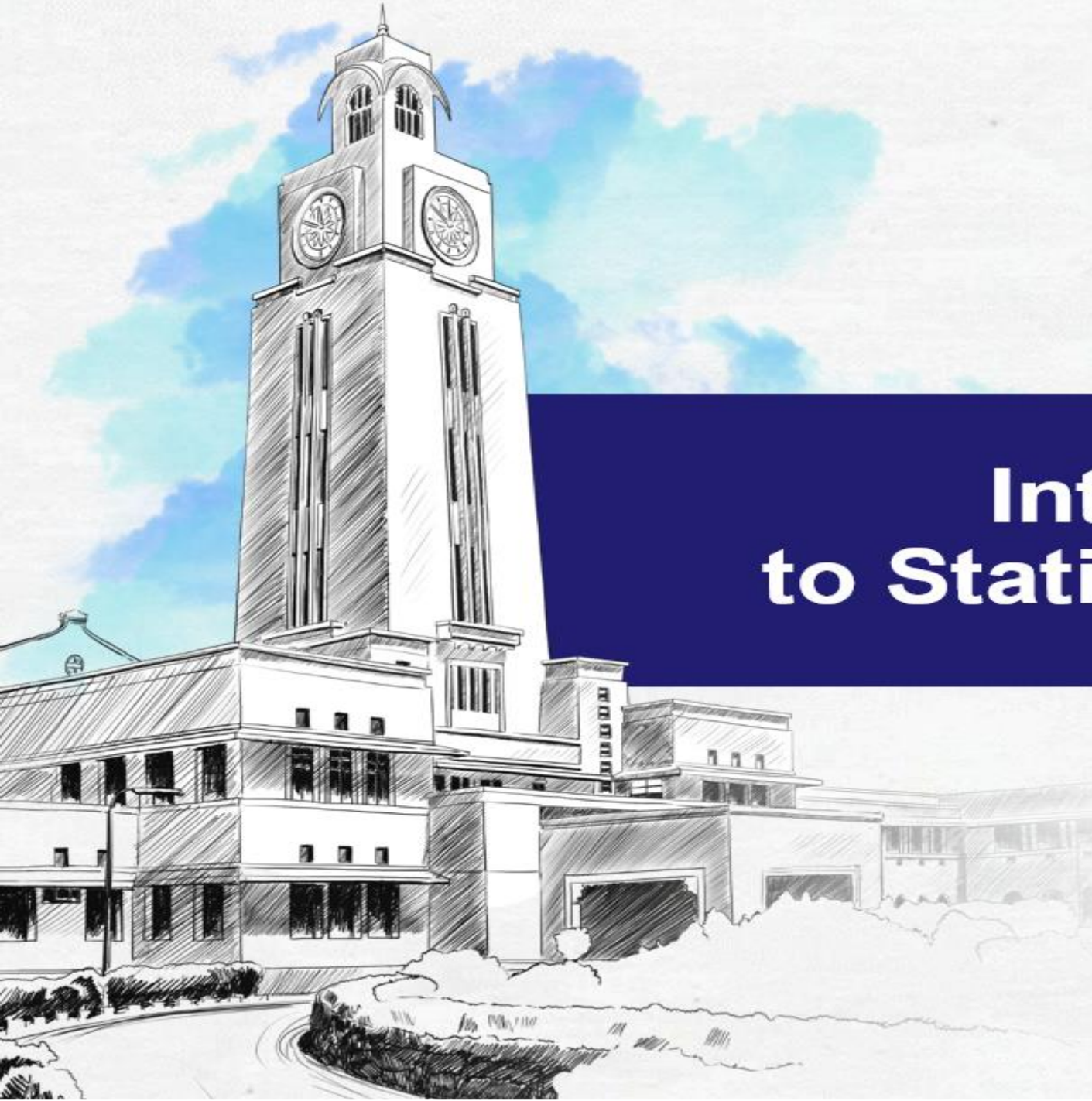


BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Statistical Methods

ISM Team

innovate achieve lead

BITS Pilani

Pilani Campus

Session 7

Testing of Hypothesis (Sampling and distribution)

(Session 7: 4th / 5th Jan 2025)

IMP Note to Self

Sampling and Distribution

Contact Session 7: Module 4: Hypothesis Testing

Contact Session	List of Topic Title	Reference
CS - 7	Sampling – random sampling and Stratified sampling Sampling distribution – Central Limit theorem, Estimation – Interval Estimation, Confidence level	T1 & T2

Definitions of some terminologies in Sampling

POPULATION (Big data or Huge or Massive data)

A population is the aggregate of facts (Data). It is the set of all elements of interest for a specific Aim and Objectives. The process of conducting a survey to collect data for the entire population is called a census.

SAMPLE (Small data)

A sample is a subset of the population. The process of conducting a survey to collect data for a sample is called a sample survey.

Definitions of some terminologies in Sampling

As one of its major contributions, Statistics uses data from a **sample** to make **estimates** and test hypotheses about the characteristics of a population in the point of view of minimising the resources (Money, Manpower, Material and Time – **3MT**). This process of Statistics is called **Inferential statistics**.

Definitions of some terminologies in Sampling

- **Parameter:**

  It is population constants, whose values generally unknown

- The various statistical characteristics or constants of a population are:
  - Population Mean (μ), Population Variance (σ^2), Population Proportion (P), Population Correlation Coefficient (ρ) etc.

- Example: Binomial distribution (n, p)

  Poisson distribution (λ)

  Normal distribution (μ, σ^2)

Definitions of some terminologies in Sampling

- **Statistic**:

  A function of sample observations

- Example
  - Sample mean ($\bar{x}$) , Sample variance (S^2), Sample proportion (P) etc.
- A **Statistic** also called an **Estimator** is used to estimate its corresponding **Parameter**. The numerical value is called estimate
- A statistic computed for different random samples forms a random variable and its distribution is called **Sampling distribution**.

Sampling

- Sampling is a process of selecting samples from a well defined **target** populations. A target population is a population characteristic in which the Analyst is interested. For instance, the coping problem of all adolescents with elders

- If the representative samples are selected from the well defined target population which are also consistent and accurate, then the estimates obtained from such samples helps in better generalization of the unknown parameters. This process is known as Inferential Statistics which forms a good scientific decision-making tool

Reasons for Sampling

Most often the population under study may be either unknown or infinite. Collecting data from such population is not feasible. Further, even if the population is countably finite, the resources required to collect data may be huge. Hence, by select a representative samples from the target population, not only the resources can be minimised but also a close estimation of Parameters can be made.

Reasons for Sampling

Collecting data from samples instead of population offers several Advantages:

1. The sample can save money.

2. The sample can save time.

3. For given resources, the sample can broaden the scope of the study.

Methods of selection of Sample

The data obtained from the samples are the observable data

Observed data = Truth + Bias + Random error

Sample design

Probability sampling

Procedure that assures all the units in the population have some **probabilities known in advance** of being chosen in a sample

Non-probability sampling

Procedures in which units in the sample are collected with no specific probability structure

Methods of selection of Sample

- Probability sampling
  - Simple Random Sampling
    - Systematic Random Sampling
    - Stratified Random Sampling
    - Multistage Random Sampling
- Non – Probability sampling
  - Purposive / Judgement Sampling
  - Convenience Sampling
  - Quota Sampling
  - Snowball Sampling

Methods of selection of Sample

- **The requirement for simple random sampling is**
  - Homogeneous and finite target population
  - Samples are selected unit by unit with equal probability
- **The requirement for systematic random sampling is**
  - Sampling frame (Listing of sampling units from finite population)
  - All units must be arranged in a systematic order (The order may ne w.r.t alphabets, date of births, Adhaar numbers, telephone numbers.. etc)
  - Only the first unit is selected at random for the sample
  - Remaining units are selected at a regular interval of k (k = N/n)

Methods of selection of Sample

- **The requirement for Stratified random sampling is**
  - Heterogeneous target population
  - Divide the heterogeneous target population into different stratum by ensuring homogeneity within each stratum
  - Use simple random sampling within each stratum to select samples proportional to population size
  - For example if there are three stratum with respective population sizes 2000, 6000, and 1000. The sample size is 1500, then with probability proportional to population size (PPPS) the respective samples are 333, 1000 and 167. (eg., 2000x1500/9000=333)
  - Obtain the estimates from each stratum and compute an estimate by weighted average.

