Statistics &Data Analysis Concepts for Data Science and ML 7

7

Sampling and Sampling Distributions

Learning Objectives

The main objective of this chapter is for you to gain an understanding of different sampling techniques and the importance of sampling distribution in data analysis. After completing this chapter, you should be able to:

- **Understand the importance of sampling and different types of samples including the probability and non-probability sampling techniques**
- **Learn the important sampling techniques including the simple random, systematic, stratified, cluster, judgment sampling, and other sampling methods**
- **Understand the concepts of standard error for both finite and infinite populations**

Learning Objectives…cont.

- **Understand the concept of sampling distribution and how it is used to draw conclusions about the population**
- **Learn the sampling distributions of two statistics – sample mean (x) and sample proportion (p)**
- **Learn the importance of central limit theorem and describe the distribution of sample means using this theorem**
- **Understand the distribution of sample proportion and solve problems involving proportions**

Sampling: Basic Concepts

- A ***population*** denotes the entire measurements that are theoretically possible (or the universe).
- A ***sample*** is a part of the population. In most statistical studies, we collect sample data to draw a conclusion about the population.
- ***Sampling*** is a systematic way of selecting a few items from the population.
- The *purpose of sampling* is to draw a conclusion or make a decision about the population parameters using the information contained in the sample statistics
- A population is described by its parameters known as ***population parameters***
- A sample is described by its statistics known as ***sample statistics***

Population Parameters and Sample Statistics

The ***population parameters*** are:

μ: the population mean

N: the population size

σ^2: the population variance

σ: the population standard deviation

A sample is described by its statistic. The ***sample statistics*** are:

$\bar{x}$: the sample mean

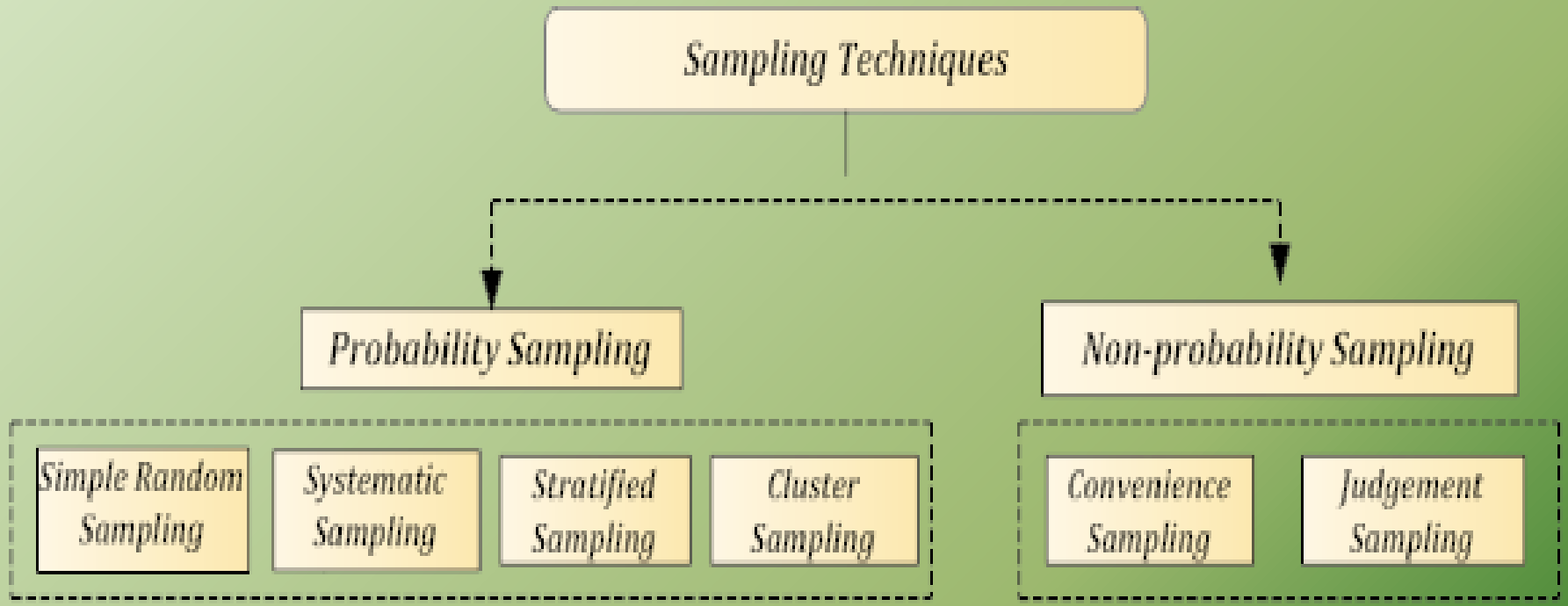
n : the sample size

s^2 : the sample variance

s : the sample standard deviation

Sample statistics are used to estimate population parameters. Note that in most cases, we don't know the population parameters (for example, the population mean, μ) and therefore, the population mean, μ must be estimated using the sample statistic, $\bar{x}$.

Sampling Techniques



Sampling Techniques…cont.

***Simple Random Sampling** is a sampling technique in which every sample has equal probability of being selected and every item in the sample has equal probability of being selected.*

Example 1:Draw all possible samples each of size n = 2 from a population of N=5. The five items in the population are identified by A, B, C, D, E.

All possible samples of size n=2 can be determined using the combination formula:

$$\binom{N}{n} = \frac{N!}{n!(N-n)!} = 10$$

Note N=5, n=2

The 10 possible samples are listed below:

AB BD AC BE AD CD AE CE BC DE

- Probability of each sample being selected is 1/10 or, P (A B) = P(AC) = - - - = P(DE)= 1/10
- Probability that each item in the sample will be selected is:

P(A) =4/10 or 40%. Similarly, P(B) = 40% = P (C) = P(D) = P(E)

The above results satisfy both conditions of simple random sampling (every sample has equal probability of being selected and every item in the sample has an equal probability of being selected).

Drawing Random Samples

The random samples are drawn from a population of interest

- ***using a table of random numbers (computer generated), or***
- ***using computer software such as, EXCEL or MINITAB***

Selecting a set of random numbers using the random number table

Select a random sample of 15 checking accounts from a list of 1000 accounts where the accounts are numbered sequentially from 1 to 1000 using the of random numbers shown on the next page.

To draw a random sample of 15 (n=15) from a population of 1000 accounts (N=1000), follow the steps below.

- **Start in any position, such as row 8 and column 3 of the random number table (next slide)**
- **Select a list of 15 random numbers by reading either across the row or down the column.**

Table of Random Numbers

Row	Col1	Col2	Col3	Col4	Col5	Col6	Col 7	Col8
1	12785	99985	46746	51452	50406	61229	63378	86436
2	40951	11508	76859	58093	20288	11376	20435	82228
3	15208	21791	78527	61821	32491	73946	36750	30036
4	25209	5633	822	55887	88145	13777	8927	44126
5	50556	79612	92213	35982	28416	78567	18095	29057
6	4562	15620	13855	79475	79669	17115	23956	62566
7	10316	34927	82992	27844	56148	11981	80583	98368
8	48962	8931	**77251**	16916	33614	7187	86386	12402
9	36287	52928	70684	41051	23529	3722	9069	72202
10	83418	90623	78304	4007	52705	82505	21244	36507
11	71480	52958	62868	55371	85155	38450	92304	36651
12	29945	24758	86671	77813	44851	85208	13148	99626
13	3665	75510	85123	75042	67234	8711	83912	20031
14	5463	41437	61181	63693	85952	19573	96066	76250
15	5204	94929	36820	12299	7427	65740	96865	72323
.								
.								

Read across the row, start in row 8 and column 3 and the first set of five digit numbers is 77251. Place a decimal between the third and fourth digits and round this value to the nearest integer. For our example, the number would be 772.51, which would become 773 after rounding. Thus, account number 773 is the first account selected for our sample.

Drawing Random Samples...continued

The next set of five digits is 16916, which would be 169 after rounding. Continuing in this manner, a sample of 15 accounts selected for our sample would be

773, 169, 336, 719, 864, 124, 363, 529, 707, 411, 235, 372 907, 722, and 834

Another way of selecting a random sample of 15 accounts:

Select three-digit numbers, reading across the row, and starting arbitrarily in any row. Suppose we start again in row 8 and column 3 of the random number table on the previous page and select three digits, reading across the row. The three digit numbers are listed below.

