-5-
Sampling and distribution of sample means

5.1 Introduction

Let us consider the following process: Taking underground water samples from different locations in a field and analyzing for total dissolved salts content. To this aim, the location must be divided into several sections having the same area (say 4 m^2). If we have 500 such sections (Population), we will choose 50 of them (Sample) to take underground water specimens. The choice is made by giving each section a serial number: 1, 2, 3, etc. Then 50 computer generated random numbers between 1 and 500 are produced: 224, 23, 28, 326, 489, 450, 238, 105, 167, 469, etc. Water specimens will be taken out of sections corresponding to these numbers. Random integers between A and B are generated by the EXCEL function RANDBETWEEN (A, B).

The question to be answered in the present chapter is the following: To what extent do the results obtained for the samples represent the population?

5.2 The distribution of sample means

5.2.1 Basic concepts

In the preceding example, it is obvious that each time, a different computer run will generate a new set of random numbers. Hence each time, the sample obtained will have a different mean and a different standard deviation. The sample mean will be denoted by $\bar{x}$; it represents a random variable of mean value μ_x and standard deviation $\sigma_{\bar{x}}$. This concept is best understood by the following example.

Consider a processing unit containing 6 reactors.
The following table shows the time elapsed a reactor must be revamped.

Converter	A	B	C	D	E	F
***x* days**	240	300	255	270	264	270

From these data, the mean μ = 265.5 and the standard deviation σ = 19.94

This is a small population such that no sampling is necessary. Let us assume, however, that a sample of 3 reactors is going to be chosen. There are C_3^6 = 20 such samples. The following table shows these samples together with the mean value of each.

Sample	ABC	ABD	ABE	ABF	ACD	ACE	ACF	ADE	ADF	AEF
Mean	265	270	268	270	255	253	255	258	260	258
Sample	BCD	BCE	BCF	BDE	BDF	BEF	CDE	CDF	CEF	DEF
Mean	275	273	275	278	280	278	263	265	263	268

The mean value of all sample means can be calculated. $\mu_{\bar{x}}$ = 265.5 = μ

While the standard deviation of sample means is $\sigma_{\bar{x}}$ = 8.35 < 19.94 (σ)

This example sets a very important principle: **"The mean value of sample means is equal to the mean value of population"**; that is:

$$\mu_{\bar{x}} = \mu \tag{5.1}$$

As for the standard deviation of sample means, it is always smaller than the standard deviation of population. This means that the values of sample means are less scattered about their mean value than the values of *x* for population.

If the size of sample *n* > 30, then the following relation holds for $\sigma_{\bar{x}}$:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{5.2}$$

In the above example, since *n* = 3 only, the latter relation cannot be applied. The value of calculated by this relation would have been:

$\frac{19.94}{\sqrt{3}}$ = 11.5 ≠ the calculated value of 8.35

Example 5.1

A factory produces 40000 sacks of PE beads of average weight = 50 kg and standard deviation = 0.5 kg. Samples of size 400 are chosen. Calculate the mean and standard deviation of sample means.

Solution:

$\mu_{\bar{x}} = \mu$ = 50

And $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.5}{\sqrt{400}}$ = 0.025 kg

5.2.2 Distribution of sample means in a population that is normally distributed

If the values of the random variable *x* are normally distributed in the population, then the mean values of sample means will also be normally distributed with the same mean value but less scattered since $\sigma_{\bar{x}} < \sigma$ (Equations 5.1 and 5.2). Figure (5.1) depicts this situation.

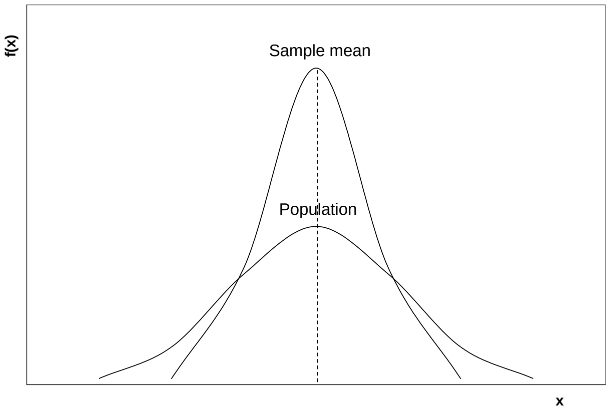

Fig (5.1): Distribution of population and sample means

5.2.3 Distribution of sample means in a population that is not normally distributed

Even if the random variable is not normally distributed over the population, it can be proved that, provided $n > 30$, the sample means will still be normally distributed and equations (5.1) and (5.2) will still hold. The following figure illustrates this principle.