Methods of selection of Sample

- **The requirement for Multistage random sampling is**
  - Heterogeneous target population spread across wide geographical area in the form of clusters(eg. Country, State, District etc.)
  - Select samples (clusters) at different stages
- **The requirement for Purposive (judgment) sampling is**
  - Sample selection is based on some purpose
  - The units which are judged as able to serve a specific purpose are only taken for the sample
- **The requirement for Convenience sampling is**
  - Sample selection based on some defined criteria which is convenient to the Investigator who collects data

Methods of selection of Sample

- **The requirement for Quota sampling is**
  - Heterogeneous target population
  - Divide the heterogeneous target population into different stratum by ensuring homogeneity within each stratum
  - Select samples by fixing the quota based on the convenience of investigator or researcher
- **The requirement for Snowball (Chain link) sampling is**
  - Heterogeneous target population
  - Sample selection is based on the recommendation of already selected units for the sample.

Sampling Error

- which occurs when the sample is not a representative of the population (not selected using probability basis).
- Improper sampling techniques used for selecting the samples
- Sampling error is proportionate to sampling variation

Sampling Variation

Population of Wages of employees of an organization

1861	2495	1000	2497	1865	791	2090	2637	1327	1678
1680	2858	795	2495	2496	2501	1160	1480	1860	2490
2090	2840	2490	2640	659	827	2646	2638	2643	868
1327	1866	1861	2486	2865	3011	2494	1489	1865	2855
2840	2499	2093	2660	1165	2600	2085	2640	2998	1861
2956	2495	2865	1865	3000	3019	1670	2858	2642	1680
3038	3000	1313	596	656	3240	590	2501	2485	3015
2092	1679	3024	2497	2825	2630	2070	2900	1861	2636
2495	2637	2497	1159	2640	3050	870	2896	2500	2638
926	2860	1481	875	2482	1860	2086	934	3200	2490

Sampling Variation

Selection of different samples of varied sizes

Sample 1

3000 2486 820 1678 2070 2638 2490 1865 1000 2090 596 3200

Sample 2

2840 2858 3000 2490 2998 3050 2070 2896 3200 2490 3280

Sample 3

2858 3240 2497 2865 656 2093 934 1861 868 795

Sample 4

2086 1000 2497 596 656 875 2085 934 1313

Sample 5

820 1313 3000 2640 596 2640 2600 2495 934 2500

Sampling Variation

Selection of different samples of varied sizes

Sample 6

2840 2499 1327 1861 2495 3024 3038 2497

Sample 7

2858 2490 868 1670 1480 2643 1480 1680 2085 2490

Sample 8

2495 2858 1861 2092 2499 3000 2660 1000 1679 926 2660

Sample 9

795 791 3200 2085 2638 2497 2486 1159 2640

Sample 10

3019 3240 3200 3050 3000 3015 2900 2896 2998

Sampling Variation

Compute Sample Mean, SD and Variance of these samples

Sample No.	Sample size	Mean	SD	Variance
1	12	1994.42	843.23	711036.83
2	11	2830.18	349.94	122458.00
3	10	1866.70	988.57	977270.64
4	9	1338.00	704.36	496123.01
5	10	1953.80	920.44	847209.79
6	8	2447.63	590.64	348855.61
7	10	1974.40	638.05	407107.80
8	11	2157.27	715.1	511368.01
9	9	2032.33	891.53	794825.74
10	9	3035.33	117.4	13782.76
Overall	100	**2162.24**	**732.26**	**536204.71**

Sampling Variation

- The term **"sampling variation"** refers to the fact that the sampling estimates vary from one sample to another forming a random variable.
- **Sampling variability** will **decrease** as the **sample size increases.**
- The samples must be randomly chosen. The size of the samples should be optimum (neither too small nor too large). The required optimum sample size are comupted based on some statistical procedures like proportion, SD etc.

Sampling Variation

Do you consider these sample means, sample SDs and smple variance as variable?

If yes, should we not describe the distribution of these variables?

The distribution of the sample estimates is called sampling distribution

The distribution of sample means is called Sampling distribution of mean

The distribution of sample variance is called Sampling distribution of Variance

Definition

- The probability distribution of a statistic (sample estimate) is called **sampling distribution**.

- The sampling distribution of a statistic depends on the distribution of the population, the size of the sample, and the method of sample selection

Sampling Distribution of $\bar{x}$

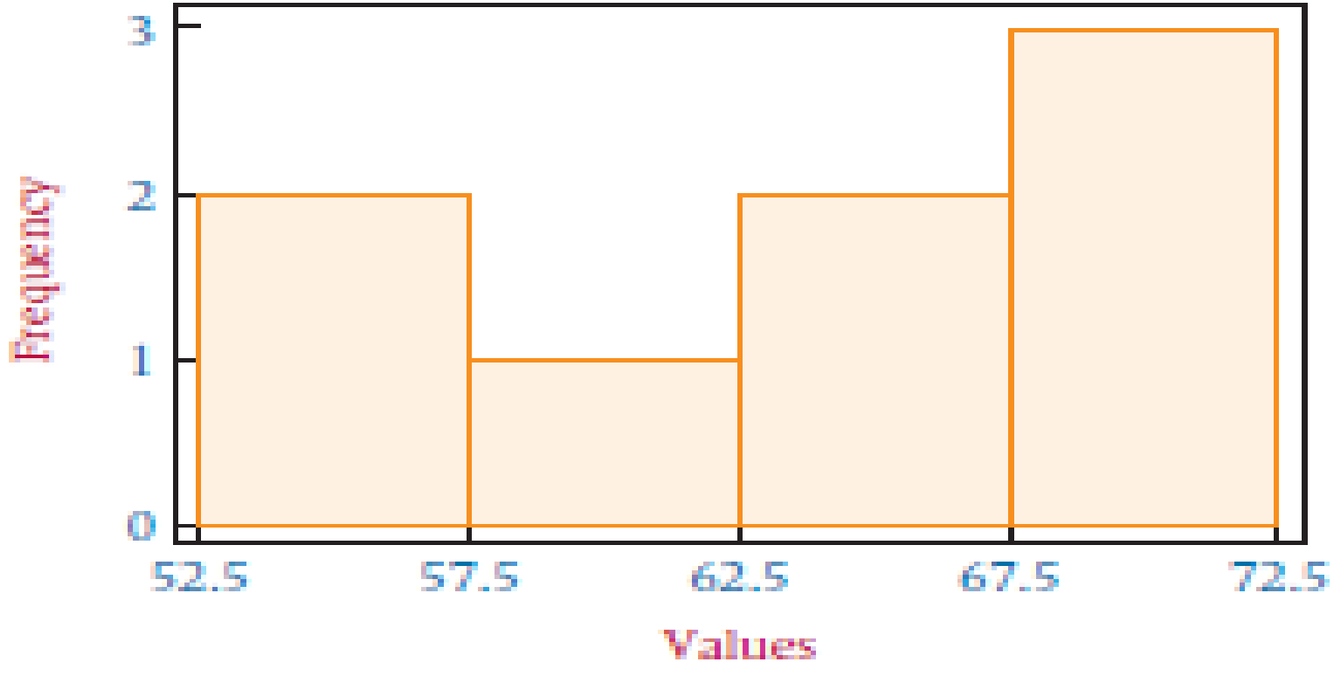
- The sample mean is one of the more common statistic used in the inferential statistics.
- The **distribution** of the values of the sample mean ($\bar{x}$) in repeated **samples** is called the **sampling distribution of $\bar{x}$**
- One way to examine the distribution possibilities is to take a population with a particular distribution, randomly select samples of a given size, compute the sample means, and attempt to determine how the means are distributed.