772, 511, 691, 633, 614, 718, 786, 386, 124, 023, 628, 752,928, 706, and 844

Drawing Random Samples using EXCEL

Use EXCEL to generate random samples, using the random number generation option. To generate 50 random checking accounts from a total 1000 accounts, follow the steps below.

From the EXCEL main menu select,

Tools → Data Analysis → Random Number Generation
Click OK

The Random Number Generation dialog box will be displayed. Complete the dialog box as shown below.

Number of Variables	**1**
Number of Random Numbers	**50**
Distribution	**Uniform** (select from the drop down list)
Between	**1** and **1000**
Output Range	Click anywhere on the worksheet where you want to store the results
Click	**OK**.

This will generate 50 random numbers from uniform distribution and store the results in the column you specified. You should round the generated values

Other Probability Sampling Techniques

Systematic Sampling: ***In a systematic sampling,***

- ***the samples are drawn at a pre-specified number or at some pre-specified time***
- ***the N items in the population are partitioned into m groups by dividing the size of the population N by the desired sample size n. That is,***

$$m = \frac{N}{n}$$

where, m is rounded to the nearest integer. To obtain a systematic sample, the first item to be selected is chosen at random from m items in the first partitioned group in the population frame, and the rest of the samples are obtained by selecting every m^{th} item thereafter from the entire population frame listing. This method is more convenient and practical than simple random sampling.

Stratified Sampling

In this method of drawing samples:

- **the population is divided into different groups or *strata*, according to some common characteristic (e.g., department, age, income level, industry type, location, etc.).**
- **a simple random sample is taken from each group and the results from the separate simple random samples are then combined.**

This technique is more efficient than simple random sampling or systematic sampling because it ensures representation of individuals or items across the entire population, which in turn ensures a greater precision in estimating the population parameters.

Cluster Sampling

In cluster sampling:

- ***the N individuals or items in the population are divided into several clusters, such that each cluster is representative of the entire population.***
- ***a random sampling of clusters is then taken and all individuals or items in each selected cluster are then studied.***

Some applications are: area sampling, where clusters are city blocks or other well defined areas.

Sampling Distribution

Sampling distribution is the probability distribution of a sample statistic (sample statistic may be a sample mean $\overline{x}$ ***, a sample variance*** s^2***, a sample standard deviation s, or sample proportion*** $\overline{p}$***).***

In most cases, the true value of the population parameters are not known. We must draw a sample or samples and calculate the sample statistic to estimate the population parameter.

The sampling error of the sample mean is given by

$$\text{Sampling Error} = \overline{x} - \mu$$

Samples are taken to draw a conclusion about the parameters of the population. For example, to draw a conclusion about the mean of certain population, we would collect samples from this population, calculate the mean of the samples, and determine the probability distribution (shape) of the sample means. This probability distribution may follow a normal or a t-distribution. The distribution will then be used to draw conclusion about the population mean.

Sampling Distribution of the Sample Mean and Sample Proportion

Sampling distribution of the sample mean $\bar{x}$ ***is the probability distribution of all possible values of the sample mean,*** $\bar{x}$.

Sampling distribution of sample proportion, $\bar{p}$ ***is the probability distribution of all possible values of the sample proportion,*** $\bar{p}$.

The process of sampling distribution for the sample mean, $\bar{x}$ ***is illustrated below:***

Population with mean, μ to be estimated

↓

Select several samples of size n (for example, 50 samples each of size n=5)*

↓

Calculate the sample mean, $\bar{x}$ of all the samples and do a probability distribution of $\bar{x}$

Example 3: Examining the Distribution of the Sample Mean, $\overline{x}$

The assembly time of a particular electrical appliance is assumed to have a mean, $\mu = 25$ minutes, and a standard deviation, $\sigma = 5$ minutes.

(1) Draw 50 samples each of size 5 (n=5) from this population using MINITAB.

(2) Determine the average or the mean of each of the samples drawn.

(3) Draw a histogram of the sample means and interpret your findings.

(4) Determine the average and standard deviation of the 50 sample means. Interpret the meaning of these.

(5) What conclusions can you draw from your answers to (3) and (4)?

Solution to (1): Table 7.1 (a) on the next slide shows 50 samples each of size 5 using MINITAB (other statistical package can be used).

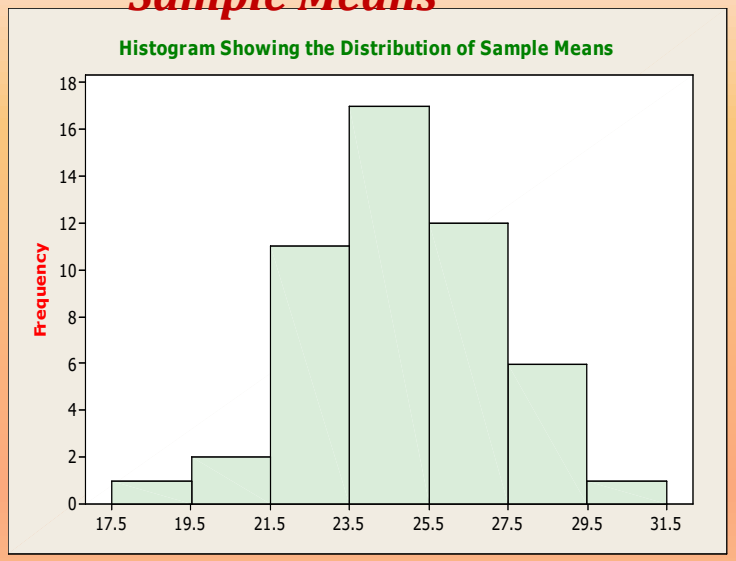
Solution to (2): The last column (Table 7.1, next slide) shows the mean of each sample drawn. Note that each row represents a sample of size 5.

Table 7.1: 50 Samples of Size n=5

Sample #						Sample Mean $\bar{x}$
1	24.13	26.53	33.99	17.09	23.39	25.03
2	26.55	30.49	30.57	26.83	28.46	28.58
3	15.39	26.33	23.04	21.12	26.82	22.54
4	15.99	27.09	27.73	24.95	21.90	23.53
5	26.16	26.16	21.37	36.40	27.25	27.47
6	26.68	26.55	26.72	26.24	26.31	26.50
7	30.37	15.32	25.11	24.10	31.68	25.32
8	31.23	24.27	33.72	30.80	25.31	29.06
9	24.87	16.56	31.46	31.51	16.64	24.21
10	26.14	31.61	25.19	24.10	17.42	24.89
11	25.91	20.27	23.67	28.76	23.38	24.40
12	22.42	19.28	22.35	27.72	31.13	24.58
13	23.87	25.11	27.19	30.79	23.85	26.16
14	32.50	34.10	32.14	27.76	22.86	29.87
15	28.30	26.64	30.33	19.87	25.09	26.05
16	24.08	24.40	28.41	30.83	8.00	23.14
17	21.34	21.17	27.31	30.24	34.34	26.88
18	20.42	19.11	16.03	19.80	17.07	18.48
19	25.87	25.46	17.18	23.54	26.71	23.75
20	18.21	18.68	27.04	27.35	21.66	22.59
21	30.36	18.51	24.01	24.35	28.54	25.15
:						
:						
50	26.01	24.35	21.94	16.89	23.73	22.58

Solution to (3): Figure 7.1 shows the histogram of the sample means shown in the last column of Table 7.1. The histogram shows that the sample means are normally distributed. Figure 7.1 is an example of the sampling distribution of the sample means .

Figure 7.1: Sampling Distribution of the Sample Means



Solution to (4): The mean and standard deviation of the sample means shown in the last column of Table 7.1 were calculated. These values are shown below.

Mean and Standard Deviation of Sample Means

Descriptive Statistics: Sample Mean	
Mean	StDev
24.978	2.285

The mean of the sample means is 24.98, which indicates that $\overline{x}$ values are centered at approximately the population mean, $\mu = 25$. However, the standard deviation of 50 sample means is 2.285, which is much smaller than the population standard deviation, $\sigma = 5$. Thus, we conclude that $\overline{x}$ — or, the sample mean values — have much less variation than the individual observations.

