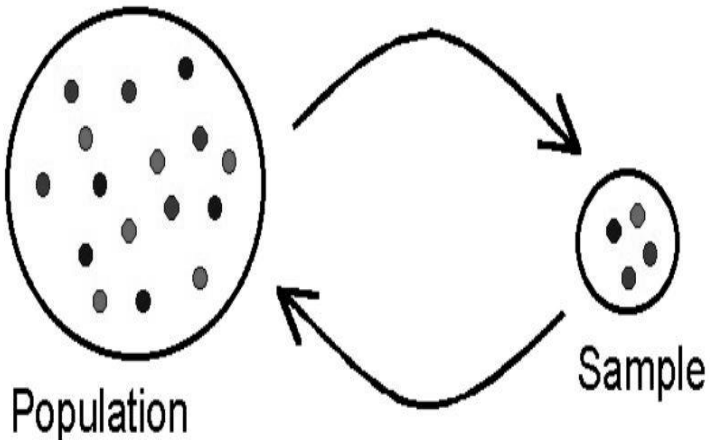
Sampling Distributions And Hypothesis Testing

5.1 Introduction

Sampling is a statistical method of obtaining representative data (observations) from a group. We have been using sampling concepts in our day to day lives knowingly or unknowingly; for instance we take a handful of rice to check the rice quality of the full lot. This is an example of random sampling from a large population.



Population (Universe): The group of objects (individuals) under study is called population or universe. Universe may be finite or infinite.

Sample: A part containing objects (individuals), selected from the population is called a sample.

Random Sampling: The selection of objects (individuals) from the universe in such a way that each object (individual) of the universe has the same chance of being selected is called random sampling. Lottery system is the most common example of random sampling.

Simple Sampling: Simple sampling is a special case of random sampling in which each event has same probability of success or failure.

Note: Every random sampling need not be simple. For example if balls are drawn without replacement from a bag of balls containing different balls; the probability of success changes in every trial. Thus the sampling though random is not simple.

Hypothesis: A hypothesis is an assumption based on insubstantial evidences that lends itself to further testing and experimentation. For example a farmer claims significant increase in crop production after using a particular fertilizer and after a season of experimenting, his hypothesis may be proved true or false. Any hypothesis may be accepted or rejected as per specific confidence levels and must be admissible to refutation.

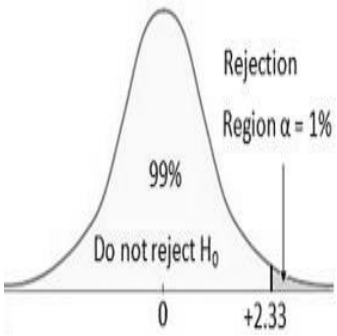
Null Hypothesis: A hypothesis which is tested for possible rejection under the assumption of being true is known as null hypothesis. Usually the null hypothesis is stated as 'There is no relationship between two quantities'. It is denoted by H_0.

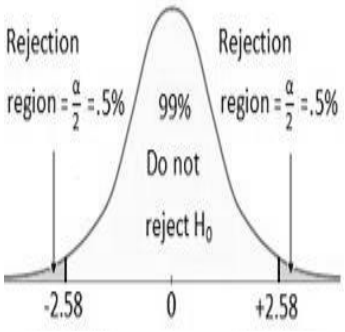
Alternative Hypothesis: It is the opposite statement of null hypothesis and denoted by H_1.

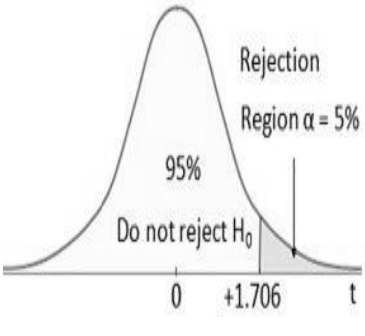
Significance levels(α)**:** The probability levels below which we reject a hypothesis H_0 are called levels of significance. Most common significance levels employed in hypothesis testing are $\alpha = 5\%$, $\alpha = 1\%$, $\alpha = 0.27\%$ in which critical (rejection) regions occupy 5% , 1% and 0.27% areas of normal curve respectively.

One Tailed and Two Tailed Tests: While testing statistical significance levels; one-tailed test and a two-tailed test are used for accepting or rejecting a hypothesis. One-tailed tests are used for asymmetric distributions (reference value is unidirectional) which have a single tail; such as the chi-square distribution.

A two-tailed test is appropriate if the estimated value may lie on both sides of reference value. Two-tailed tests are only applicable when the probability curve has two tails; such as normal distribution.







One-tailed test; $\alpha = 1\%$ Two-tailed test; $\alpha = 1\%$ One-tailed test; $\alpha = 5\%$

Reference *z* values (two- tailed) for various levels are given in the table below:

Significance level	Confidence level	z value
0.27 %	99.73%	3.00
1.00%	99.00%	2.58
4.55%	95.45%	2.00
5.00%	95.00%	1.96
9.89%	90.11%	1.65

Testing at1% level of confidence means that assumption has 99% chance of being true as rejection zone is just 1% of the total area under the normal curve. Similarly testing at 5% would provide 95% confidence level of acceptance.