Example

- Suppose a small finite population consists of only $N = 8$ numbers:

  54, 55, 59, 63, 64, 68, 69, 70

- Using an Excel-produced histogram, we can see the shape of the distribution of this population of data.



- Suppose we take all possible samples of size $n = 2$ from this population with replacement.

Example

The result is the following pairs of data

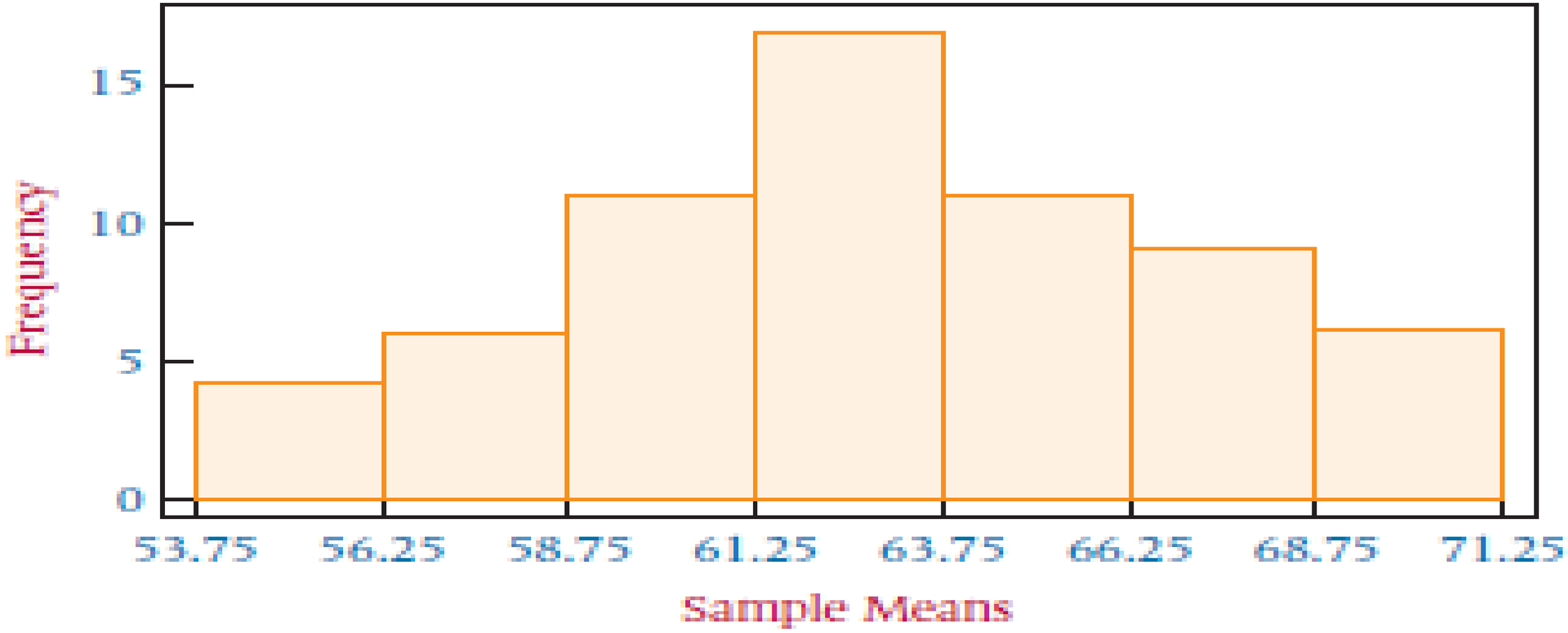
(54,54)	(55,54)	(59,54)	(63,54)
(54,55)	(55,55)	(59,55)	(63,55)
(54,59)	(55,59)	(59,59)	(63,59)
(54,63)	(55,63)	(59,63)	(63,63)
(54,64)	(55,64)	(59,64)	(63,64)
(54,68)	(55,68)	(59,68)	(63,68)
(54,69)	(55,69)	(59,69)	(63,69)
(54,70)	(55,70)	(59,70)	(63,70)
(64,54)	(68,54)	(59,54)	(70,54)
(64,55)	(68,55)	(69,55)	(70,55)
(64,59)	(68,59)	(69,59)	(70,59)
(64,63)	(68,63)	(69,63)	(70,63)
(64,64)	(68,64)	(69,64)	(70,64)
(64,68)	(68,68)	(69,68)	(70,68)
(64,69)	(68,69)	(69,69)	(70,69)
(64,70)	(68,70)	(69,70)	(70,70)

The means of each of these samples follow

54	54.5	56.5	58.5	59	61	61.5	62
54.5	55	57	59	59.5	61.5	62	62.5
56.5	57	59	61	61.5	63.5	64	64.5
58.5	59	61	63	63.5	65.5	66	66.5
59	59.5	61.5	63.5	64	66	66.5	67
60	61.5	63.5	65.5	66	68	68.5	69
61.5	62	64	66	66.5	68.5	69	69.5
62	62.5	64.5	66.5	67	69	69.5	70

Example

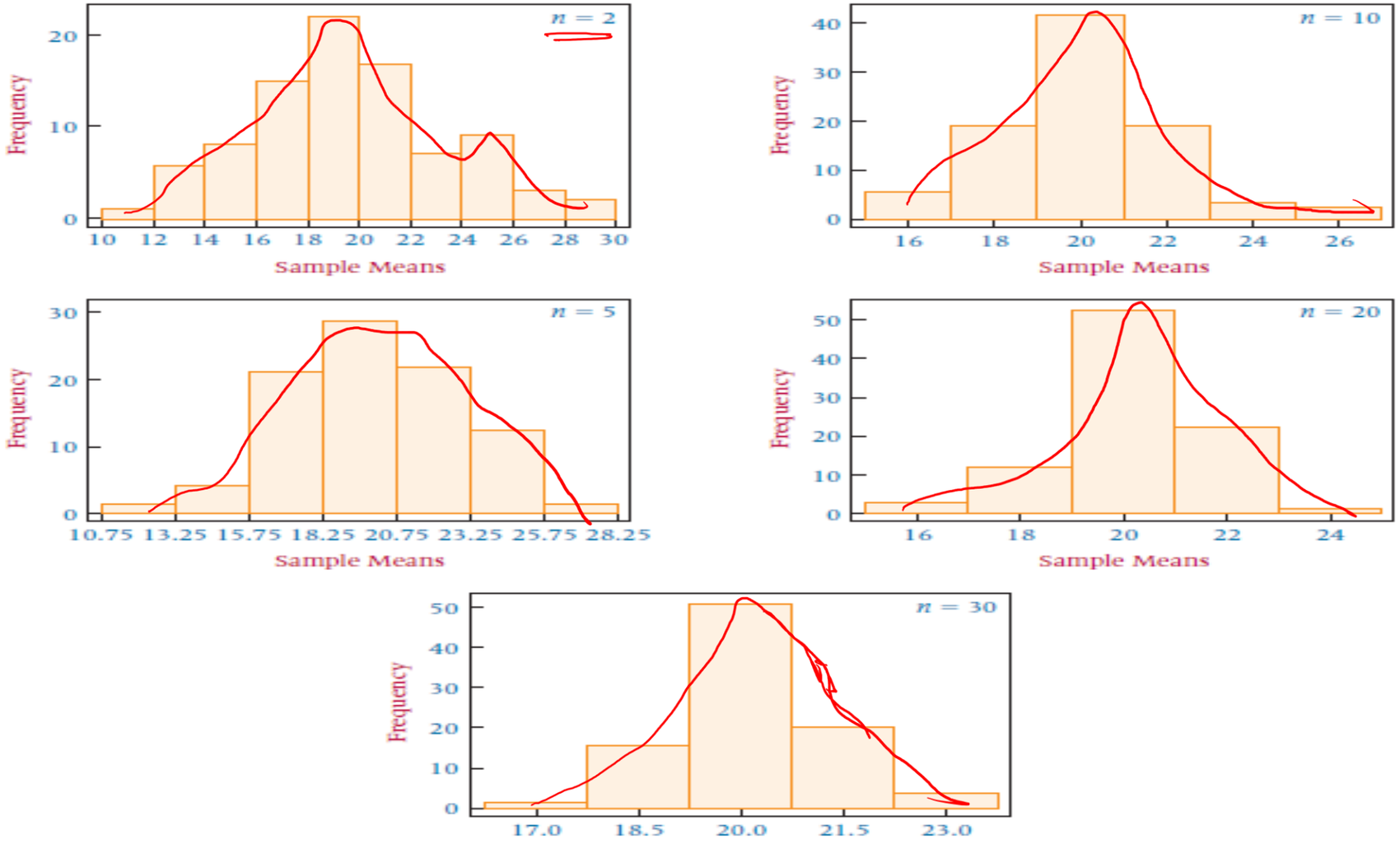
- Again using an Excel-produced histogram, we can see the shape of the distribution of these sample means.