Solution to (5): Based on parts (3) and (4), we conclude that the sample means, $\bar{x}$ follows a normal distribution, and this distribution is much narrower than the population of individual observations, which has a standard deviation, $\sigma = 5$. This is apparent from the standard deviation of $\bar{x}$ value, which is 2.285 (see Table 7.2). In general, the mean and standard deviation of the random variable $\bar{x}$ are given by

Mean of the sample mean, $\bar{x}$ is $\mu_{\bar{x}} = \mu$

The standard deviation of the sample mean, $\bar{x}$ is $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$

For our example, $\mu = 25$, $\sigma = 5$, and n=5. Using these values

$$\mu_{\bar{x}} = \mu = 25 \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{5}} = 2.236$$

Standard Deviation of The Sample Mean or The Standard Error

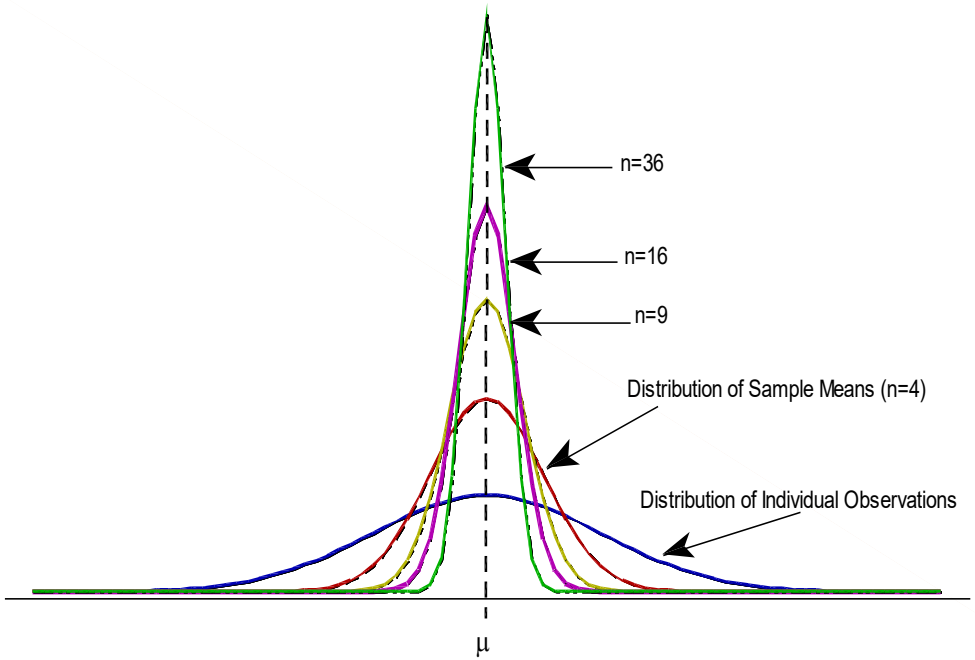
The standard deviation of the sample mean $\sigma_{\bar{x}}$ is often called the standard error of the mean and is given by

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

This equation shows that the standard deviation of the sample means, $\bar{x}$ (or the sampling distribution of the random variable $\bar{x}$) varies inversely as the square root of the sample size. Since the standard deviation of the mean is a measure of scatter of the sample means, it provides the precision that we can expect of the mean of one or more samples.

Figure 7.2 (next slide) shows a comparison between the probability distribution of individual observations and the probability distributions of means of samples of various sizes drawn from the underlying population. Note that as the sample size increases, the standard error becomes smaller and hence the distribution becomes more peaked.

Figure 7.2: Probability Distribution of Sample Means (n=4, 9, 16, and 36) Compared to Individual Observations



As more samples are taken, the standard error decreases, thus providing greater precision.

Standard Error

Two different formulas are used to calculate the standard error or the standard deviation of the sample mean . Standard deviation of $\bar{x}$, $\sigma_{\bar{x}}$ *or the standard error*

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

for an infinite population

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

for a finite population

Example 3: Standard Error

Suppose, for a population, the standard deviation, σ = 25. Calculate the standard error when the sample size, n = 50, 100, 150, and 200. Comment on your results.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{25}{\sqrt{50}} = 3.54$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{25}{\sqrt{100}} = 2.50$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{25}{\sqrt{150}} = 2.04$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{25}{\sqrt{200}} = 1.77$$

The standard error decreases as the sample size n, increases.

Example 4: Standard Error

(a) Suppose, the population standard deviation, σ = 10 and the sample size, n = 50. Calculate the standard error.

Solution

$$\sigma_{\bar{x}} = \frac{10}{\sqrt{50}} = 1.41$$

(b) Suppose, n=50, $\sigma = 50, N = 50,000$. Calculate the standard error.

Solution

The standard error can be calculated using the formula for the finite population.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$ where, $\sqrt{\frac{N-n}{N-1}}$ is a finite population correction factor and if $\frac{n}{N} < 0.05$ do not use this factor.

The purpose of this factor is to reduce the error. But, if the sample size is too small compared to the population size, the factor becomes almost 1.0 and does not help to reduce the error. For our example:

$$\frac{n}{N} = \frac{50}{50,000} = 0.001 < 0.05$$ therefore, $$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{50}} = 1.41$$

SAMPLING FROM DIFFERENT POPULATIONS: SAMPLING DISTRIBUTION OF THE SAMPLE MEAN

Draw samples of various sizes from different distributions (normal, exponential, uniform, etc.), calculate the means of all the samples, and do a probability distribution of the sample means, $\bar{x}$.

Objective: **The purpose of this exercise is to observe the shape of the sampling distribution. We would like to know what happens to the shape of the distribution when**

(a) the samples are drawn from different populations, and

(b) the sample size increases.

The results will lead to a very important conclusion that is critical to solving problems involving sampling distribution. The following distributions are considered for demonstration purpose.

The following distributions are considered for demonstration purpose.

(1) Normal (2) Exponential (3) Uniform

Drawing Samples From Different Distributions and Observing The Shape of The Distribution of The Sample Means

Steps

Draw 100 samples of various sizes [n=1, n=5, n=20, n=35, and n=50] from each of the above populations (we used MINITAB to do this)

In each case, calculate the mean $(\bar{x})$ of the samples and do the sampling distribution of the sample means $(\bar{x})$

Investigate the shape of distribution of the sample means, $(\bar{x})$

Sampling from a Normal Distribution

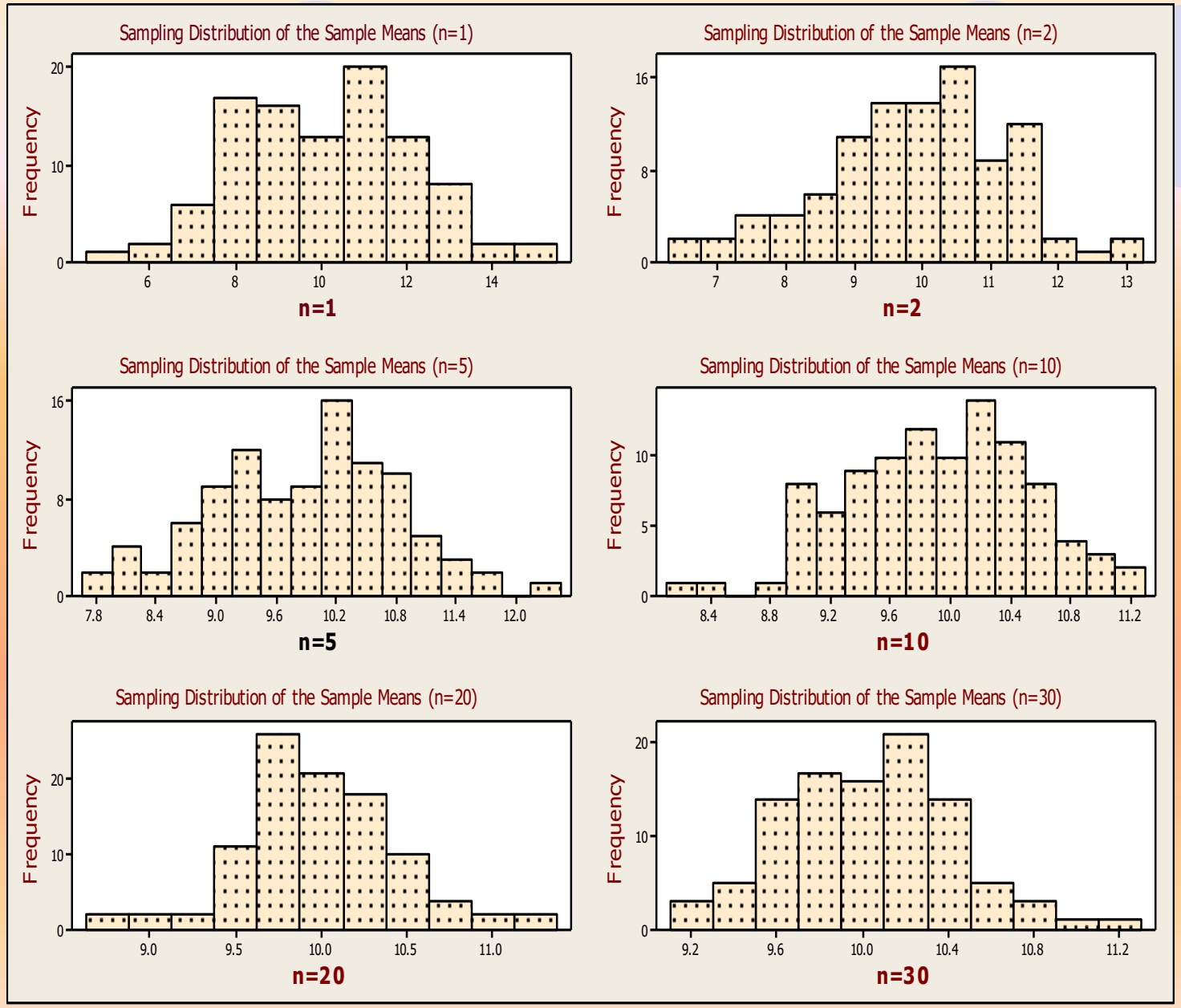
- *Using a statistical package such as, MINITAB, generate 100 samples of size n=1, n=2, n=5, n=10, n=20, n=30, and n=50 from the normal distribution with mean μ=10 and standard deviation σ=2.*

