of heartbeats per minute can be constructed as follows:

$$\boxed{\begin{array}{c} n = 49 \\ \bar{X} = 90 \quad , \ \sigma = 10 \\ (1 - \alpha)\ 100\ \% = 90\% \end{array}}$$

Step(1)

$(1-\alpha)$ 100 % = 90% implies that $Z_{1-\left(\frac{\alpha}{2}\right)}$ = 1.645

Step(2)

The $(1-\alpha)$100% = 90% confidence interval (CI) for the population mean (μ) of the number of heartbeats per minute can be calculated as follows:

$$\mathrm{CI} = \bar{X} \pm Z_{1-\left(\frac{\alpha}{2}\right)} \frac{\sigma}{\sqrt{n}} = 90 \pm \left[(1.645)\left(\frac{10}{\sqrt{49}}\right)\right] = 90 \pm 2.35 = (87.65\,,92.35)$$

Conclusion

With 90% confidence interval we can say that the mean number of heartbeats per minute for all subjects in the population (μ) is between 87.65 and 92.35.

6.8 Estimation for the Binomial Distribution

6.8.1 Point Estimation

In this section, point estimation for the parameter p of a binomial distribution using the random variable $\hat{p}$ (sample proportion) is discussed.

Equation 6.16

Let X be a binomial random variable with parameters n and p, that is, $X \sim B(n, p)$. Thus, for any random sample of size n the sample proportion $\hat{p}$ is an unbiased estimator of the population proportion p, that is, $\mathrm{E}(\hat{p}) = p$. The standard error of this proportion is given exactly by $\mathrm{se}(\hat{p}) = \sqrt{p\,q/n}$ and is estimated by $\mathrm{se}\,(\hat{p}) = \sqrt{\hat{p}\,\hat{q}/n}$ where $\hat{q} = 1 - \hat{p}$.

Question: How to calculate the value of the sample proportion $(\hat{p})$?

Answer

$$\boxed{\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = X/n}$$

where

$$X_i = \begin{cases} 1 \quad , & \text{if the } i^{th} \text{ unit have the specified characteristic} \\ 0 \quad , & \text{if the } i^{th} \text{ unit does not have the specified characteristic} \end{cases}$$

Note that: The sample proportion $(\hat{p})$ is the point estimator for the population proportion (p).

EXAMPLE 6.43

Cancer Consider the problem of estimating the prevalence of malignant melanoma in 45- to 54-year-old women in the United States. Suppose a random sample of size 5000 women is selected from this age group, of whom 28 are found to have the disease. Suppose that the prevalence of the disease in this age group = p and let the random variable X_i represent the disease status for the i^{th} woman, then for i = 1, . . . , 5000 we have

$$X_i = \begin{cases} 1 \quad , & \text{if the } i^{th} \text{ woman has the disease} \\ 0 \quad , & \text{if the } i^{th} \text{ woman does not has the disease} \end{cases}$$

Answer the following:

(a) Estimate the prevalence of malignant melanoma in this age group (p)?

Solution

- $n = 5000$
- $X = \sum_{i=1}^{n\,=\,5000} X_i$ = the number of women with malignant melanoma among the 5000 women = 28

Thus, the prevalence of malignant melanoma in this age group (a among 45- to 54-year-old women) p can be estimated by using $(\hat{p})$ as follows:

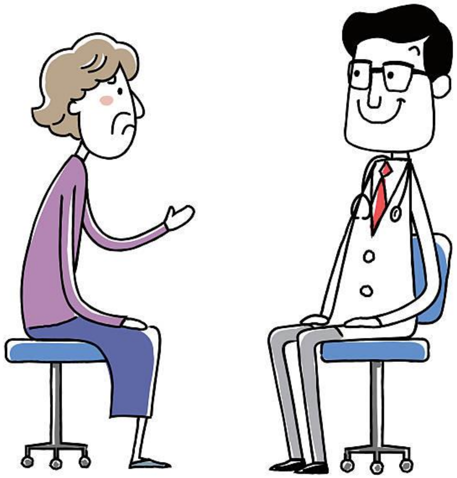
$$\boxed{\hat{p} = \frac{X}{n} = \frac{\sum_{i=1}^{n\,=\,5000} X_i}{n} = \frac{28}{5000} = 0.0056 \text{ (or } 0.56\%).}$$

(b) Calculate the value of the standard error of the sample proportion ($\hat{p}$)?

Solution

The estimated standard error can be calculated as follows:

$$\begin{aligned} \text{se}\,(\hat{p}) &= \sqrt{\hat{p}\,\hat{q}/n} \\ &= \sqrt{(0.0056)(1-0.0056)/5000} \\ &= \sqrt{(0.0056)(0.9944)/5000} \\ &= 0.001055 \\ &\approx 0.0011 \end{aligned}$$



6.8.2 Interval Estimation—Normal-Theory Method

In the previous section, the point estimation of the population parameter (p) of a binomial distribution was covered. In this section, we will learn how can an interval estimate (confidence interval) of the population parameter (p) can be obtained.