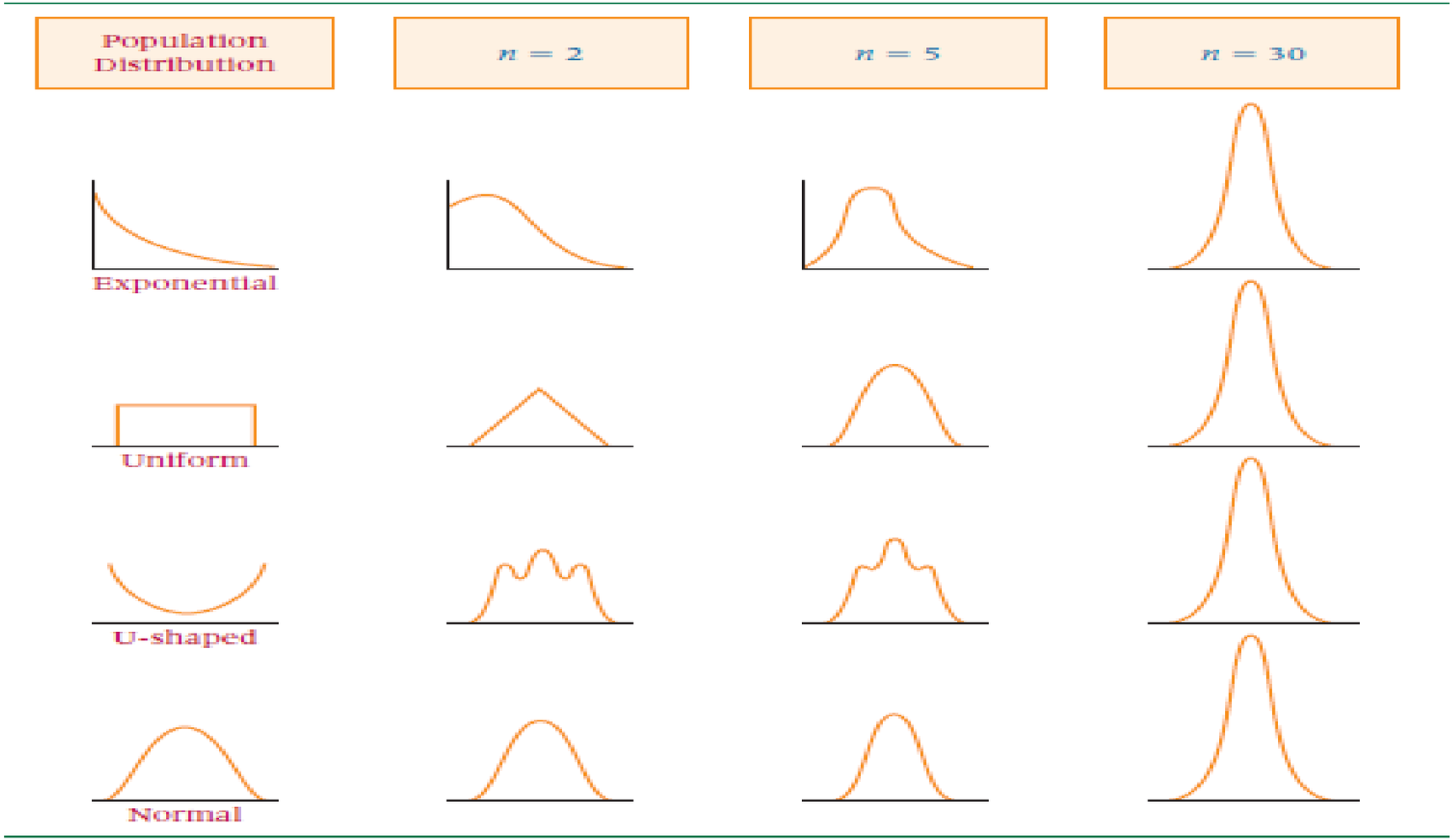
Conclusions

- Observe that the shape of the histogram for sample means is quite unlike the shape of the histogram for the population.
- The sample means appear to “pile up” toward the middle of the distribution and “tail off” toward the extremes.
- As sample sizes become much larger, the sample mean distributions begin to approach a **Normal distribution** and the variation among the means decreases.

Sample Means from 90 Samples Ranging in Size from n = 2 to n = 30 from a Uniformly Distributed Population with a = 10 and b = 30



Shapes of the Distributions of Sample Means



Central Limit Theorem

- If samples of size *n* are drawn randomly from a population that has a mean of μ and a standard deviation of σ, the sample means, $\bar{x}$, are approximately normally distributed for sufficiently large sample sizes ($n \geq 30$) regardless of the shape of the population distribution.

- If the population is normally distributed, the sample means are normally distributed for any size sample.

- From mathematical expectation: $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$

Z score for sample means

- The central limit theorem states that sample means are normally distributed regardless of the shape of the population for large samples and for any sample size with normally distributed populations.
- Thus, **sample means** can be **analyzed** by using ***z* scores**
- The formula to determine z scores for individual values from a normal distribution:

$$Z = \frac{X - \mu}{\sigma}$$

- If sample means are normally distributed, the z score formula applied to sample means would be

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}}$$

- The standard deviation of the statistic of interest is $\sigma_{\bar{x}}$, sometimes referred to as the **standard error of the mean.**

Example

The average age of a vehicle registered in the United States is 8 years, or 96 months. Assume the standard deviation is 16 months. If a random sample of 36 vehicles is selected, find the probability that the **mean** of their age is between 90 and 100 months.