- *Calculate the sample mean for each of the samples and construct the histograms for the sample*

- *Calculate the descriptive statistics for each sample*

- *The histograms in Figure 7.3 (next two slides) show that if the samples are drawn from a normal population, the distribution of the sample means also follows a normal distribution*

- *Table 7.3 shows the descriptive statistics for different sample sizes. You can see how the standard error of the mean (SE Mean) decreases as we increase the sample size*


Sampling Distribution of the Sample Means (n=1)
Frequency
n=1
Sampling Distribution of the Sample Means (n=2)
Frequency
n=2
Sampling Distribution of the Sample Means (n=5)
Frequency
n=5
Sampling Distribution of the Sample Means (n=10)
Frequency
n=10
Sampling Distribution of the Sample Means (n=20)
Frequency
n=20
Sampling Distribution of the Sample Means (n=30)
Frequency
n=30

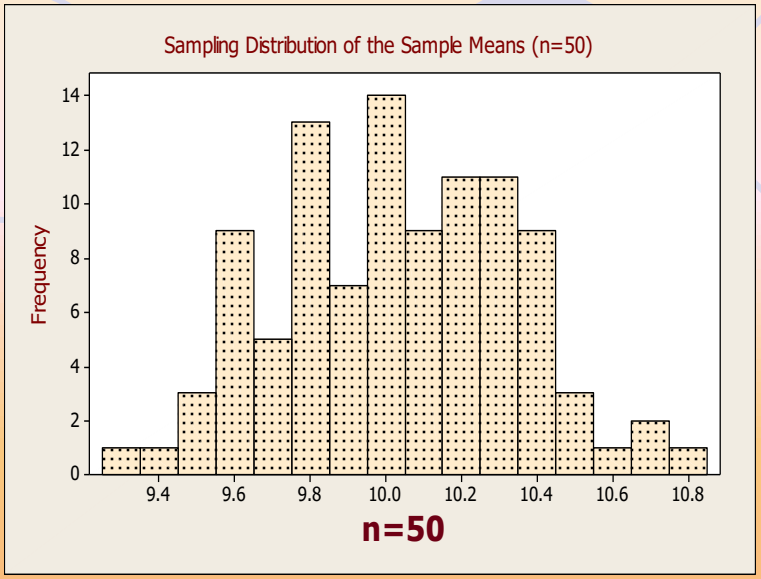


Figure 7.3: Sampling Distribution of the Sample Means (Data from a Normal Distribution)

Table 7.3: Descriptive Statistics for Various Sample Size from a Normal Distribution

Descriptive Statistics: n=1, n=2, n=5, n=10, n=20, n=30, n=50

Variable	N	Mean	Median	TrMean	StDev	SE Mean
n=1	100	10.100	10.129	10.111	1.777	0.178
n=2	100	10.119	10.004	10.133	1.454	0.145
n=5	100	9.9531	9.8935	9.9444	0.7886	0.0789
n=10	100	10.007	9.981	9.993	0.595	0.060
n=20	100	9.9978	9.9454	9.9897	0.4465	0.0447
n=30	100	9.9738	9.9252	9.9655	0.3717	0.0372
n=50	100	9.9946	9.9750	9.9916	0.2942	0.0294

Note how the standard error of the mean (SE Mean) decreases as we increase the sample size

Sampling from an Exponential Distribution

- **Using a statistical package, generate 500 samples of size n=1, n=2, n=5, n=10, n=20, n=30, and n=50 from an exponential distribution with mean =1.0**

- **Calculate the sample mean for each of the samples and construct the histograms for the sample means**

- **Calculate the descriptive statistics for each sample size**

- **Examine the shape of the distribution for different sample sizes. Figure 7.4 shows that the distribution of the sample mean follows a normal distribution as the sample size increases.**

- **Examine the descriptive statistics of the samples shown in Table 7.4. This table shows the descriptive statistics for various sample sizes. Note how the standard error of the mean (SE Mean) decreases as we increase the sample size.**

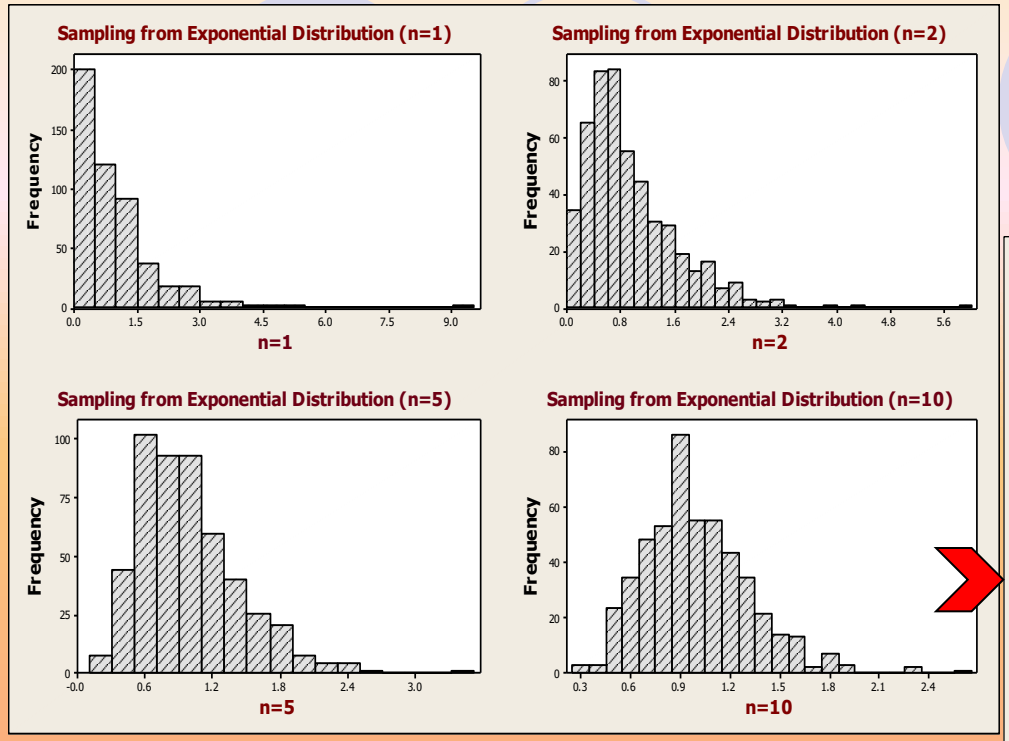


Figure 7.4: Sampling Distribution of the Sample Means (Data from Exponential Distribution)