Type I and Type II Errors: When we test a statistic at specified confidence level, there are chances of taking wrong decisions due to small sample size or sampling fluctuations etc.
Type I error is the incorrect rejection of a true null hypothesis, i.e. we reject H_0, when it is true.

Type II error is the incorrect acceptance of a false null hypothesis, i.e. we accept H_0, when it is false.

Testing of Hypothesis:

Testing of statistical hypothesis is a procedure designed for accepting or rejecting a hypothesis on the basis of some preset values.

Step1: Plant Null Hypothesis H_0 and alternate hypothesis H_1(optional) where H_0 is the hypothesis of no difference, i.e. H_0 presumes that there is no significant difference between observed value and expected value.

Step2: Find the most befitting test statistic for the analysis.

Step3: Take a random sample and compute the test statistic.

Step4: H_0 is accepted if the value of test statistic lies in acceptance zone and rejected if it falls in critical (rejection) region at the desired significance level.

5.2 Sampling Distributions

A sampling distribution is a distribution of all of the possible values of a statistic; computed from randomly drawn samples of the same size from a population.

Some commonly used notations in sampling distributions are given below:

	Population	**Sample**
Size	N	n
Mean	μ	$\overline{x}$
Variance	σ^2	s^2
Standard Deviation	σ	s

Suppose we take various samples each of size n from a population. If p and q be the probabilities of success and failure of each member of the sample, then the binomial distribution given by $(p+q)^n$ provides the sampling distribution of the number of successes in the sample with mean np and variance npq.

- Mean (expected value) of number of successes $= np$
- Standard deviation $= \sqrt{npq}$.
- Probable occurrence range at 99.73% confidence level i.e. 0.27% significance level is given by: $np \pm 3\sqrt{npq}$
- Probable occurrence range at 99% confidence level i.e. 1% significance level is given by: $np \pm 2.58\sqrt{npq}$
- Probable occurrence range at 95% confidence level i.e. 5% significance level is given by: $np \pm 1.96\sqrt{npq}$

In case of proportion of successes, mean and standard deviation of proportion of successes are obtained by dividing each statistic by n.

- Mean (expected value) of proportion of successes $= \frac{np}{n} = p$
- Standard deviation $= \frac{\sqrt{npq}}{n} = \sqrt{\frac{pq}{n}}$
- Probable occurrence range of the proportion at 99.73% confidence level i.e. 0.27% significance level is given by: $p \pm 3\sqrt{\frac{pq}{n}}$
- Probable occurrence range of the proportion at 99% confidence level i.e. 1% significance level is given by: $p \pm 2.58\sqrt{\frac{pq}{n}}$
- Probable occurrence range of the proportion at 95% confidence level i.e. 5% significance level is given by: $p \pm 1.96\sqrt{\frac{pq}{n}}$

Standard Error: The standard deviation of the sampling distribution of a statistic is known as Standard Error (S.E.).
Precision: Reciprocal of standard error is known as precision.

Probable Error: It is taken as 0.67449 times the standard error and is used sometimes to explain the concept of sampling errors to layman or unprofessional people.

5.3 Sampling of Attributes for large samples ($n > 30$)

Characteristics like language, religion, habits (traits) etc. cannot be measured in numbers as they are attributes. Sampling of attributes means testing how many in a population possess a particular attribute (trait) or whether the two populations share an attribute (trait) in common and to how much confidence level.

When sample size (n) is very large i.e. greater than 30 and neither p nor q are very small, the binomial distribution tends to normal distribution and therefore we choose the variate z as test statistic.

Following procedure is adopted for testing the significance of large samples in terms of attributes.

Step1: Postulate the null hypothesis (H_0); if required.

Step2: If x is the observed number of successes in a sample and z is the standard normal variate, then $z = \frac{x-\mu}{\sigma}$, i.e. $z = \frac{x-np}{\sqrt{npq}}$

Step3: Accept or reject H_0 as per given values of parameter z at different significance levels:

Significance Level	H_0 accepted	H_0 rejected
0.27%	$\|z\| < 3$	$\|z\| > 3$
1%	$\|z\| < 2.58$	$\|z\| > 2.58$
5%	$\|z\| < 1.96$	$\|z\| > 1.96$

In case of absence of any specified significance level, we may consider 0.27% level, i.e. take acceptance range as $-3 < z < 3$.

Example 1 A coin is tossed 400 times and turns up head 216 times. Discuss whether the coin may be unbiased one.