Solution

Given: $\mu = 96$ months, & $\sigma = 16$, $n = 36 > 30 \rightarrow CLT$

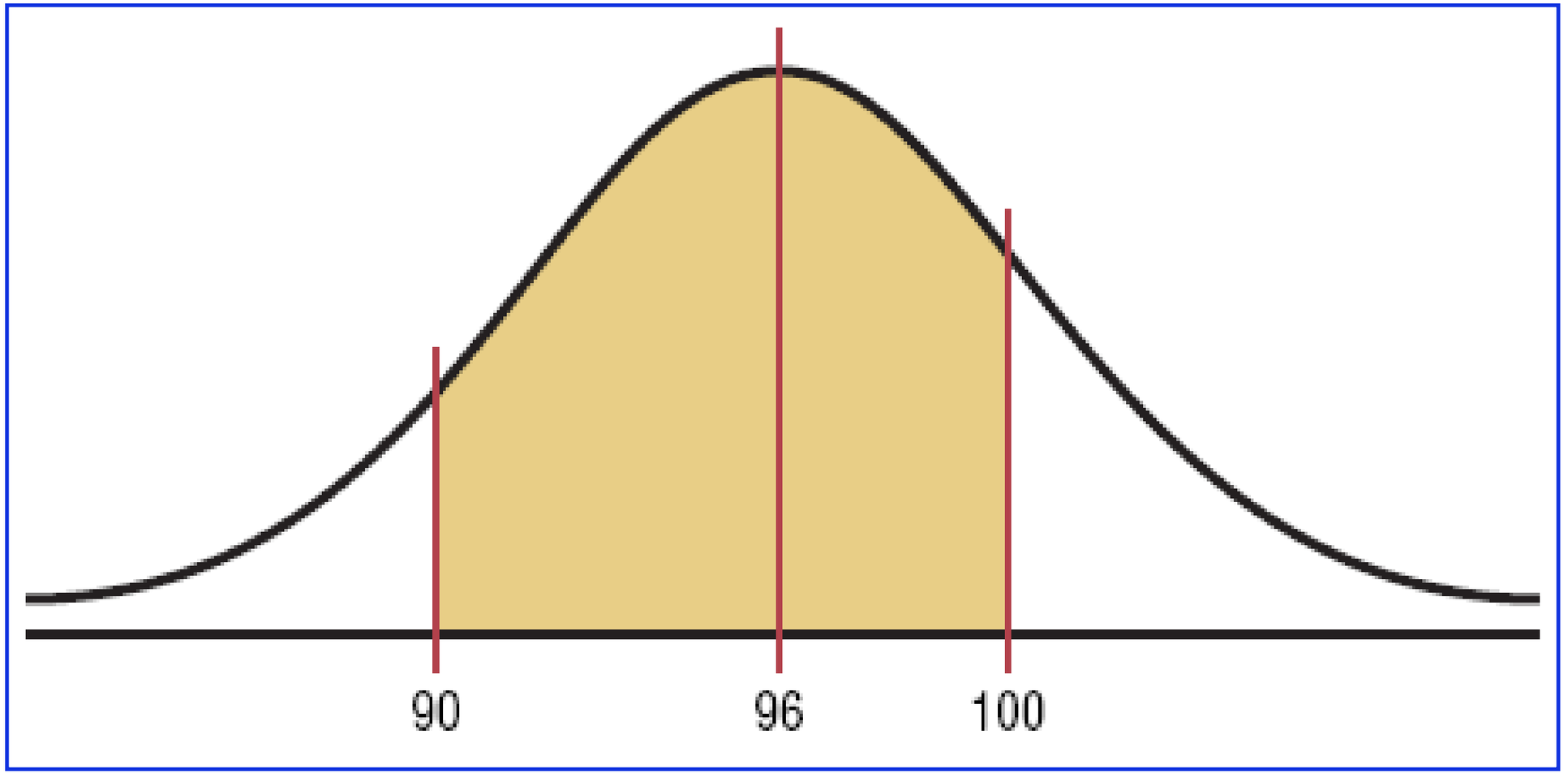
$$z = \frac{x - \mu}{\sigma}, z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$$z = \frac{90 - 96}{\frac{16}{\sqrt{36}}} = -2.25 \qquad z = \frac{100 - 96}{\frac{16}{\sqrt{36}}} = 1.50$$

$$P(90 < \bar{x} < 100) = \quad P(-2.25 < z < 1.5)$$

$$= 0.9332 - 0.0122 = 0.9210$$

Example



Example

Suppose the mean expenditure per customer at a tire store is $85.00, with a standard deviation of $9.00.

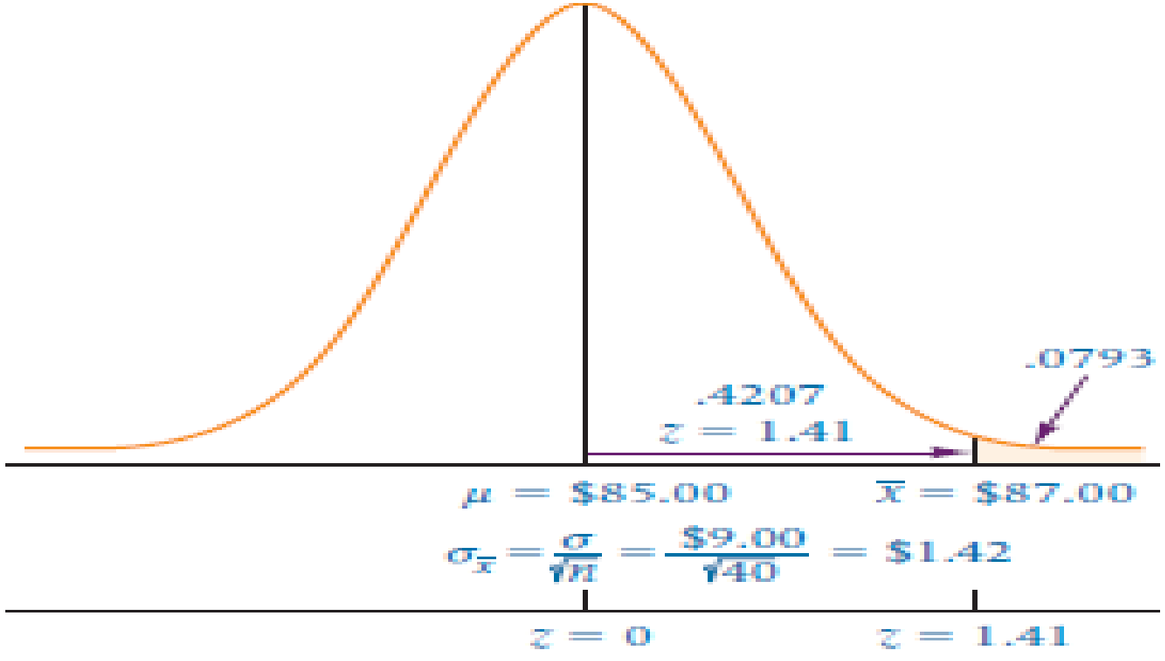
If a random sample of 40 customers is taken, what is the probability that the sample average expenditure per customer for this sample will be $87.00 or more?

Solution

Because the sample size is greater than 30, the central limit theorem Can be used, and the sample means are normally distributed.

μ= $85 σ= $9

$$z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\$87.00-\$85.00}{\frac{\$9.00}{\sqrt{40}}} = \frac{\$2.00}{\$1.42} = 1.41$$



STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361

Example

Suppose that during any hour in a large department store, the average number of shoppers is 448, with a standard deviation of 21 shoppers.

What is the probability that a random sample of 49 different shopping hours will yield a sample mean between 441 and 446 shoppers?

Solution

For this problem, μ= 448, σ= 21, and n = 49. The problem is to determine P(441<= $\bar{x}$ <=446).

The following

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

$$z_1 = \frac{441-448}{\frac{21}{\sqrt{49}}} = \frac{-7}{3} = -2.33\ , \quad z_2 = \frac{446-448}{\frac{21}{\sqrt{49}}} = \frac{-2}{3} = -0.67$$

Now, P(-2.33<z<-0.67)

=P(0.67<z<2.33)

=F(2.33)-F(0.67)

=0.9901-0.74851

=0.24159

Sampling from a Finite Population

- The earlier example was based on the assumption that the population was infinitely or extremely large.
- In cases of a finite population, *a statistical adjustment can be made to the z formula for sample means*. The adjustment is called the **finite correction factor** $\sqrt{\frac{N-n}{N-1}}$
- Following is the z formula for sample means when samples are drawn from finite populations. $z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}$

Rules for finite population

- As the size of the finite population becomes larger in relation to sample size, the finite correction factor approaches 1.

- In theory, whenever researchers are working with a finite population, they can use the finite correction factor.