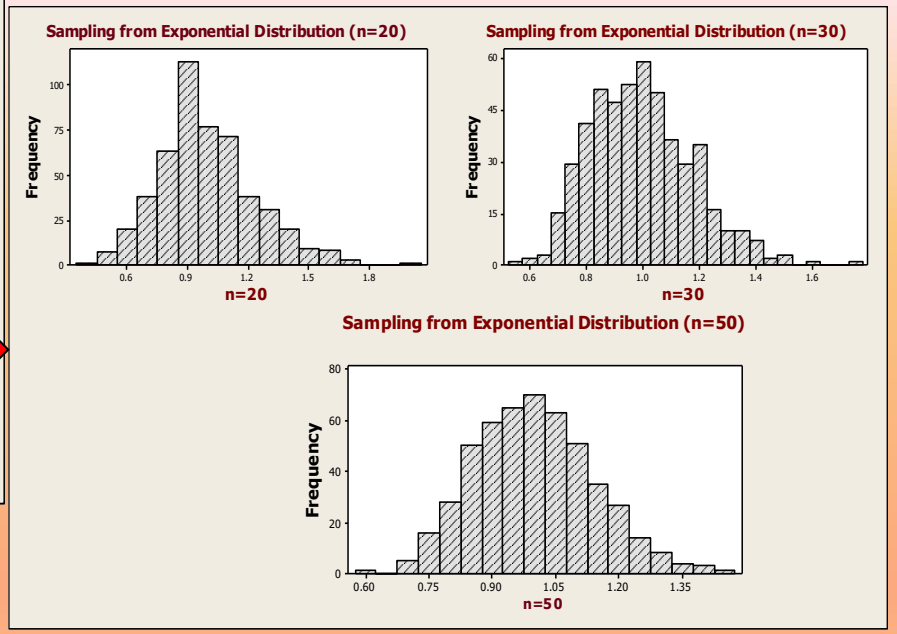


Table 7.4: Descriptive Statistics for Various Sample Size from an Exponential Distribution

Descriptive Statistics: n=1, n=2, n=5, n=10, n=20, n=30, n=50

Variable	N	Mean	Median	TrMean	StDev	SE Mean
n=1	500	1.0355	0.7279	0.9308	0.9957	0.0445
n=2	500	1.0450	0.9200	0.9935	0.6919	0.0309
n=5	500	1.0074	0.9403	0.9877	0.4140	0.0185
n=10	500	0.9912	0.9775	0.9811	0.3081	0.0138
n=20	500	0.9865	0.9614	0.9793	0.2325	0.0104
n=30	500	0.99554	0.98172	0.99102	0.18519	0.00828
n=50	500	0.99317	0.98350	0.99112	0.14716	0.00658

Sampling from a Uniform Distribution

- Using a statistical package, generate 500 samples of size n=1, n=2, n=5, n=10, n=20, n=30, and n=50 from a uniform distribution with a lower value of a=0.0 and upper value, b=1.0
- Calculate the sample mean for each of the samples and construct the histograms for the sample means
- Calculate the descriptive statistics for each sample size
- Examine the shape of the distribution for different sample sizes. Figure 7.5 shows that the distribution of the sample mean follows a normal distribution as the sample size increases.
- Examine the descriptive statistics of the samples shown in Table 7.5. This table shows the descriptive statistics for various sample sizes. Note how the standard error of the mean (SE Mean) decreases as we increase the sample size.

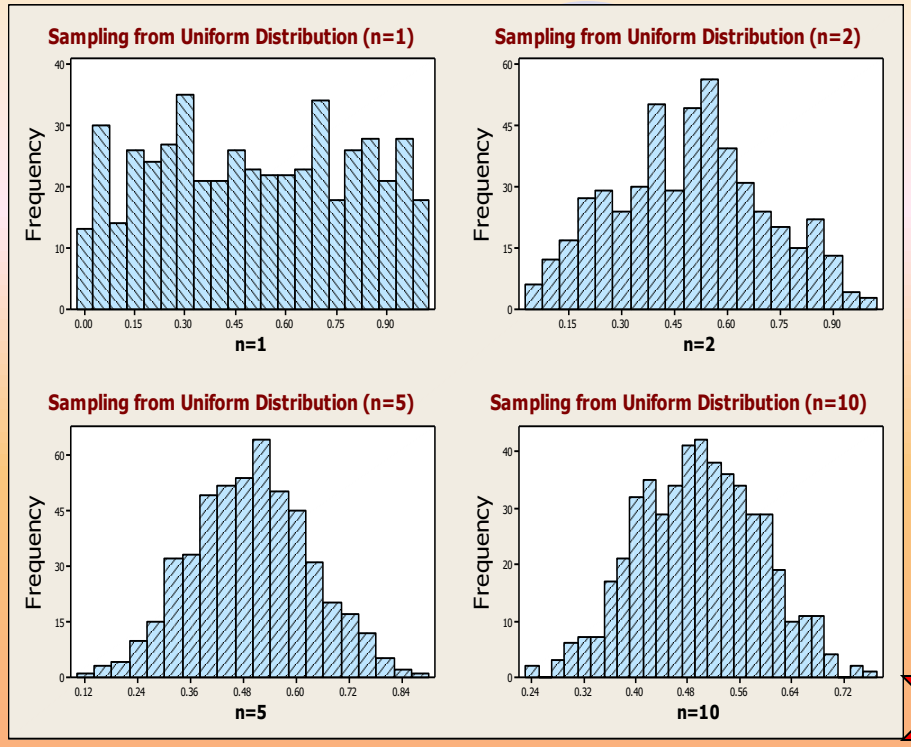


Figure 7.5: Sampling Distribution of the Sample Means (Data from a Uniform Distribution)

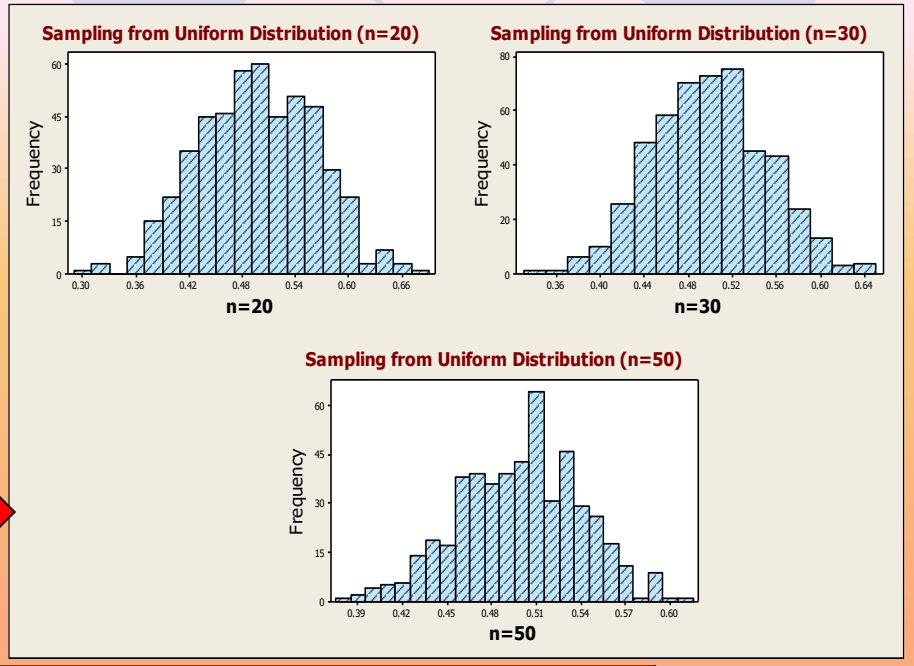


Table 7.5: Descriptive Statistics for Various Sample Size from a Uniform Distribution

Descriptive Statistics: n=1, n=2, n=5, n=10, n=20, n=30, n=50

Variable	N	Mean	Median	TrMean	StDev	SE Mean
n=1	500	0.4945	0.4960	0.4942	0.2839	0.0127
n=2	500	0.48960	0.48790	0.48946	0.21001	0.00939
n=5	500	0.49528	0.49549	0.49454	0.13143	0.00588
n=10	500	0.49464	0.49831	0.49499	0.09084	0.00406
n=20	500	0.49764	0.50435	0.49864	0.06236	0.00279
n=30	500	0.50120	0.50337	0.50155	0.04951	0.00221
n=50	500	0.50011	0.50041	0.49977	0.03984	0.00178

Conclusion from the previous examples: ***No matter what the shape of the distribution is, the distribution of the sample means ($\bar{x}$) approaches a normal distribution as the sample size increases. For a large sample (n ≥30), the shape is almost always symmetrical, i.e., follows a normal distribution.***

This leads to an important theorem in statistics known as ***Central Limit Theorem***

Central Limit Theorem (CLT) : As the sample size, n increases, the distribution of the sample mean approaches a normal distribution.