Solution: Let H_0: coin is unbiased

Here $n = 400$, if p denotes probability of success; i.e. getting a head,

then $p = \frac{1}{2}$, $q = \frac{1}{2}$, ∴ Expected number of heads is $np = 400 \times \frac{1}{2} = 200$

Again if x denotes observed number of heads; then the test statistic

$$z = \frac{x-np}{\sqrt{npq}}; \text{ i.e. } z = \frac{216-200}{\sqrt{400\times\frac{1}{2}\times\frac{1}{2}}} = 1.6 < 1.96 < 2.58$$

∴ H_0 is acceptable at both 5% and 1% levels of significance.

Hence our assumption that coin is unbiased is true

Example2 A die is rolled 900 times and turns up five or six 338 times. Discuss whether the die may be unbiased one.

Solution: Let H_0: die is unbiased

Here $n = 900$, if p denotes probability of success; i.e. getting 5 or 6,

then $p = \frac{1}{3}$, $q = \frac{2}{3}$, ∴ Expected number of successes is $np = 900 \times \frac{1}{3} = 300$

Again if x denotes observed number of successes; then the test statistic

$$z = \frac{x-np}{\sqrt{npq}} \text{ ; i.e. } z = \frac{338-300}{\sqrt{900\times\frac{1}{3}\times\frac{2}{3}}} = 2.69$$

$\therefore$ $2.69 < 3$ but $2.69 > 1.96$ and also $2.69 > 2.58$

Hence H_0 is acceptable at 0.27% level of significance but fails at both 1% and 5% significance levels.

Example3 A random sample of 500 oranges was taken from a large consignment and 65 were found to be bad. Show that the standard error of the proportion of bad ones in a sample of this size is 0.015 and deduce that the percentage of bad oranges in the consignment lies between 8.5 and 17.5.

Solution: Let p denote the proportion of bad oranges in the given sample

Then $p = \frac{65}{500} = 0.13$, $q = 1 - p = 0.87$

Standard Deviation (S.D.) $= \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.13\times0.87}{500}} = 0.015$

$\therefore$ Standard Error (S.E.) = S.D. $= 0.015$

Probable limit of bad oranges in the consignment is given by:

$$p \pm 3\sqrt{\frac{pq}{n}} = 0.13 \pm 3(0.015)$$

$= 0.085$ to 0.175

$\therefore$ Probable percentage of bad oranges in the consignment is 8.5% to 17.5%

Example4 A random sample of 100 bolts was taken from the lot manufactured by a machine and 10 were found to be defective. Find the 95% confidence limits for the proportion of defective bolts produced by the machine.

Solution: Let p denote the proportion of defective bolts in the given sample

Then $p = \frac{10}{100} = 0.1$, $q = 1 - p = 0.9$

Standard Deviation (S.D.) $= \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.1\times0.9}{100}} = 0.03$

Probable limit of defective bolts in the lot at 95% confidence level is given by:

$$p \pm 1.96\sqrt{\frac{pq}{n}} = 0.1 \pm 1.96(0.03) = 0.1 \pm 0.0588 = 0.0412 \text{ to } 0.1588$$

$\therefore$ Probable percentage of proportion of defective bolts in the lot at 95% confidence level is 4.12% to 15.88%

Example5 A sample of 900 days is taken from metrological records of a district and 100 of them are found to be foggy. What is the probable percentage of foggy days in the district?

Solution: Let p denote the probability of a foggy day in the district, then

$$p = \frac{100}{900} = \frac{1}{9},\ q = 1 - p = \frac{8}{9}$$

$$\text{Standard Deviation (S.D.)} = \sqrt{npq} = \sqrt{900 \times \frac{1}{9} \times \frac{8}{9}} = 9.43$$

Probable limit of foggy days at 99.73% confidence level is given by:

$np \pm 3\sqrt{npq} = 100 \pm 3(9.43) = 100 \pm 28.29$; i.e. 71.71 to 128.29

Probable percentage is $\frac{71.71}{900} \times 100$ to $\frac{128.29}{900} \times 100$ i.e. 7.97% to 14.25%

5.3.1 Comparing Proportions of Large Samples from Two Different Populations in Terms of Attributes

For an attribute (trait), let proportions p_1 and p_2 be given from two large samples of sizes n_1 and n_2 respectively from two different populations. We may want to test whether two populations are similar regarding the specified attribute (trait).

Working methodology:

1. Set up the hypothesis H_0: The two populations are similar regarding the specified attribute (trait)
2. Find common proportion of two populations for the specified attribute as: $p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2},\ q = 1 - p$
3. Compute combined standard error of two populations as: $e = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$
4. Compute the statistic parameter $z = \frac{p_1 - p_2}{e}$
5. Accept or reject H_0 as per given values of parameter z at different significance levels:

Significance Level	H_0 accepted	H_0 rejected
0.27%	$\|z\| < 3$	$\|z\| > 3$
1%	$\|z\| < 2.58$	$\|z\| > 2.58$

5%	$\|z\| < 1.96$	$\|z\| > 1.96$

Example 6 In a sample of 600 men from a certain city, 450 are found to be smokers. In another sample of 700 men from another city, 450 are found smokers. Do the data indicate that the cities are significantly different with respect to the habit of smoking among men?