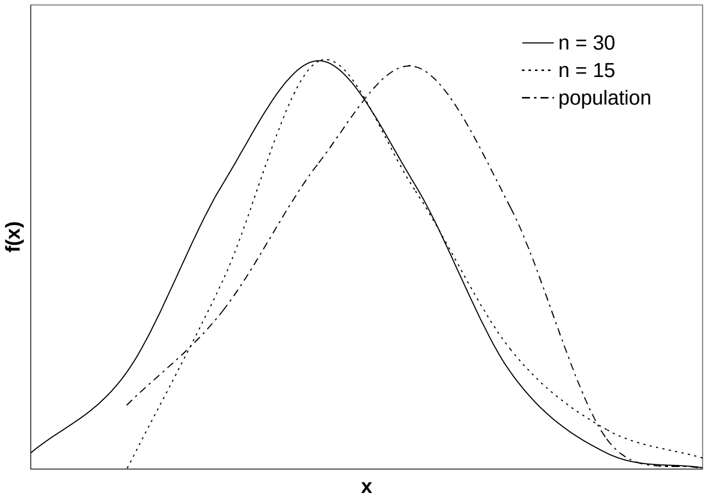

Fig (5.2): Effect of size on distribution of means

5.3 Estimation of the population mean: Levels of confidence

5.3.1 Point estimate of the population mean: Large samples

Consider a population of unknown mean μ with $n > 30$. If we choose a random sample of mean value $\bar{x},$ this value will not necessarily be equal to the mean of the population. As previously stated, the mean values of samples will be normally distributed about μ. It is said that $\bar{x}$ is a **point estimate** of the population mean μ.

The question to be asked is the following: To what extent does the sample mean differ from the population mean?

Let the difference between the two means be a, then the following inequality holds: $-a < \bar{x} - \mu < a$

The probability of such occurrence is: $P\,(-a < \bar{x} - \mu < a\,)$.

Dividing all sides by the standard deviation of sample means $\sigma_{\bar{x}}$, this probability becomes:

$$P\left(\frac{-a}{\sigma_{\bar{x}}} < \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} < \frac{a}{\sigma_{\bar{x}}}\right)$$

Since the means are normally distributed about μ, then this can be written in the following form:

$$P\left(\frac{-a}{\sigma_{\bar{x}}} < z < \frac{a}{\sigma_{\bar{x}}}\right) \quad \text{or} \quad P(-z_{crit} < z < z_{crit}) \quad \text{or}$$

$$P\left(-z_{crit} < \frac{\bar{x}-\mu}{\sigma_{\bar{x}}} < z_{crit}\right) \tag{5.3}$$

The value of this probability is termed the **level of confidence L**. This is an arbitrary level that expresses the extent to which the sample mean will differ from the population mean. The usual values of L are 0.9, 0.95 and 0.99. Each of these levels corresponds to a definite value of $z_{crit} = a / \sigma_{\bar{x}}$ that can be obtained from the normal tables. The following table shows these values.

Table 5.1: Values of z_{crit} corresponding to different confidence levels

L	0.99	0.95	0.90
***z*crit**	2.58	1.96	1.64

The inequality in equation (5.3) can be written as: $-z_{crit}.\sigma_{\bar{x}} < \bar{x} - \mu < z_{crit}.\sigma_{\bar{x}}$

Or, from equation (5.2):

$$-z_{crit}.\frac{\sigma}{\sqrt{n}} < \bar{x} - \mu < z_{crit}.\frac{\sigma}{\sqrt{n}} \tag{5.4}$$

So, if the value of σ is known, we can estimate, to any level of confidence, the deviation of sample mean from population mean.

Since it is seldom possible to know the value of σ, the standard deviation of population, it is safe enough to use the standard deviation of sample instead (s). This way, the above equation becomes:

$$-z_{crit}.\frac{s}{\sqrt{n}} < \bar{x} - \mu < z_{crit}.\frac{s}{\sqrt{n}}$$

Better written as: $\bar{x} - z_{crit}.\frac{s}{\sqrt{n}} < \mu < \bar{x} + z_{crit}.\frac{s}{\sqrt{n}}$ (5.5)

The interval defined by equations (5.4) or (5.5) is **a confidence interval for mean value of population.**

Finally, it is worth mentioning that the value $\alpha = 1 - L$ is called, in statistical terminology: **Level of significance**.