- **The Central Limit Theorem has been proclaimed as "the most important theorem in statistics"[1] and "perhaps the most important result of statistical theory."**
- **The Central Limit Theorem can be proven to show the "amazing result" that the mean values of the sum of a large number of independent random variables are normally distributed.**
- **The probability distribution resulting from "a large number of individual effects . . . would tend to be Gaussian."**

[1] **Ostle, Bernard and Mensing, Richard W., *Statistics in Research*, Third Edition, The Iowa State University Press, Ames, Iowa, 1979, p. 76.**

For a sample size of n ≥ 30 (large sample), we can always use the normal distribution to draw conclusions from the sample data.

For a large sample, the sampling distribution of the sample mean ($\bar{x}$) follows a normal distribution and **the probability that the sample mean ($\bar{x}$) is within a specified value of the population mean (μ) can be calculated using the following formulas:**

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$ **(for an infinite population)**

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}$$ **(for a finite population)**

Example 5

A population has a mean, μ = 100 and standard deviation, σ = 16. What is the probability that a sample mean will be within ± 2 of the population mean for each of the following sample sizes:

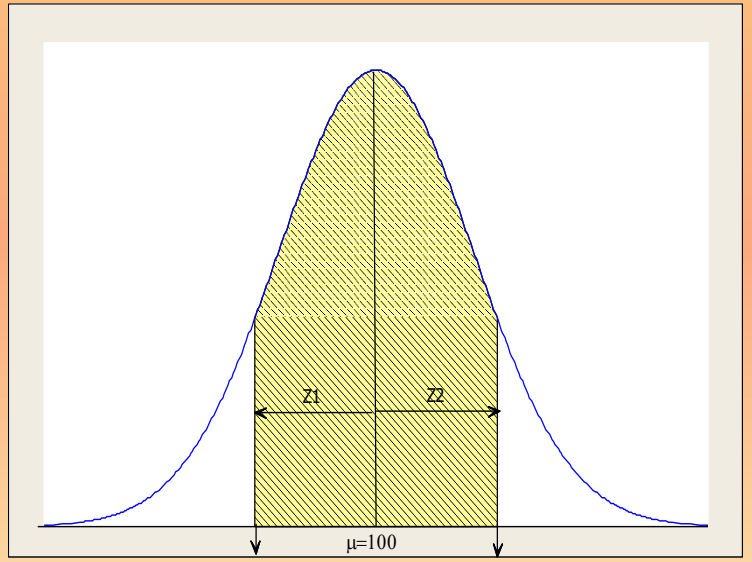
(a) n= 50 (b) n= 100 (c) n= 200 (d) n = 400

Solution:

$\mu_{\bar{x}} = 100$ $\sigma = 16$

$\bar{x} - \mu = \pm 2$ (or, the sample mean $\bar{x}$ is within 98 and 102.)

(a) n = 50. The required probability is the shaded area in the figure below.



$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{2}{16/\sqrt{50}} = 0.88$$

z = 0.88 is equivalent to an area of 0.3106 (from the standard normal table).

Therefore, the required probability = (2) (0.3106)

= 0.6212 or, 62.12%

or, there is 62.12% chance that the sample mean is within ±2 of the population mean.

This probability is also written as:

$$p(98 \le \bar{x} \le 102) = 0.6212$$

Alternate way of Solving Example 5

The probability can be evaluated as shown below (refer to the figure on page 35)

$$z_1 = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{98 - 100}{16/\sqrt{50}} = 0.88 \Rightarrow 0.3106$$

Note z = 0.88 is equivalent to an area of 0.3106 (obtained from the normal table)

$$z_2 = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{102 - 100}{16/\sqrt{50}} = 0.88 \Rightarrow 0.3106$$

Required probability = 0.3106 + 0.3106 = 0.6212 or, 62.12%

$$p(98 \le \bar{x} \le 102) = 0.6212$$

(b) Repeat the above problem for n= 100, n=200, n=400

n = 100 Required probability = 0.7888 or, 78.88%

n = 200 Required probability = 0.9232 or, 92.32%

n = 400 Required probability = 0.9876 or, 98.76%

You should verify these results

Note that larger the sample size n, higher the probability that the sample mean ($\bar{x}$) will be within ± 2 of μ.

Example 6

The mean price per gallon of regular gasoline sold in the U.S. was $1.20 (March 1997). Assume that the population mean price per gallon was = $1.20 and the population standard deviation was σ = 0.10. Suppose that a random sample of 50 gasoline stations is selected and a sample mean price per gallon is computed for the data collected from 50 gasoline stations.

[a] Show the sampling distribution of the sample mean $\bar{x}$ ***, where*** $\bar{x}$ ***is the sample mean price per gallon for 50 stations***.

μ= $1.20

σ = 0.10 n = 50 Therefore, $E(\bar{x}) = \mu = 1.20$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.10}{\sqrt{50}} = 0.014$

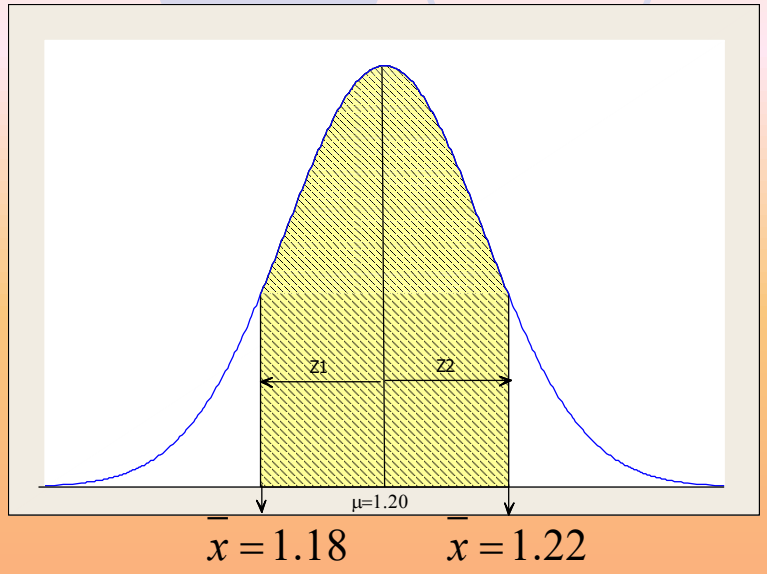
Note: the sampling distribution of $\bar{x}$ will follow a normal distribution as n > 30. The normal distribution is described by mean (μ) and $\sigma_{\bar{x}}$.

Here, $E(\bar{x}) = \mu$ and $\sigma_{\bar{x}} = \sigma / \sqrt{n}$

(b) What is the probability that the simple random sample of size 50 will provide a sample mean within 2 cents or 0.02 of the population mean?

To calculate the probability, refer to the figure on the next slide

Example 6...cont.



The required probability is the shaded area in figure. The probability can be calculated as,

$$z_1 = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{1.18 - 1.20}{0.10/\sqrt{50}} = -1.41 \Rightarrow 0.4207$$

$z_1 = -1.41$ corresponds to an area of 0.4207 (from the standard normal table)

$$z_2 = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{1.22 - 1.20}{0.10/\sqrt{50}} = +1.41 \Rightarrow 0.4207$$

Required probability = P (1.18 ≤ $\bar{x}$ ≤ 1.22) =0.4207 + 0.4207
=0.8414 or, 84.14%.

(c) What is the probability that the simple random sample will provide a sample mean within 1 cent or 0.01, of the population mean?

P (1.19 ≤ $\bar{x}$ ≤1.21) = 0.5224
= 52.24%

Calculations are similar to (b). You should verify the result.

Example 7

Suppose that a population contains the following 5 values (N=5):

2, 3, 5, 4, 6

Calculate the population mean. Then select all possible samples of size 2 (n=2) from this population. Calculate the mean of all samples. Next, calculate the mean of all sample means. Show that the mean of all sample means is equal to the population mean; that is, E ($\bar{x}$) = μ. What is this property called?

Solution: The population mean is:

$$\mu = \frac{\sum x}{N} = \frac{2+3+5+4+6}{5} = 4$$

All samples of size 2 (n=2) possible from this population of 5 (N=5) can be calculated using the combination formula

$$\binom{N}{n} = \frac{N!}{n!(N-n)!} = \frac{5!}{2!(5-2)!} = 10$$

Thus, there are 10 possible samples of size 2. These samples are listed below (note that 5 items in the population are 2, 3, 5, 4, 6).