Solution: Let H_0: The two populations are similar regarding smoking habits among men.

Here $n_1 = 600$, $p_1 = \frac{450}{600} = 0.75$

$n_2 = 900$, $p_2 = \frac{450}{700} = 0.64$

$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{600 \times 0.75 + 700 \times 0.64}{600 + 700} = \frac{449}{650} = 0.69$

$q = 1 - p = 1 - 0.69 = 0.31$

$\therefore\ e = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{0.69 \times 0.31\left(\frac{1}{600} + \frac{1}{700}\right)} = 0.026$

$z = \frac{p_1 - p_2}{e} = \frac{0.75 - 0.64}{0.026} = 4.23$

$|z| > 3$, $\therefore$ H_0 is rejected, i.e. the difference between two populations is highly significant in terms of smoking habits among men both at 1% and 5% levels of significance.

Example 7 In a large city A , 20% of a random sample of 900 school boys had defective eye-sight. In another large city B, 15.5% of a random sample of 1600 school boys had defective eye-sight. Is the difference between two proportions significant?

Solution: Let H_0: The two populations are similar regarding school boys having defective eye sight.

Here $n_1 = 900$, $p_1 = \frac{20}{100} = \frac{1}{5} = 0.2$

$n_2 = 1600$, $p_2 = \frac{15.5}{100} = \frac{31}{200} = 0.155$

$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{900 \times 0.2 + 1600 \times 0.155}{900 + 1600} = 0.1712$

$q = 1 - p = 1 - 0.1712 = 0.8288$

$\therefore\ e = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{0.1712 \times 0.8288\left(\frac{1}{900} + \frac{1}{1600}\right)} = 0.01568$

$$z = \frac{p_1 - p_2}{e} = \frac{0.2 - 0.155}{0.01568} = 2.8699$$

$|z| < 3$, $\therefore$ H_0 is accepted at 0.27% level of significance, i.e. The difference between two populations is not significant in terms of defective eye-sights.

Remark: H_0 should be rejected at 1% or 5% significance levels.

Example 8 Following data gives proportion of dark coloured people in two cities.

City	Sample size	Percentage of dark coloured people
A	250	42
B	450	48

Can the difference between two percentages taken as sampling fluctuations?

Solution: Let H_0: The two populations are similar regarding smoking habits among men.

Here $n_1 = 250$, $p_1 = 42\%$

$n_2 = 450$, $p_2 = 48\%$

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{250 \times 42 + 450 \times 48}{250 + 450} = 45.86\%$$

$$q = 1 - p = 54.14\%$$

$$\therefore e = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{45.86 \times 54.14\left(\frac{1}{250} + \frac{1}{450}\right)} = 3.93\%$$

$$z = \frac{p_1 - p_2}{e} = \frac{42 - 48}{3.93} = -1.53$$

$|z| < 1.96$, $\therefore$ H_0 is accepted, i.e. The difference between two populations is not significant in terms of dark coloured people both at 1% and 5% levels of significance.

5.4 Sampling Distributions of Sample means

The sampling distribution of the mean refers to the pattern of sample means, observed by different samples drawn from the population at large.

Result *I* : If all possible samples of size n are drawn without replacement from a finite population of size N and if μ and σ denote population mean and standard deviation respectively;

$\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$ denote mean and standard deviation respectively of sampling distribution, then $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$

Result *II* : If all possible samples of size n are drawn with replacement from a finite population of size N and if μ and σ denote population mean and standard deviation respectively;

$\mu_{\overline{x}}$ and $\sigma_{\overline{x}}$ denote mean and standard deviation respectively of sampling distribution, then $\mu_{\overline{x}} = \mu$ and $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$

Example 9 Suppose a population consists of five numbers: 1,3,5,7,9. Find sampling distribution of sample means. Also calculate mean and standard deviation of the sample means, if random samples of two numbers are drawn

$i.$ without replacement $ii.$ with replacement

Verify Results I and II by comparing these statistics with population mean and population standard deviation.

Solution: Population mean $(\mu) = \frac{1}{N}\sum_1^N x_i$

$\mu = \frac{1}{5}[1 + 3 + 5 + 7 + 9] = 5$

Population variance (σ^2)$= \frac{1}{N}\sum_1^N (x_i - \mu)^2$

$\sigma^2 = \frac{1}{5}[(1-5)^2 + (3-5)^2 + 0 + (7-5)^2 + (9-5)^2] = \frac{40}{5} = 8$, $\therefore \sigma = \sqrt{8}$

$i.$ Total number of possible samples each of size two without replacement is ${}^5C_2 = 10$

Sample	Sample mean ($\overline{x}$)
1,3	2
1,5	3
1,7	4
1,9	5
3,5	4
3,7	5
3,9	6
5,7	6
5,9	7
7,9	8

Sampling distribution of Sample mean ($\overline{x}$)