For values of L different from Table (5.1), the value of z_{crit} is obtained from the EXCEL function NORM.S.INV with entry $\frac{1+L}{2}$

For example, for a confidence level of 0.98, z_{crit} is obtained by introducing the entry 1.98/2 = 0.99 into the function NORM.S.INV. We get z_{crit} = 2.326

Example 5.2

Data belonging to a sample of 35 mortar cubes tested for compressive strength (MPa) showed that the mean value is 24.2 MPa and the standard deviation = 3.75 MPa. Estimate the value of the mean strength of the batch using a 95% confidence level**.**

Solution:

$\bar{x}$ = 24.2 and s = 4.75 MPa

Applying equation (5.5), we get for L = 0.95:

$$24.2 - \frac{1.96 \times 4.75}{\sqrt{35}} < \mu < 24.2 + \frac{1.96 \times 4.75}{\sqrt{35}}$$

or: **$22.63 < \mu < 25.77$**

This result means that **we are 95% sure that the mean value of population lies between 22.63 and 25.77 MPa.**

Note that if we increase the level of confidence to 99%, then the value of z_{crit} = 2.58, and application of equation (5.5) will give $22.13 < \mu < 26.27$, which is wider a range.

5.3.2 Effect of confidence level on width of confidence interval of mean

In general, the higher the confidence level used, the higher will be the error in predicting the. This error = $\frac{z_{crit} \cdot s}{\sqrt{n}}$

Consider for example a sample of size = 30 drawn from a population. Let the standard deviation of sample = 3. Figure (5.3) clearly shows that increasing the value of L will increase the confidence interval width yielding more uncertain estimations for the population mean. That is why it is common to use 0.95 as reasonable confidence level.

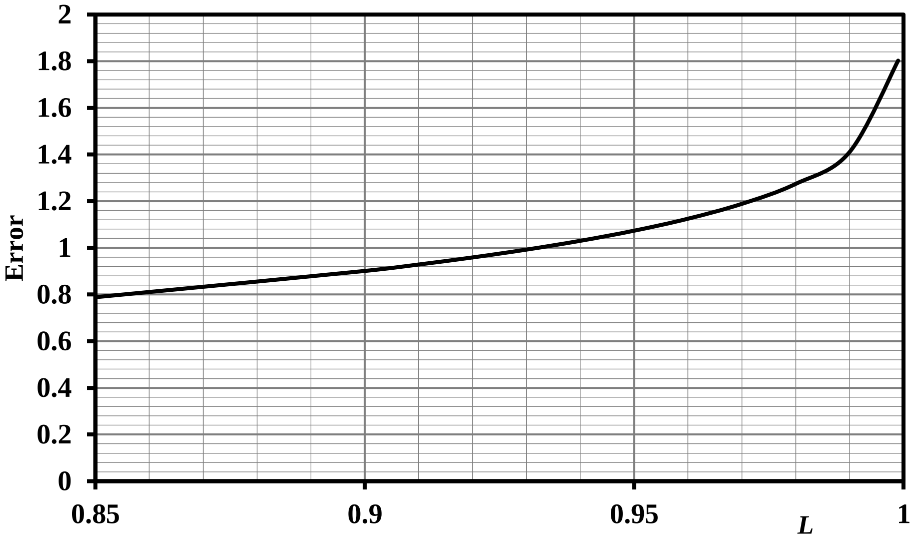

Fig (5.3): Effect of confidence level on the accuracy of predicting the population mean

5.3.3 The case of small samples

If $n < 30$, then the normal approximation for sample means is no longer valid; we must use another distribution. This is known as **the *t* – distribution**. This is a symmetrical distribution like the normal distribution. Its density function is given by the following law:

$$f(t) = \frac{\Gamma\left(\frac{n}{2}\right).\left(1+\frac{t^2}{n-1}\right)^{-\frac{n}{2}}}{\sqrt{(n-1).\pi.\Gamma(\frac{n-1}{2})}} \qquad (5.6)$$

This distribution gives values of a parameter *t* instead of z_{crit} in equation (5.4), which is a function of both *α* and *n* in the form: $t\ (\alpha,\ n-1)$, where $n-1$ is the number of degrees of freedom (*d.f.*)

In case of small samples, equation (5.5) is written as follows:

$$\bar{x}-t.\frac{s}{\sqrt{n-1}} < \mu < \bar{x}+t.\frac{s}{\sqrt{n-1}} \qquad (5.7)$$

The EXCEL (office 2010 or 2013) function T.INV2t can be readily used to display the required value of *t* for a known value of *α* and *d.f.* = *n* – 1

There is however one major assumption related to the use of the *t* – distribution; namely, that the population is normally distributed.