- A rough rule of thumb for many researchers is that, if the sample size is **less** than **5%** of the finite population size or ***n/N* < 0.05**, the finite correction factor does **not** significantly modify the solution.

Example

A production company's 350 hourly employees average 37.6 years of age, with a standard deviation of 8.3 years.

If a random sample of 45 hourly employees is taken, what is the probability that the sample will have an average age of less than 40 years?

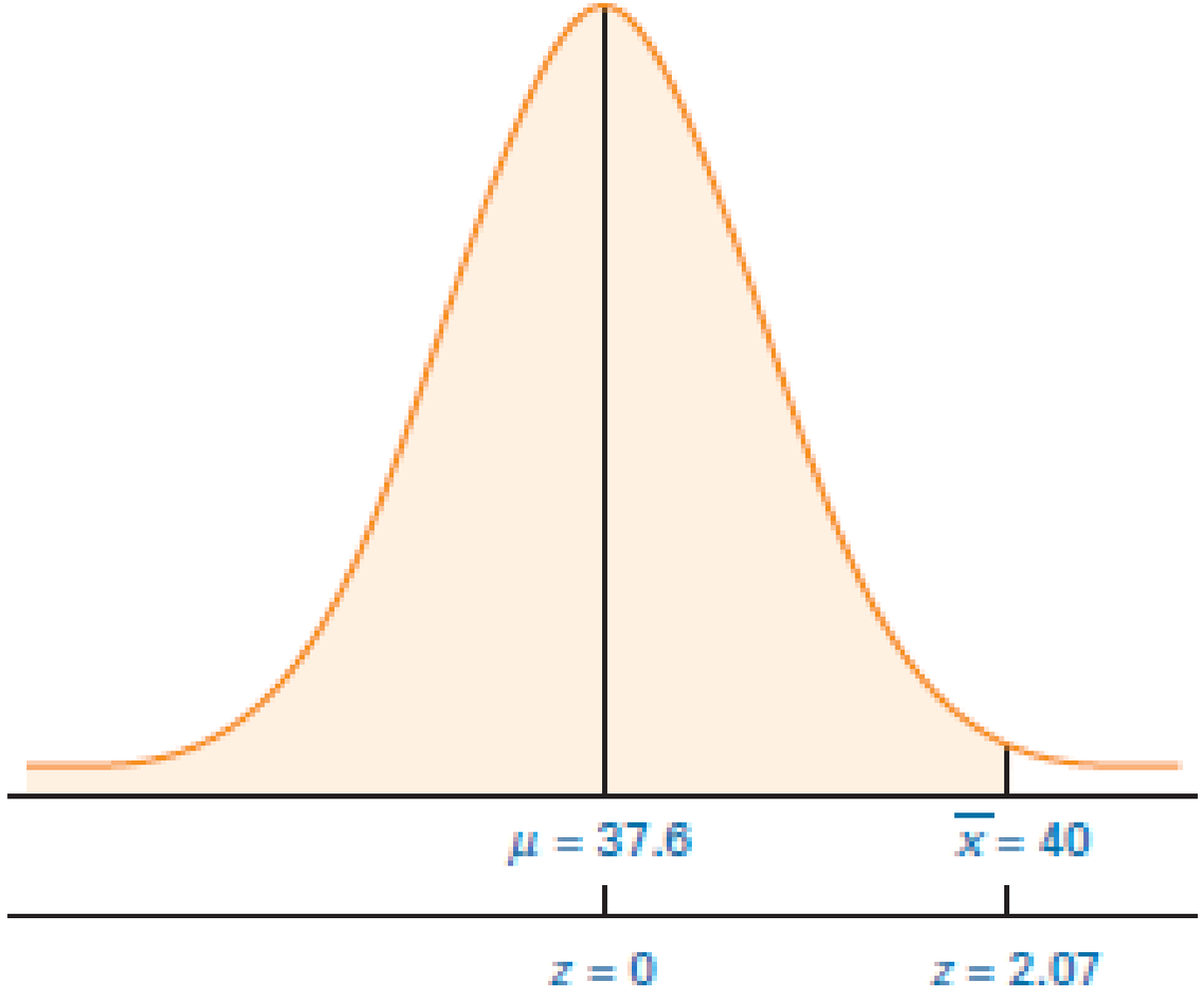
Solution

- The population mean is 37.6, with a population standard deviation of 8.3.
- The sample size is 45, but it is being drawn from a finite population of 350; that is, n = 45 and N = 350.
- The sample mean under consideration is 40
- Using the z formula with the finite correction factor gives

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}} = \frac{40 - 37.6}{\frac{8.3}{\sqrt{45}}\sqrt{\frac{350-45}{350-1}}} = \frac{2.4}{1.157} = 2.07$$

…Solution

- This z value yields a probability of .4808.
- Therefore, the probability of getting a
- sample average age of less than
- 40 years is **.4808 + .5000 = .9808.**



Sampling Distribution Of Sample Proportion

- If research results in ***countable*** items such as how many people in a sample have a flexible work schedule, the sample proportion is often the statistic of choice.

SAMPLE PROPORTION

$$\hat{p} = \frac{x}{n}$$

Where

x= number of items in a sample that have the characteristic

n= number of items in the sample

Example

- In a sample of 100 factory workers, 30 workers might belong to a union.

- The value of sample proportion for this characteristic, union membership, is

$$30/100 = 0\ .30$$

How does a researcher use the sample proportion in analysis?

- The central limit theorem applies to sample proportions in that the normal distribution approximates the shape of the distribution of sample proportions

- If $n*p > 15$ and $n*q > 15$ (p is the population proportion and $q = 1 - p$).

- The mean of sample proportions for all samples of size n randomly drawn from a population is p (the population proportion) and the standard deviation of sample proportions is $\sqrt{\frac{pq}{n}}$

- sometimes referred to as the **standard error of the proportion**

Z – Formula for Sample Proportions

For $n*p > 1$ 5 and $n*q > 15$

$$Z = \frac{\hat{p} - P}{\sqrt{\frac{pq}{n}}}$$

where

$P = Population\ proportion$

$\hat{p} = Sample\ proportion$

$q = 1\text{-}p$

n = Sample size

Example

Suppose 60% of cancer patient are cured by new drug. What is the probability that a random sample of size 120 patients 50% or more will cured by new drug?

Solution

$$Z = \frac{\hat{p} - P}{\sqrt{\frac{pq}{n}}} = \frac{0.5 - 0.6}{\sqrt{\frac{0.5 * 0.5}{120}}} = -2.24$$

The probability corresponding to *z*

= -2.24

For *z* > -2.14 (the tail of the distribution) = 1- P(z<-2.24)

= 1- 0.**0125= 0.9875**

Forms of Statistical Inference

- Three forms of statistical inference
  - Point estimation
  - Interval estimation
  - Hypothesis testing

Point Estimate

- A **point estimate** is a statistic taken from a sample that is used to estimate a population parameter.
- A point estimate is only as good as the representativeness of its sample.
- If other random samples are taken from the population, the point estimates derived from those samples are likely to vary.

Point Estimation

Statistical measure	Sample estimates		Population parameter
Mean	$\bar{x}$	→	μ
Variance	s^2	→	σ^2
Proportion	p	→	P
Correlation	r	→	ρ

Interval Estimate

- Because of variation in sample statistic, point estimate some times may not be a reliable estimate for population mean, more so when the target population from where the samples drawn is heterogeneous.
- In order to overcome this, by making use of point estimates, an interval with lower and upper bound for the parameter or difference in parameters can be computed at certain level of confidence, so that the value of the unknown parameter or difference in parameters are covered within this limit.