$\overline{x}$	2	3	4	5	6	7	8
Frequency	1	1	2	2	2	1	1

Mean of sample means ($\mu_{\overline{x}}$) $= \frac{1}{n}\sum_1^n \overline{x}$

$= \frac{1}{10}[2 + 3 + 2(4) + 2(5) + 2(6) + 7 + 8] = 5$

Variance of sample means

$\sigma_{\overline{x}}^2 = \frac{1}{10}[(2-5)^2 + (3-5)^2 + 2(4-5)^2 + 0 + 2(6-5)^2 + (7-5)^2 + (8-5)^2] = \frac{30}{10} = 3$ Standard deviation of sample means ($\sigma_{\overline{x}}$) $= \sqrt{3}$

$\therefore\ \mu_{\overline{x}} = 5 = \mu$, also $\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{\sqrt{8}}{\sqrt{2}}\sqrt{\frac{5-2}{5-1}} = \sqrt{3} = \sigma_{\overline{x}}$

Hence Result *I* is verified.

ii. Total number of possible samples each of size two with replacement is $5 \times 5 = 25$

Samples of 2 numbers with replacement are shown below; individual sample mean is given under each sample.

(1,1) 1	(1,3) 2	(1,5) 3	(1,7) 4	(1,9) 5
(3,1) 2	(3,3) 3	(3,5) 4	(3,7) 5	(3,9) 6
(5,1) 3	(5,3) 4	(5,5) 5	(5,7) 6	(5,9) 7
(7,1) 4	(7,3) 5	(7,5) 6	(7,7) 7	(7,9) 8
(9,1) 5	(9,3) 6	(9,5) 7	(9,7) 8	(9,9) 9

Mean of sample means is given by:

$\mu_{\overline{x}} = \frac{1}{25}[1 + 2(2) + 3(3) + 4(4) + 5(5) + 4(6) + 3(7) + 2(8) + 9] = 5$

Variance of sample means is given by:

$\sigma_{\overline{x}}^2 = \frac{1}{25}[(1-5)^2 + 2(2-5)^2 + 3(3-5)^2 + 4(4-5)^2 + 5(5-5)^2 +$

$4(6-5)^2 + 3(7-5)^2 + 2(8-5)^2 + (9-5)^2]$

$= \frac{1}{25}[16 + 18 + 12 + 4 + 0 + 4 + 12 + 18 + 16] = \frac{100}{25} = 4$

$\therefore$ Standard deviation of sample means ($\sigma_{\overline{x}}$) $= 2$

$\mu_{\overline{x}} = 5 = \mu$ and $\sigma_{\overline{x}} = 2 = \frac{8}{\sqrt{2}} = \frac{\sigma}{\sqrt{n}}$, hence Result *II* is verified.

Result *III* : If a population is normally distributed with mean and variance μ and σ^2 respectively; also $\overline{x}$ denotes sample mean for a random sample (x_1, x_2,⋯, x_n); then sampling distribution of $\overline{x}$ is also normally distributed with mean μ and variance $\frac{\sigma^2}{n}$, i.e. If $x \sim N(\mu, \sigma^2)$ then $\overline{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

Proof: Since the population is normally distributed, for any object x_i of the population;

$E(x_i) = \mu$ and $Var(x_i) = \sigma^2$

$$\therefore E(\overline{x}) = E\left(\frac{x_1 + x_2 + \cdots + x_n}{n}\right)$$

$$= E\left(\frac{x_1}{n}\right) + E\left(\frac{x_2}{n}\right) + \cdots + E\left(\frac{x_n}{n}\right)$$

$$= \frac{1}{n}\mu + \frac{1}{n}\mu + \cdots + \frac{1}{n}\mu = \frac{n\mu}{n} = \mu$$

Also $Var(\overline{x}) = Var\left(\frac{x_1 + x_2 + \cdots + x_n}{n}\right)$

$$= Var\left(\frac{x_1}{n}\right) + Var\left(\frac{x_2}{n}\right) + \cdots + Var\left(\frac{x_n}{n}\right)$$

$$= \frac{1}{n^2}\sigma^2 + \frac{1}{n^2}\sigma^2 + \cdots + \frac{1}{n^2}\sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$$\because Var(ax_i) = a^2 Var(x_i)$$

$\therefore$ Standard deviation of sampling distribution ($\sigma_{\overline{x}}$) $= \frac{\sigma}{\sqrt{n}}$

Standard Error: The standard deviation of the sampling distribution is called the standard error $\therefore$ Standard error of sampling distribution is $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$

Remark: We have assumed here that σ is known. However if σ is not known, we take σ to be equal to the standard deviation of the sample.

Properties of the sampling distribution of sample means:

- A sample drawn from a normally distributed population follows normal distribution,

  $\therefore$ Z-value for the distribution of $\overline{x}$ is given by $\frac{\overline{x} - \mu}{\sigma_{\overline{x}}} = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

- The mean $\mu_{\overline{x}}$ of the sample means will be the same as population mean from which the samples were drawn, i.e. $\mu_{\overline{x}} = \mu$
- The variance $\sigma_{\overline{x}}^2$ of the sampling distribution of $\overline{x}$ will be equal to the variance of the population divided by the sample size i.e. $\sigma_{\overline{x}}^2 = \frac{\sigma^2}{n}$

Result *IV* : Central Limit Theorem: As the sample size gets large enough (30 or higher); the sampling distribution becomes approximately normal regardless of shape of population.