Example 5.3

Six specimens of lubricating oil were chosen from the daily production of a factory and tested for viscosity, the values were: 105, 112, 97, 102, 107, 107 cP. At 95% confidence level, estimate the mean value of viscosity in the produced batch.

Solution:

The average value of sample was calculated as 105 cP
The standard deviation was calculated as 5.1 cP
For a level of significance = 1 – 0.95 = 0.05 and *d.f.* = 6 – 1 = 5, we get form t – table: *t* = 2.571
Replacing in equation (5.7), we get:

$$105-\frac{2.57\times 5.1}{\sqrt{5}} < \mu < 105+\frac{2.57\times 5.1}{\sqrt{5}}$$ or $\mathbf{99.13 < \mu < 110.86}$

5.3.4 Sample size

Suppose that we wish to choose a sample from a population and ensure, at a given confidence level, that the population mean will not differ from the sample mean by more than a certain amount.

Let the maximum deviation be *D*, hence: $|\ \bar{x}-\mu\ | = D$

Hence, $D = \frac{z_{crit}\sigma}{\sqrt{n}}$

$$n = \frac{z_{crit}^2 . \sigma^2}{D^2} \qquad (5.8)$$

Example 5.4

It is required to choose a sample from a stream of industrial waste water. The standard deviation of the percent TDS is known to be 3. At a confidence level of 95%, how many specimens should we take so that the sample mean percent would not deviate from the population mean by more than 1?

Solution:

Here: $\sigma = 3$, $D = 1$ and $z_{crit} = 1.96$

Substituting in equation (5.7), we get:

$n = \frac{1.96^2 \times 3^2}{1^2} = 34.57 \approx$ **35 specimens**

5.4 Sampling of proportions

In many cases, the most important factor is the proportion of successes in any case at hand rather than their number. For example, it is known that factories can tolerate a certain number of defective items in their production line, provided it does not exceed a certain ratio.

In general, the proportion of "success" in the population will be denoted π whereas the proportion of failure $\tau = 1 - \pi$.

For example, in a production line of porcelain ware the proportion of first grade products is 0.4. So, $\pi = 0.4$ and $\tau = 0.5$.

If now we take a sample of say, 5 specimens, then all odds are possible. That is the number of first grade product can take any value from 0 to 5. We denote the proportion of success (first grade) in sample by *p* and failure by *q*. The likelihood of any proportions of successes will follow a binomial distribution as shown by the following table:

Table 5.2: Binomial distribution of success probability in a sample

No of successes	No of failures	*p*	*q*	Probability $f(x_i)$	$p.f(x_i)$	$x_i^2.f(x_i)$
0	5	0	1	$(0.6)^5$	0	0
1	4	0.2	0.8	${}^5C_1 \times 0.6^4 \times 0.4$	0.05184	0.010368
2	3	0.4	0.6	${}^5C_2 \times 0.6^3 \times 0.4^2$	0.13824	0.055296
3	2	0.6	0.4	${}^5C_3 \times 0.6^2 \times 0.4^3$	0.13824	0.082944
4	1	0.8	0.2	${}^5C_4 \times 0.6 \times 0.4^4$	0.06144	0.049152
5	0	1	0	0.4^5	0.01024	0.01024
Total				**1**	**0.4**	**0.208**

Note that the total sum of probabilities is 1 and the mean value of proportions of success in different samples is 0.4, which is equal to the proportion of successes in population (π).

On the other hand, the standard deviation of success proportions in samples can be obtained from equation (1.8):

$$\sqrt{\Sigma\, x_i^2 . f_i - \bar{x}^2} = \sqrt{0.208 - 0.4^2} = 0.219$$

Now, the standard deviation of population σ_p of a random variable for a binomial distribution was obtained from $\sqrt{npq} = \sqrt{\mu q}$. In case the variable is a proportion, this equation has to be rewritten to include the mean value of population proportion π (instead of μ) and the complementary proportion τ instead of q.

$$\sigma_p = \sqrt{\pi . \tau} \qquad (5.9)$$

Now, from equation (5.2), the standard deviation of sample means is:

$$\sigma_s = \sqrt{\pi . \tau / n} \qquad (5.10)$$

In the example at hand, $\sigma_s = \sqrt{\dfrac{0.6 \times 0.4}{5}} = 0.219$

The foregoing example yields two important results that coincide with those previously enunciated through equations (5.1) and (5.2). That is, the mean value of sample proportion means is equal to the proportion of successes in population and their standard deviation can be obtained by an equation analogous to (5.2), namely (5.10).