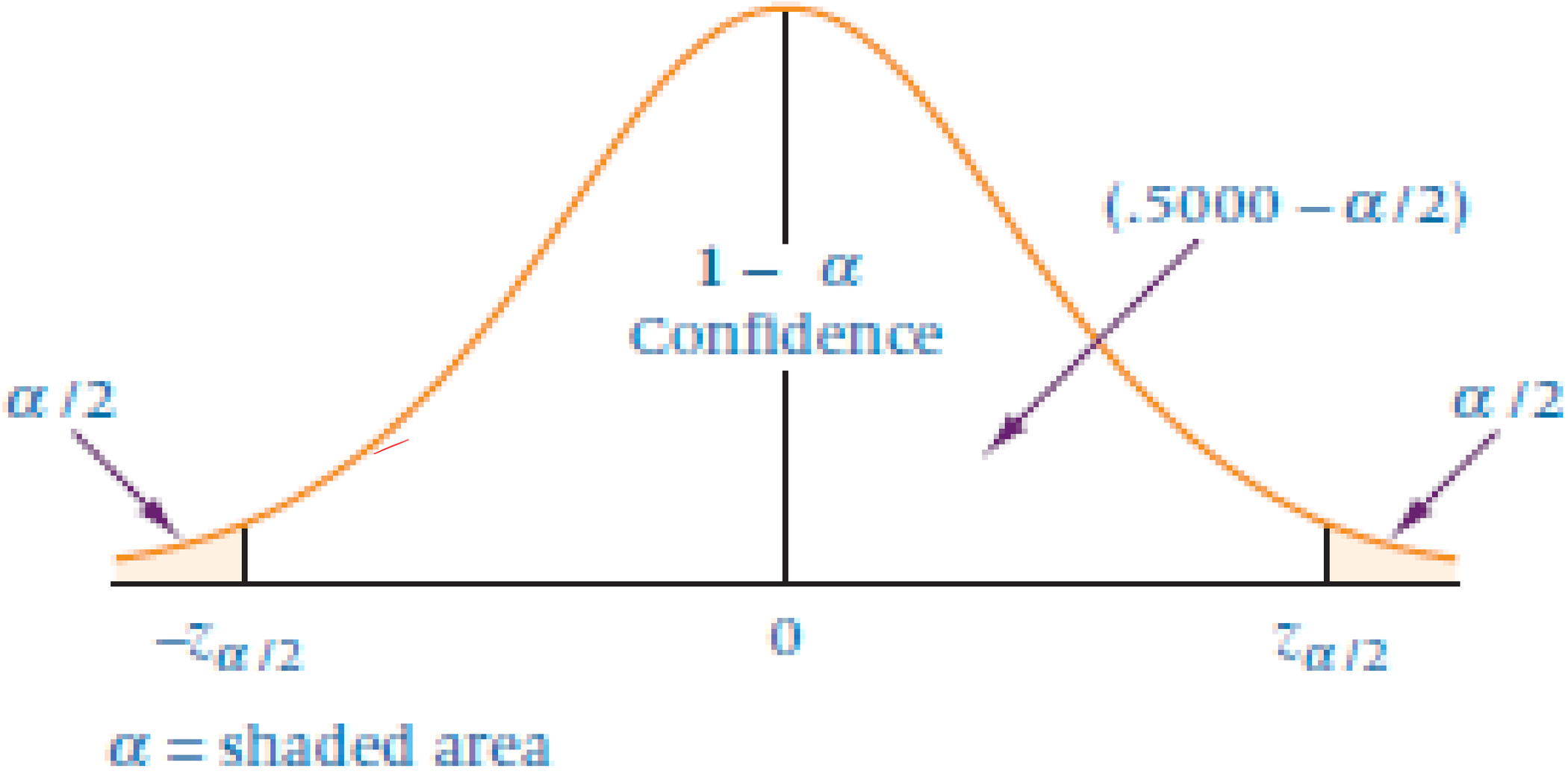
Interval Estimate

- Undoubtedly, the interval estimate with certain confidence level is the most powerful type of inference called Confidence interval.
- The confidence interval is a range of values within which the analyst can declare, with some confidence, the population parameter lies.

Confidence Interval

- Computation of 100 (1-α)% confidence interval is the most common way of finding the interval estimate, where α is the probability of type I error.
- Most commonly used confidence level close to 1, are 0.95 (95%) or 0.99 (99%).

The role of α in Confidence Intervals



The role of α in Confidence Intervals

Standard normal variable value $z_{\alpha/2}$ (Table Value)	Level of significance ∝		
	1%=0.01	**5%=0.05**	**10%=0.1**
∝/2	0.01/2 =0.005	0.05/2=0.025	0.1/2=0.05
F($z_{\alpha/2}$)=1-∝/2 P(Z<=$z_{\alpha/2}$)= 1-∝/2	P(Z≤2.58)= 0.995 then $z_{\alpha/2}$= 2.58	P(Z≤1.96)= 0.975 then $z_{\alpha/2}$= 1.96	P(Z≤1.645)= 0.95 then $z_{\alpha/2}$= 1.645

Confidence Interval

- Computation of 100 (1-α)% confidence interval is the most common way of finding the interval estimate, where α is the probability of type I error. α is also called area under curve at one of the tail-end of normal distribution. $\frac{\alpha}{2}$ is the area under normal curve at both the tail-ends.

- Most commonly used confidence level close to 1, are 0.95 (95%) or 0.99 (99%).

Confidence Interval for mean of population

The 100 (1-α)% confidence interval for mean is

$$\text{Sample estimate} \pm Z_{\alpha/2}\ \text{SE (estimate)}$$

Based on Standard normal distribution 100 (1-α)% CI for μ is

$$\bar{x} \pm Z_{\alpha/2}\ \text{SE}(\bar{x})$$

For α = 0.05, $Z_{\alpha/2} = 1.96$

For α = 0.01, $Z_{\alpha/2} = 2.58$

Confidence Interval for difference in means of two populations

The 100 (1-α)% confidence interval for difference in means of two populations is

$$\text{Difference in estimates} \pm Z_{\frac{\alpha}{2}} SE(Difference\ in\ \text{estimates})$$

Based on Standard normal distribution 100 (1-α)% CI for μ is

$$(\bar{X}_1 - \bar{X}_2) \pm Z_{\frac{\alpha}{2}} SE(\bar{X}_1 - \bar{X}_2)$$

For α = 0.05, $Z_{\alpha/2} = 1.96$

For α = 0.01, $Z_{\alpha/2} = 2.58$

Confidence Interval for difference in means of two populations

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}$$ When $n_1 = n_2$

$$SE(\bar{x}_1 - \bar{x}_2) = s_p \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$ When $n_1 \neq n_2$

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$ s_p = Pooled SD

Example

In the cellular telephone company, problem of estimating the population mean number of minutes called per residential user per month, from the sample of 85 bills it was determined that the sample mean is 510 minutes.

Suppose past history and similar studies indicate that the population standard deviation is 46 minutes.

Determine a 95% confidence interval.

Solution

The business researcher can now complete the cellular telephone problem. To determine a 95% confidence interval for $\bar{x}$=510, $\sigma = 46$, n=85 , and z=1.96, the researcher estimates the average call length by including the value of z in formula 8.1.

$$510\text{-}1.96\frac{46}{\sqrt{85}} \leq \mu \leq 510 + 1.96\frac{46}{\sqrt{85}}$$

$$510\text{-}9.78 \leq \mu \leq 510+9.78$$

$$500.22 \leq \mu \leq 519.78$$

…solution

- The confidence interval is constructed from the point estimate, which in this problem is 510 minutes, and the error of this estimate, which is 9.78 minutes.
- The resulting confidence interval is $500.22 \leq \mu \leq 519.78$.
- The cellular telephone company researcher is 95%, confident that the average length of a call for the population is between 500.22 and 519.78 minutes.

Example

A study is conducted in a company that employs 800 engineers. A random sample of 50 engineers reveals that the average sample age is 34.3 years. Historically, the population standard deviation of the age of the company's engineers is approximately 8 years.

Construct a 98% confidence interval to estimate the average age of all the engineers in this company.