Remark: For large samples ($n \geq 30$), probability distribution is taken as normal for computational purposes.

Example 10 A population has mean 0.1 and standard deviation 2.1. Find the probability that the mean of a random sample of size 900 will be negative.

Solution: Given that population mean $(\mu) = 0.1$ and standard deviation $(\sigma) = 2.1$

Since the sample size is large enough, sampling distribution is approximately normal with mean 0.1 and standard deviation $= \frac{\sigma}{\sqrt{n}} = \frac{2.1}{\sqrt{900}} = 0.07$

i.e. $\overline{x} \sim N\left(0.1, \frac{(2.1)^2}{900}\right)$ and $z = \frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\overline{x}-0.1}{\frac{2.1}{30}} \quad \Rightarrow \overline{x} = 0.1 + 0.07z$

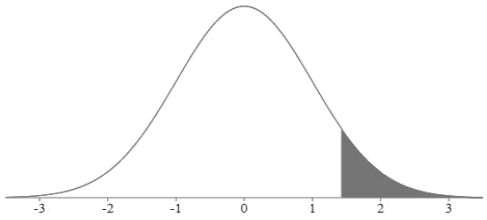
$$\therefore \quad P(\overline{x} < 0) = P(0.1 + 0.07z < 0)$$

$$= P\left(z < \frac{-0.1}{0.07}\right)$$

$$= P(z < -1.43)$$

$$= P(z > 1.43)$$

$$= 0.5 - 0.4236 = 0.0764$$



Example 11 Suppose a population has mean 10 and variance 4. What is the probability that the sample of size 36 has mean lying between 9.8 and 10.2?

Solution: Given that population mean $(\mu) = 10$ and standard deviation $(\sigma) = 2$. Since the sample size is large enough, sampling distribution is approximately normal with mean 10 and standard deviation $= \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{36}} = 0.33$

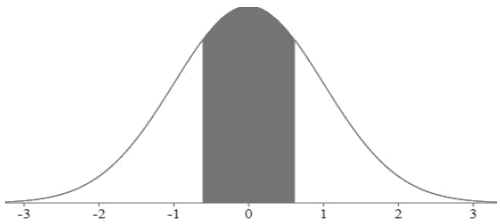
i.e. $\overline{x} \sim N\left(10, \frac{4}{36}\right)$ and $z = \frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\overline{x}-10}{\frac{2}{6}} \quad \Rightarrow \overline{x} = 10 + 0.33z$

$$\therefore \quad P(9.8 < \overline{x} < 10.2)$$

$$= P(9.8 < 10 + 0.33z < 10.2)$$

$$= P\left(\frac{9.8-10}{0.33} < z < \frac{10.2-10}{0.33}\right)$$

$$= P(-0.61 < z < 0.61)$$



$= 0.2291 + 0.2291 = 0.4582$

Example 12 A firm produces electric bulbs that have normally distributed mean burning life 800 hours with a standard deviation of 40 hours. Find the probability that a random sample of 16 bulbs will have average burning life of less than 775 hours.

Solution: Given distribution is normal with population mean $(\mu) = 800$ and standard deviation $(\sigma) = 40$, sample size $(n) = 16$

$$z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{x}-800}{10}$$

$$\Rightarrow \bar{x} = 800 + 10z$$

$$\therefore \quad P(\bar{x} < 775) = P(800 + 10z < 775)$$

$$= P\left(z < \frac{775-800}{10}\right)$$

$$= P(z < -2.5)$$

$$= P(z > 2.5) \quad \text{By symmetry of standard normal curve}$$

$$= 0.5 - P(0 < z < 2.5) = 0.5 - 0.4938 = 0.0062$$

Result *V* : Test of significance of sample mean for a large sample

Here we test the hypotheses about the sample mean of a population in the case when sample size is at least 30 (for large samples, distribution is approximately normal as per central limit theorem). Let μ and σ be mean and variance of a population. If $\bar{x}$ be the mean of a simple large sample of size n from an unknown population, we may want to test whether the sample belongs to given population.

Working rule:

1. Set up the hypothesis H_0: the sample belongs to given population
2. Compute the statistic parameter $z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$
3. Accept or reject H_0 as per given values of parameter z at different significance levels:

Significance Level	H_0 accepted	H_0 rejected
0.27%	$\|z\| < 3$	$\|z\| > 3$
1%	$\|z\| < 2.58$	$\|z\| > 2.58$
5%	$\|z\| < 1.96$	$\|z\| > 1.96$

Remark: The statistic parameter $z = \frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$ can also be used to check whether the sample taken from the given population is random or not.