Now, For large n, and from the central-limit theorem, we can see that $\hat{p} = \overline{\mathrm{X}}$ is normally distributed with mean $\mu_{\hat{p}} = p$ and variance $\sigma_{\hat{p}}^2 = \frac{\sigma^2}{n} = \frac{pq}{n} = \frac{p(1-p)}{n}$.

EQUATION 6.17

$$\hat{p} \dot{\sim} N(p, pq/n)$$ For large n

Now, an approximate $(1-\alpha) \times 100\%$ confidence interval (CI) for the population parameter (p) (population proportion) can be obtained from Equation 6.19 as follows:

EQUATION 6.19

Normal-Theory Method for Obtaining a CI for the Binomial Parameter *p* (Wald Method)

An approximate 100% × (1 − α) CI for the binomial parameter *p* based on the normal approximation to the binomial distribution is given by

$$\text{CI} = \hat{p} \pm z_{1-\alpha/2}\sqrt{\hat{p}\hat{q}/n} = \left(\hat{p} - z_{1-\alpha/2}\sqrt{\hat{p}\hat{q}/n}, \hat{p} + z_{1-\alpha/2}\sqrt{\hat{p}\hat{q}/n}\right) \text{ where } \hat{q} = 1 - \hat{p}$$

This method of interval estimation should only be used if $n\hat{p}\hat{q} \geq 5$.

EXAMPLE **COVID-19**

Suppose that a ministry of health in a certain country is interested to estimate the percent of adults living in a large city who have COVID-19. A random sample of 500 adult residents in this city are tested to determine whether they have COVID-19. Suppose that out of the 500 people tested, 421 are infected. Derive an approximate 95% confidence interval (CI) for the true proportion (p) of adult residents of this city who have COVID-19?

Solution

The $(1-\alpha)\text{x}100\% = 95\%$ confidence interval (CI) for the population proportion (p) can be constructed (derived) as follows:

Step(1)

We calculate the value of the sample proportion as follows:

$$\hat{p} = \frac{X}{n} = \frac{421}{500} = 0.842 \quad \text{implies that} \quad \hat{q} = 1 - \hat{p} = 1 - 0.842 = 0.158$$

Step(2)

We have that

$$n\hat{p}\hat{q} = (500)(0.842)(0.158) = 66.518 > 5.$$

Thus, we can use the large sample method in Equation 6.19.

Step(3)

$(1-\alpha)100\% = 95\%$; $1 - \alpha = 0.95$; $\alpha = 0.05$; $\frac{\alpha}{2} = \frac{0.05}{2} = 0.025$

Thus $1 - \frac{\alpha}{2} = 1 - 0.025 = 0.975$ implies that $Z_{1-\frac{\alpha}{2}} = Z_{0.975} = 1.96$

Step(4)

The approximate 95%confidence interval (CI) for for the true proportion (p) of adult residents of this city who have COVID-19 can be constructed as follows:

$$n = 500 \quad ; \quad \hat{p} = 0.842 \quad ; \quad \hat{q} = 0.158 \quad ; \quad (1-\alpha)\ 100\ \% = 95\%$$

$$\boxed{\text{CI} = \hat{p} \pm Z_{1-(\frac{\alpha}{2})}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

Lower Limit = $\hat{p} - Z_{1-(\frac{\alpha}{2})}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$= 0.842 - [(1.96)\sqrt{\frac{(0.842)(0.158)}{500}}]$

$= 0.842 - 0.0320$

$= 0.810$

Upper Limit = $\hat{p} + Z_{1-(\frac{\alpha}{2})}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$= 0.842 + [(1.96)\sqrt{\frac{(0.842)(0.158)}{500}}]$

$= 0.842 + 0.0320$

$= 0.874$

Then CI =(L, U) = (0.810, 0.874)

Conclusion

With 95% confidence we can say that the true proportion (p) of adult residents of this city who have COVID-19 is between 0.810 and 0.874.

Problems: 6.5 – 6.9, 6.11 – 6.17, 6.27 – 6.32, 6.52 – 6.55.

$= 95 - (2.861)\ (\frac{21}{\sqrt{20}})$

$= 95 - (2.861)\ (4.583)$

$= 95 - 13.112$

$= 81.89$

$= 95 + (2.861)\ (\frac{21}{\sqrt{20}})$

$= 95 + (2.861)\ (4.583)$

$= 95 + 13.112$

$= 108.11$

CI= (81.89 , 108.11)