The standard deviation of sample proportion means is also termed the **standard error of sample proportion.**

If, now we select a sample of *n* items out of a large population, the population success proportion will not be known. A **point estimate of population mean** will be the proportion of successes in sample (*p*) and a **point estimate for standard error** of sample proportion can be obtained from the following equation:

$$s_p = \sqrt{\frac{p.q}{n-1}} \qquad (5.11)$$

For a given confidence interval, the range of values of the mean population proportion of successes can be obtained from:

$$p - z_{crit}.s_p < \pi < p + z_{crit}.s_p \qquad (5.12)$$

Example 5.5

A sample of 100 specimens was taken out of many products. It was found that 7 of these specimens were defective. What is the estimate for the mean proportion of defective items in the production line at a 95% confidence level?

Solution:
The point estimate of the proportion of defective items p = 0.07 and the standard error of sample is:

$$s_p = \sqrt{\frac{0.07 \times 0.93}{100 - 1}} = 0.0256$$

Hence, from equation (5.12):

$0.07 - 0.025 \times 61.96 < \pi < 0.07 + 0.025 \times 61.96$ or $\mathbf{0.02 < \pi < 0.12}$

This means that, based on the sample chosen, we are 95% sure that the proportion of defective items in the production line will range from 0.02 to 0.12, that is 2% to 12%.
We note that this range is somewhat wide. Had we taken a sample of 500 specimens, the previous result would have been: $0.048 < \pi < 0.092$, which represents a better estimate.

In case of proportions, we must take into consideration the effect of sample size as previously done under (5.3.4). In that case, the equation corresponding to equation (5.8) that determines the sample size required obtaining a maximum deviation $|p - \pi|$ is:

$$n = z_{crit}^2 \pi\tau / D \qquad (5.13)$$

If π and τ are not known, the best approximation is to use a sample estimate for p and q.

Example 5.6
In the preceding example, what is the size sample required to get a maximum deviation between the mean proportion of success in population and its sample estimate of 0.01?

Solution:
π and τ being unknown, we take p = 0.07 and q = 0.93, whereas D = 0.01 and for a 95% confidence level, z_{crit} = 1.95.

Replacing in equation (5.13), we get:
$n = (1.96)^2 \times 0.07 \times 0.93 / (0.01)^2 = \mathbf{2500}$

In that case, we get: $0.06 < \pi < 0.08$

5.5 Exercise problems

(1) The concentration of salt in the effluent stream of a reactor was measured for 35 samples. The mean value was 0.4 g mole/dm^3 and the standard deviation = 0.02. At a 95% confidence level, what is the probable mean value of the salt in the effluent stream?

(2) 100 PVC pipes were tested for bending strength. The mean value of samples was found to be 8.5 MPa and the standard deviation = 0.5 MPa. What are the

lower and upper limits of the population mean, at 90% and at 95% confidence level?

(3) The following table shows the marks of a sample of 30 students in an exam (attended by 2500 students). At a 95% confidence level, what would be the probable limits of the mean value of population?

Class interval	[0 – 5)	[5 – 10)	[10 – 15)	[15 – 20)	[20 – 25)	[25 – 30]
Number of	2	4	11	7	5	1

(4) The following data represent the level of solid dust concentration (mg/m^3) gathered from 10 different locations inside a plant. At a 95% confidence level, what would be the probable mean value of dust concentration inside this plant?

56	73	66	52	60	58	44	44	50	61

(5) It is required to choose a sample from the production of an oil well. The standard deviation of the percent sulfur in the crude is known to be 0.3. At a confidence level of 95%, how many samples should we take so that the mean percent sulfur in sample would not deviate from that of population mean by more than 0.05?

(6) A sample of 200 specimens was taken out of a production line. It was found that 20 of these specimens were defective. What is the estimate for the mean proportion of defective items in the production line at a 95% confidence level?

(7) A factory produces articles for exportation. However, due to process problems, the proportion of non-conforming items in a 300 specimens sample was found to be 15%. What would be your estimate for the proportion of non-conforming articles in the whole production on a 90% confidence level?

(8) A sample of bottled edible oil taken from the market showed that a proportion of 0.12 of the products are outdated. What would be a reasonable sample size such that the sample proportion represents the population with a maximum error of 0.01? Use a 95% confidence level.

(9) A batch storage tank dispenses on the average batches of volume 0.8 m^3 and standard deviation of 0.1 m^3. The volumes dispensed are assumed to follow a normal distribution of means. How many batches should be chosen so as to ensure with 90% confidence that its mean value would not deviate by more than 1% of the true mean? If it is intended to take 50 samples, what significance level would guarantee that the samples mean would not deviate from the true mean by more than 1%?