Example 13 A sample of size 900 is having mean 3.6 mm; could it be reasonably regarded as a random sample from large population whose mean is 3.35 mm and standard deviation 2.6 mm at 1% significance level.

Solution: Let H_0: sample belongs to the given population

Here population mean $\mu = 3.35$mm, population standard deviation $\sigma = 2.6$ mm,

also sample size $n = 900$ and sample mean $\overline{x} = 3.6$ mm

Now $z = \frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{3.6-3.35}{\frac{2.6}{\sqrt{900}}} = \frac{3.6-3.35}{2.6}(30) = 2.88 > 2.58$

As $|z| > 2.58$, H_0 is rejected, i.e. difference between sample mean ($\overline{x}$) and population mean (μ) is significant at 1% level and the sample may not be considered from the same population.

Example 14 A random sample of 400 canes of vegetable oil with labeled net weight as 5kg has a mean net weight of 4.98 kg with a standard deviation of 0.22 kg. Can we accept the hypothesis of net weight 5 kg per cane on the basis of given sample, at 5 % level of significance?

Solution: Let H_0: Mean net weight of each tin is 5 kg.

Here population mean $\mu = 5$kg, sample size $n = 400$ and sample mean $\overline{x} = 4.98$ kg,

sample standard deviation $s = 0.22$kg

Now $z = \frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{4.98-5}{\frac{0.22}{\sqrt{400}}} = \frac{4.98-5}{0.22}(20) = -1.81$

As $|z| < 1.96$, H_0 is accepted at 5% significance level, i.e. Mean net weight of each tin is 5 kg.

Note: Here population standard deviation σ is not known, $\therefore$ sample standard deviation is taken as population standard deviation.

Example 15 A pharmaceutical company fills its best-selling 8 ounce jars of liquid medicine by an automatic dispensing machine. The machine is set to dispense a mean of 8.1 ounces per jar with a standard deviation of 0.22 ounce. Uncontrollable factors in the process can shift the mean away from 8.1 and cause either under-fill or overfill, both of which are undesirable. In such a case the dispensing machine is stopped and recalibrated. A quality control engineer routinely selects 30 jars from the assembly line to

check the amounts filled. On one occasion, the sample mean is 8.2 ounces with the standard deviation of 0.25 ounce. Determine if there is sufficient evidence in the sample to indicate, that the machine should be recalibrated, at the 5% level of significance.

Solution: Let H_0: machine need not be recalibrated

Here population mean $\mu = 8.1$ ounce, population standard deviation

$\sigma = 0.22$ ounce.

Also sample size $n = 30$ and sample mean $\bar{x} = 8.2$ ounce

$$\therefore z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{8.2-8.1}{\frac{0.22}{\sqrt{30}}} = \frac{8.2-8.1}{0.22}\left(\sqrt{30}\right) = 0.45 \times 5.48 = 2.47$$

As $|z| > 1.96$, H_0 is rejected, i.e. difference between sample mean ($\bar{x}$) and population mean (μ) is significant at 5% level and the machine needs to be recalibrated.

Result *VI* : Test of significance of difference between means of two large samples

(a) If samples are from two different populations

Let $\bar{x}_1$ and $\bar{x}_2$ be sample means of two populations of sizes n_1 and n_2 from two different normally distributed populations having means μ_1; μ_2 and variances σ_1^2 ; σ_2^2 respectively;

i.e. $\bar{x}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right)$ and $\bar{x}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$

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

Also $Var(\bar{x}_1 - \bar{x}_2) = Var(\bar{x}_1) + Var(\bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

$\therefore$ Standard Error $(e) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

z-value for the distribution of $\bar{x}_1 - \bar{x}_2$ is given by $z = \frac{(\bar{x}_1-\mu_1)-(\bar{x}_2-\mu_2)}{e}$

(b) If samples are from same population

If $\bar{x}_1$ and $\bar{x}_2$ be two sample means of sizes n_1 and n_2 taken from a normally distributed population having mean μ and variance σ^2

Then $E(\bar{x}_1 - \bar{x}_2) = E(\bar{x}_1) - E(\bar{x}_2) = \mu - \mu = 0$

Also $Var(\bar{x}_1 - \bar{x}_2) = Var(\bar{x}_1) + Var(\bar{x}_2) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}$

$\therefore$ Standard Error $(e) = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}$

z-value for the distribution of $\bar{x}_1 - \bar{x}_2$ is given by $z = \frac{\bar{x}_1 - \bar{x}_2}{e}$

Example 16 A random sample of 150 villages was taken from a district A having standard deviation 32 and average population per village was found to be 440. Another random sample of 250 villages from district B with a standard deviation of 56 gave an average population of 480 per village. Is the difference between the averages of two populations significant? Give reasons.