Example 7…cont.

sample #	Sample of size 2 (n=2)	Mean of the sample ($\bar{x}$)
1	2, 3	2.5
2	2,5	3.5
3	2,4	3.0
4	2,6	4.0
5	3,5	4.0
6	3,4	3.5
7	3,6	4.5
8	5,4	4.5
9	5,6	5.5
10	4,6	5.0

$\bar{x}$

From this table, the mean of all sample means can be calculated as

$$E(\bar{x}) = \frac{2.5 + 3.5 + 3.0 + \ldots\ldots\ldots + 5.0}{10} = 4$$

This proves that $E(\bar{x}) = \mu$

This property is known as unbiasedness. It means that the sample mean is the unbiased predictor of the population mean.

Example 8

What is a ***point estimate***? What are the formulas for point estimates of a population mean, population standard deviation, and population proportion?

Solution:

The purpose of a point estimate is to estimate the value of a population parameter using a sample statistics. The population parameters are μ, σ , p etc.

Example 8…cont.

(a) The point estimate of the population mean (μ) is the sample mean ($\bar{x}$) where,

$$\bar{x} = \frac{\sum x}{n}$$

(b) The point estimate of the population standard deviation (σ) is the sample standard deviation (s)

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} \qquad \text{or,} \qquad s = \sqrt{\frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}}$$

(c) The point estimate of a population proportion (p) is the sample proportion ($\bar{p}$)

where, $\bar{p} = \frac{x}{n}$ x = no. of successes
n = sample size

Example 9

Suppose that the mean of a population is 200 with a standard deviation of 50. A simple random sample of size 100 (n=100) is selected from this population. The sample mean $\bar{x}$ will be used to estimate the population mean μ.

(a) Find the expected value of $\bar{x}$; that is, E ($\bar{x}$).

(b) Find the standard deviation of $\bar{x}$ or the standard error of the mean.

Example 9...cont.

(c) What distribution the sample mean would follow?
(d) What does the sampling distribution of $\bar{x}$ show?

Solution:

Given μ = 200 σ= 50 n = 100

(a) E ($\bar{x}$) = μ or, E($\bar{x}$) = μ = 200

(b) $$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{50}{\sqrt{100}} = 5$$

(c) The sampling distribution of $\bar{x}$ would follow a normal distribution (because n >30) with mean E ($\bar{x}$) = 200 and $\sigma_{\bar{x}} = 5$

(d) The sampling distribution of the sample mean, $\bar{x}$ shows the probability distribution of all possible sample means that can be observed with random samples of size 100.

This distribution can be used to compute the probability that $\bar{x}$ is within a specified value from μ We can use the normal distribution to calculate the probabilities because the sample size is large (n is >30)

Example 10

The mean price of a particular brand of a digital camera is $200 with a standard deviation of 50. Suppose these are the population mean and standard deviation for this brand of camera; that is, μ = 200 and σ = 50. From this population, a random sample of size 100 (n=100) is selected to estimate the population mean price, μ.

(a) Find the probability that the sample mean will be within ± 5 of the population mean.

Solution: μ = 200 σ = 50 n = 100

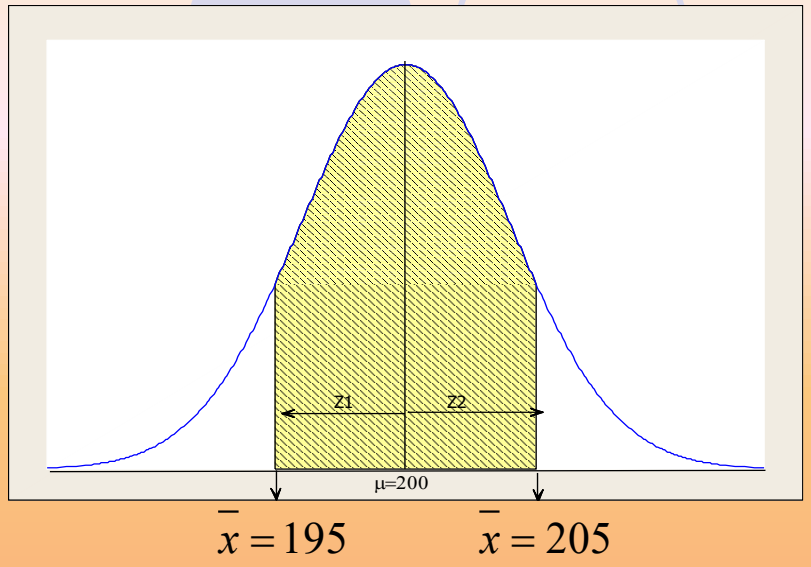
Since the sample size is greater than 30, we can use the normal distribution to find the probabilities (according to the central limit theorem).

The probability that the sample mean will be within± 5 of the population mean

$\bar{x}$ - μ = 195-200= -5 and 200-195 = +5 (see the slide on the next page)

The required probability is the shaded area shown in the next slide.

Example 10...cont.



$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{5}{\frac{50}{\sqrt{100}}} = 1.0$$

z = 1.0 is equivalent to an area of 0.3413 (from the z-table); the required probability = 2 X 0.3413 = 0.6826 = 68.26%

Note: ± 5 of the population mean can be seen as the area between 195 and 205.

Alternatively, (see figure above)

$$z_1 = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{195 - 200}{50/\sqrt{100}} = -1.0 \Rightarrow 0.3413$$

(z= -1.0 corresponds to an area of 0.3413 from the z-table)

$$z_2 = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{205 - 200}{50/100} = 1.0 \Rightarrow 0.3413$$

(z= 1.0 corresponds to an area of 0.3413 from the z-table)

Required probability is the area of the shaded region which is 0.3413 + 0.3413 = 0.6826 or, $P(195 \le \bar{x} \le 205) = 0.6826$

Sampling Distribution of Sample Proportion, $\overline{p}$

- *Use the sample proportion* $\overline{p}$ *to make statistical inferences about the population proportion (p).*
- *The sampling distribution of* $\overline{p}$ *is the probability distribution of all possible values of the sample proportion (p).*
- *The sampling distribution of* $\overline{p}$ *can be approximated by a normal distribution whenever the sample size is large.*

The basis of the sampling distribution of the proportion is the Binomial distribution. The sample size (n) can be considered large whenever:

np ≥5 (where, n is the number of trials and p is the probability of success in the Binomial distribution).

and, n(1-p) ≥5

The sample proportion is calculated as $\overline{p} = \frac{x}{n}$

where, x= no. of successes, n= sample size.

Useful results for sample proportion:

Population proportion = p Population proportion = $\bar{p}$

Expected value of $\bar{p}$ $E(\bar{p}) = p$

Standard deviation of $\bar{p}$: $\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$ for an infinite population

$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}$ for a finite population

For large sample (n ≥30); the sampling distribution of the sample proportion ($\bar{p}$) follows a normal distribution. The sampling distribution of the sample proportion is given by

$$z = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

for an infinite population

$$z = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}}$$

for a finite population

Example 11

Suppose a sample of n = 100 is taken from a population with p = 0.40.

(a) What is the expected value of $\bar{p}$?

$$E(\bar{p}) = p = 0.40$$

(b)What is the standard deviation of $\bar{p}$?

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(0.40)(0.60)}{100}} = 0.0490$$

(c) ***Show the sampling distribution of $\bar{p}$.***

The sampling distribution of $\bar{p}$ will follow a normal distribution with

$$E(\bar{p}) = p = 0.40 \quad \text{and} \quad \sigma_{\bar{p}} = 0.0490$$

(d) What does the sampling distribution of $\bar{p}$ show?

The sampling distribution of $\bar{p}$ shows the probability distribution for sample proportion, $\bar{p}$.

Example 12

The quality control department believes that 30% of the firm's defective parts are supplied from one vendor. A simple random sample of 100 parts will be used to estimate this proportion.

Note: n = 100 p = 0.30

(a) What is the sampling dist. of $\overline{p}$ for this study?