Solution: Let H_0: The differences between averages of two populations is not significant,

i.e. $\mu_1 = \mu_2$

Here $n_1 = 150$, $\bar{x}_1 = 440$, $\sigma_1 = 32$

$n_2 = 250$, $\bar{x}_2 = 480$, $\sigma_2 = 56$

$$e = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{32^2}{150} + \frac{56^2}{250}} = 4.4$$

$$\therefore z = \frac{(\bar{x}_1 - \mu_1) - (\bar{x}_2 - \mu_2)}{e} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{e}$$

$= \frac{440-480}{4.4} = -9.09$, under the assumption $\mu_1 = \mu_2$

$|z| \gg 3$, thus differences between two averages is highly significant.

$\therefore$ H_0 is rejected and the averages of two populations cannot be taken as same.

Example 17 Random samples of 500 and 400 are having means 11.5 and 10.9 respectively. Can the two samples be regarded as drawn from the population of standard deviation 5?

Solution: Let H_0: The two samples are drawn from the same population with standard deviation $(\sigma) = 5$

Here $n_1 = 500$, $\bar{x}_1 = 11.5$

$n_2 = 400$, $\bar{x}_2 = 10.9$

$$e = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} = \sqrt{\frac{25}{500} + \frac{25}{400}} = 0.335$$

$$\therefore z = \frac{\bar{x}_1 - \bar{x}_2}{e} = \frac{11.5 - 10.9}{0.335} = 1.79$$

$|z| < 1.96$, $\therefore$ H_0 is accepted at 5% level of significance, i.e. the two samples can be considered to be drawn from same population.

Example18 50 new entrants in a class are found to have an average height of 135cm and 30 old one have an average height of 140cm with a class standard deviation of 8cm. Does this indicate that mean height of old students is greater than that of new entrants?

Solution: Let H_0: Average height of new entrants is same as old students

Here standard deviation of the class is $\sigma = 8$

$$n_1 = 50\,, \quad \overline{x}_1 = 135$$

$$n_2 = 30\,, \quad \overline{x}_2 = 140\,,$$

$$e = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} = \sqrt{\frac{64}{50} + \frac{64}{30}} = 1.85$$

$$\therefore z = \frac{\overline{x}_1 - \overline{x}_2}{e} = \frac{135-140}{1.85} = -2.7$$

$|z| < 3$, $\therefore$ H_0 is accepted at 0.27% significance level, i.e. average height of new entrants is same as old students.

Remark: H_0 will have to be rejected at 1% and 5% levels of significance.

Exercise 5

1. A coin is tossed 400 times and head turns up 225 times. Discuss whether the coin is biased or unbiased at 5% level of significance.
2. A random sample of 600 oranges was taken from a large consignment and 60 were found to be rotten. Show that the standard error of the proportion of bad ones in a sample of this size is 0.1 and deduce that the percentage of bad oranges in the consignment almost lies between 6.3 and 13.7
3. In a city *A* 20% of a random sample of 900 school children wore spectacles and in another city *B* 18.5% of a random sample of 1600 school children used to wear spectacles. Is the difference between the proportions significant?
4. In a sample of 500 people from a state 280 take tea and rest take coffee. Can we assume that tea and coffee are equally popular in the state?
5. A sample of 900 members is found to have a mean of 3.4cm. Can it be reasonably regarded as truly random sample from a large population with mean 3.25cm and S.D. 1.61cm.

6. A sample of 100 electric bulbs produced by a manufacturer M_1 showed a mean life time 1190 hours with a standard deviation of 90 hours. Another sample of 75 bulbs produced by manufacturer M_2 showed a mean life time 1230 hours with a standard deviation of 120 hours. Is there a difference between the mean life times of two brands at 5% level of significance?
7. The means of two large samples of 1000 and 2000 members are 168.75 cm and 170cm respectively. Can these be regarded as drawn from the same population of standard deviation 6.25 cm.
8. A stenographer states that he can take dictation at the rate of 120 words per minute. Can we accept his claim on the basis of 100 trials in which he showed a mean of 116 words with standard deviation of 15 words ?
9. A sample of height of 6400 soldiers has a mean of 67.85 inches and a standard deviation of 2.56 inches, while a random sample of heights of 1600 sailors has a mean of 68.55 inches and a standard deviation of 2.52 inches. Does this indicate that the sailors are on average taller than the soldiers?
10. A random sample of 400 students has an average weight of 55 kg. Can we say that the sample comes from a population with mean 58 kg. with a variance of 9 kg. ?
11. In a big city two samples of people are drawn. First sample of size 100, the average daily income of people is 210$ with a standard deviation 10$ and in the second sample of size 150 persons, average daily income is 220$ with a standard deviation of 11$. Test if there is any significant difference in average incomes.

Answers

1. Biased
3. $z = 0.37$ ∴ the difference is not significant
4. $z = 6.56$ ∴ the difference is highly significant
5. $z = 2.8$ ∴ it cannot be regarded as a random sample.
6. Yes
7. No
8. $z = 2.67$ ∴ The claim is not acceptable
9. Yes
10. No
11. $|z| = 7.14$ ∴ the difference is highly significant.