The sampling dist. of $\overline{p}$ will be normal because

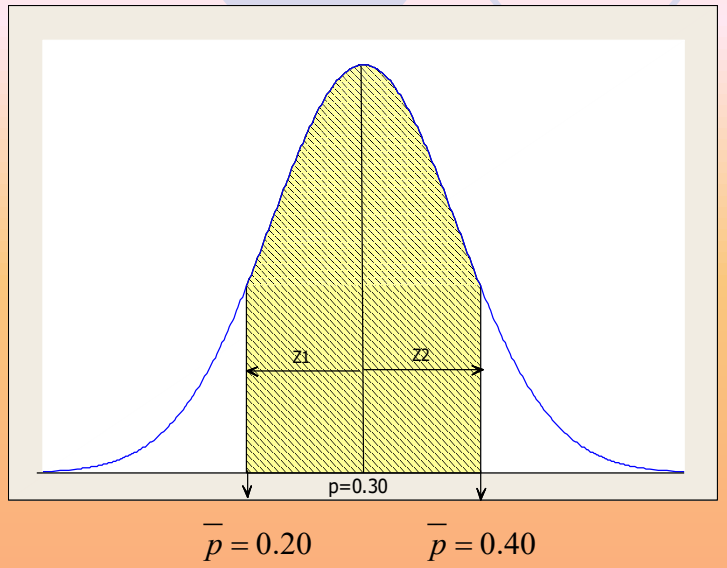
np = 100 (0.30) = 30 and n(1-p) = 100 (0.70) = 70

Note that the sampling distribution of $\overline{p}$ follows a normal distribution for a large sample if np ≥5 and, n (1-p) ≥5 then the sample size is considered to be large.

(b) What is the probability that the sample proportion $\overline{p}$ will be between 0.20 and 0.40?

Example 12…cont.

The required probability is the shaded area in the figure.



$$z_1 = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.20 - 0.30}{\sqrt{\frac{(0.30)(0.70)}{100}}} = -2.18 \Rightarrow 0.4854$$

$z_1 = -2.18$ corresponds to an area of 0.4854

$$z_2 = \frac{0.40 - 0.30}{\sqrt{\frac{(0.30)(0.70)}{100}}} = +2.18 \Rightarrow 0.4854$$

$z_1 = +2.18$ corresponds to an area of 0.4854

Therefore, the required probability; $p(0.20 \le \bar{p} \le 0.40) = 0.4854 + 0.4854 = 0.9708$

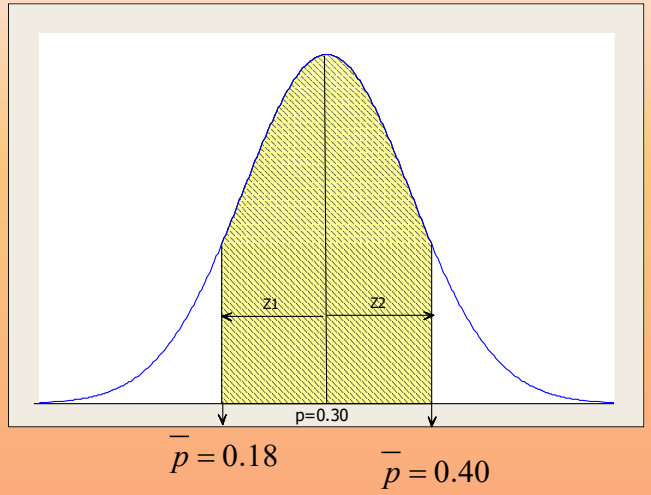
(c) What is the probability that the proportion defective is between 25% and 35%?

$$p(0.25 \le \bar{p} \le 0.35) = 0.7242$$

Example 12…cont.

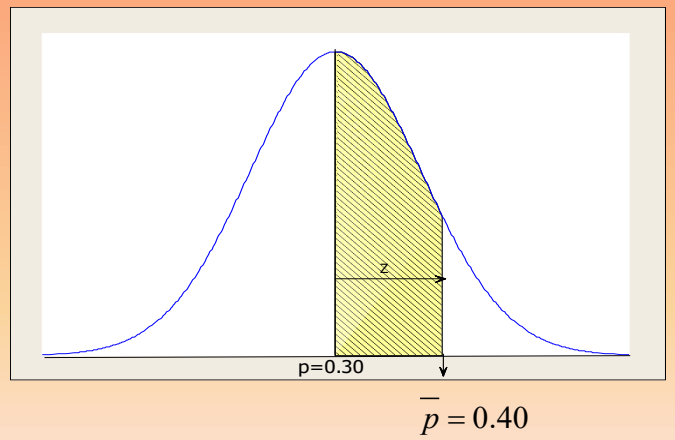
(d) Find the probability that will be between

(i) 0.18 and 0.40 or, $p(0.18 \le \overline{p} \le 0.40) = ?$



Find z1 and z2 using the formula in part (b) and find their values from the normal table. Add the results to get the required probability. The required probability is given below. You should verify the result.

$$p(0.18 \le \overline{p} \le 0.40) = 0.9810$$



(ii) 0.30 and 0.40 $p(0.30 \le \overline{p} \le 0.40) = ?$

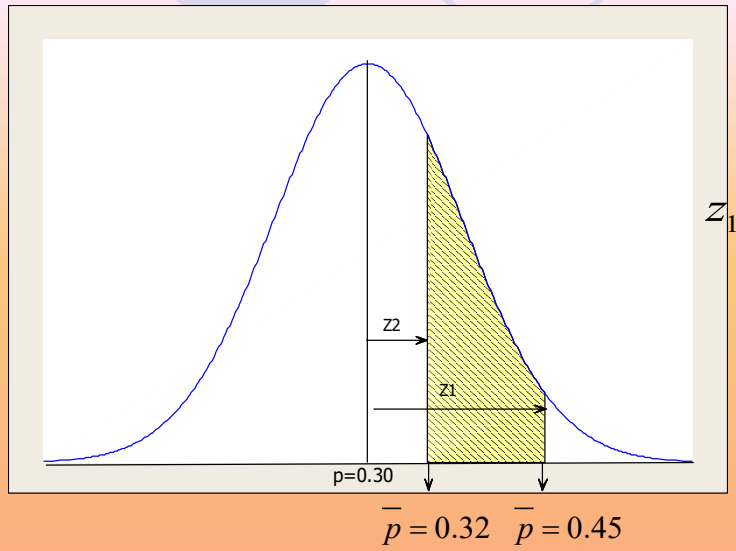
This probability is,

$$p(0.30 \le \overline{p} \le 0.40) = 0.4854$$

Example 12…cont.

(iii) 0.32 and 0.45

$$p(0.32 \le \bar{p} \le 0.45) = ?$$



$$z_1 = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.45 - 0.30}{\sqrt{\frac{(0.30)(0.70)}{100}}} = 3.27 \approx 0.50$$

$z_1 = 3.27$ corresponds to an area of approximately 0.5

$$z_2 = \frac{0.32 - 0.30}{\sqrt{\frac{(0.30)(0.70)}{100}}} = +0.44 \Rightarrow 0.1700$$

$z_2 = +0.44$ corresponds to an area of 0.1700

The required probability;

$$p(0.32 \le \bar{p} \le 0.45) = 0.5 - 0.1700 = 0.33$$

Standard Normal Distribution Table

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.00	0.004	0.008	0.012	0.016	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.091	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.148	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.17	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.195	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.219	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.258	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.291	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.334	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.377	0.379	0.381	0.383
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.398	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.437	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.475	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.483	0.4834	0.4838	0.4842	0.4846	0.485	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.489
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.492	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.494	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.496	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.497	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.498	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.499	0.499

Sampling and Sampling Distribution

Standard Error &Sampling Distribution of the Sample Mean

Expected Value of $\overline{x}$

$$E(\overline{x}) = \mu$$

Standard deviation of $\overline{x}$, $\sigma_{\overline{x}}$ or the Standard error

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$ for an infinite population

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$ for a finite population

---

$\sigma_{\overline{x}} =$ where, n= sample size, N= population size, σ is the population standard deviation

The factor,

$$\sqrt{\frac{N-n}{N-1}}$$

is known as the finite population correction factor

if $\frac{n}{N} < 0.05$ (the ratio of the sample to population size is <0.05), do not use the finite population correction factor as it does not help reduce the standard error)

Sampling Distribution of $\overline{x}$ follows a normal distribution for $n \geq 30$

$$z = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$ For an infinite population

---

$$z = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}$$ For a finite population

Sampling Distribution of the Sample Proportion

Expected value of $\overline{p}$: $E(\overline{p}) = p$

$\overline{p}$ = sample proportion, p = population proportion

Standard deviation of $\overline{p}$:

$$\sigma_{\overline{p}} = \sqrt{\frac{p(1-p)}{n}}$$ for an infinite population

---

$$\sigma_{\overline{p}} = \sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}$$ for a finite population

Sampling Distribution of a proportion ($\overline{p}$)

Sampling distribution of sample proportion follows a normal distribution for $n \geq 30$

$$z = \frac{\overline{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

$$z = \frac{\overline{p} - p}{\sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}}$$

The first formula is for infinite population; the second is for a finite population.