Solution

- This problem has a finite population. The sample size, 50, is greater than 5% of the population, so the finite correction factor may be helpful.
- In this case N = 800, n = 50, $\bar{x}$= 34.3 and σ= 8
- The z value for a 98% confidence interval is 2.33

$$\bar{x} - z\alpha/_2 \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \leq \mu \leq \bar{x} - z\alpha/_2 \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

$$34.30\text{-}2.33\frac{8}{\sqrt{50}} \sqrt{\frac{800-50}{800-1}} \leq \mu \leq 34.30\text{+}2.33\frac{8}{\sqrt{50}} \sqrt{\frac{800-50}{800-1}}$$

$$34.30\text{-}2.33\frac{8}{\sqrt{50}} \sqrt{\frac{750}{799}} \leq \mu \leq 34.30\text{+}2.33\frac{8}{\sqrt{50}} \sqrt{\frac{750}{799}}$$

$$34.30\text{-}2.55 \leq \mu \leq 34.30\text{+}2.55$$

$$31.75 \leq \mu \leq 36.85$$

Confidence Interval for Population Proportion

The 100 (1-α)% confidence interval for mean is

$$\text{Sample estimate} \pm Z_{\alpha/2}\ \text{SE (estimate)}$$

Based on Standard normal distribution 100 (1-α)% CI for P is

$$\hat{p} \pm Z_{\frac{\alpha}{2}} SE(\hat{p})$$

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Confidence Interval for difference in means of two populations

The 100 (1-α)% confidence interval for difference in proportions of two populations is

$$\text{Difference in estimates} \pm Z_{\frac{\alpha}{2}} SE(Difference\ in\ \text{estimates})$$

Based on Standard normal distribution 100 (1-α)% CI for μ is

$$(p_1 - p_2) \pm Z_{\frac{\alpha}{2}} SE(p_1 - p_2)$$

Confidence Interval for difference in means of two populations

When $n_1 = n_2$

$$SE(p_1 - p_2) = \sqrt{\left(\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}\right)}, \quad q_1 = 1 - p_1, \quad q_2 = 1 - p_2$$

$$SE(p_1 - p_2) = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$ When $n_1 \neq n_2$

where

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}, q = 1 - p$$

Example

A study of 87 randomly selected companies with a telemarketing operation revealed that 39% of the sampled companies used telemarketing to assist them in order processing.

Using this information, how could a researcher estimate the *population* proportion of telemarketing companies that use their telemarketing operation to assist them in order processing?

Solution

The sample proportion, $\hat{p}$=0.39, is the point estimate of the population proportion, p. For n=87 and $\hat{p}$=0.39, a 95% confidence interval can be computed to determine the interval estimation of p. The z value for 95% confidence is 1.96. The value of $\hat{q}$=1- $\hat{p}$=1-0.39=0.61. The confidence interval estimate is

$$0.39 - 1.96\sqrt{\frac{(0.39)(0.61)}{87}} \leq p \leq 0.39 + 1.96\sqrt{\frac{(0.39)(0.61)}{87}}$$

$$=(0.39\text{-}0.10)\leq p\leq (0.39+0.10)$$

$$=0.29\leq p\leq 0.49$$

Example

Coopers & Lybrand surveyed 210 chief executives of fast-growing small companies. Only 51% of these executives had a management succession plan in place. A spokesperson for Cooper & Lybrand said that many companies do not worry about management succession unless it is an immediate problem. However, the unexpected exit of a corporate leader can disrupt and unfocus a company for long enough to cause it to lose its momentum.

Use the data given to compute a 92% confidence interval to estimate the proportions

Solution

The point estimate is the sample proportion given to be 0.51. It is estimated that 0.51, or 51% of all fast-growing small companies have a management succession plan. Realizing that the point estimate might change with another sample selection, we calculate a confidence interval.

The value of n is 210; $\hat{p} = 0.51$ and $\hat{q} = 1 - \hat{p} = 0.49$. Because the level of confidence is 92%, the value of $z_{0.4}$=1.75.

The confidence interval is computed as

$$0.51 - 1.75\sqrt{\frac{(0.51)(0.49)}{210}} \leq p \leq 0.51 + 1.75\sqrt{\frac{(0.51)(0.49)}{210}}$$

$$=(0.51\text{-}0.06) \leq p \leq (0.51+0.06)$$

$$=0.45 \leq p \leq 0.57$$

Home Work Problems

- **Question :**

- Car mufflers are constructed by nearly automatic machine. One manufacturer finds that, for any type of car muffler, the time for a person to set up and complete a production run has a normal distribution with mean 1.82 hours and standard deviation 1.20.

- What is the probability that the sample mean of the next 40 runs will be from 1.65 to 2.04 hours ?

- **Question :**
- Engine bearings depend on a film of oil to keep shaft and bearing surfaces separated. Insufficient lubrication causes bearings to be overloaded. The insufficient lubrication can be modeled as a random variable having a mean 0.6520 ml and standard deviation 0.0125 ml.

- The sample mean of insufficient lubrication will be obtained from a random sample of 60 bearings.

- What is the probability that sample mean $\bar{x}$ will be between 0.600 ml and 0.640 ml ?

- **Question :**

- A random sample size of n = 100 is taken from a population with $\sigma = 5.1$.
- Given that the sample mean is $\bar{x} = 2.16$,

- construct a 95% confidence interval for the population mean μ.

- Question :
- With reference to the data in section 2.1 (of R1) , we have
- n = 50 , $\bar{x}$ = 305.58 nm, and s^2 = 1366.86(hence, s=36.97 nm),

- Construct a 99% confidence interval for the population mean of all nanopillars.

366	333	296	304	276	336	289	234	253	292
367	323	309	284	310	338	297	314	305	330
368	391	315	305	290	300	292	311	272	312
369	355	346	337	303	265	278	276	373	271
308	276	364	390	298	290	308	221	274	343

Exercise

A survey was taken of U.S. companies that do business with firms in India. One of the questions on the survey was: Approximately how many years has your company been trading with firms in India?

A random sample of 44 responses to this question yielded a mean of 10.455 years. Suppose the population standard deviation for this question is 7.7 years.

Using this information, construct a 90% confidence interval for the mean number of years that a company has been trading in India for the population of U.S. companies trading with firms in India.

Solution

Here, n= 44, $\bar{x}$= 10.455 and σ= 7.7. To determine the value of $z_{\alpha/2}$, divide the 90% confidence in half, or take .5000 - α/2= .5000 - .0500= 0.45 where α= 10%.

Z table yields a z value of 1.645 for the area of .45

The confidence interval is

$$x - z\frac{\sigma}{\sqrt{n}} \leq \mu \leq x + z\frac{\sigma}{\sqrt{n}}$$

$$10.455\text{-}1.645\left(\frac{7.7}{\sqrt{44}}\right) \leq \mu \leq 10.455\text{+}1.645\left(\frac{7.7}{\sqrt{44}}\right)$$

$$10.455\text{-}1.910 \leq \mu \leq 10.455\text{+}1.910$$

$$8.545 \leq \mu \leq 12.365$$

Exercise

A clothing company produces men's jeans. The jeans are made and sold with either a regular cut or a boot cut. In an effort to estimate the proportion of their men's jeans market in Oklahoma City that prefers boot-cut jeans, the analyst takes a random sample of 212 jeans sales from the company's two Oklahoma City retail outlets. Only 34 of the sales were for boot-cut jeans.

Construct a 90% confidence interval to estimate the proportion of the population in Oklahoma City who prefer boot-cut jeans.

Solution

The sample size is 212, and the number preferring boot-cut jeans is 34.

The Sample proportion is $\hat{p} = \frac{34}{212} = 0.16$. A point estimate for boot-cut jeans in the population is 0.16, or 16%

The Z value for a 90% level of confidence is 1.646, and the value of $\hat{q} = 1 - \hat{p} = 1 - 0.16 = 0.84$

The confidence interval estimate is

$$0.16 - 1.646\sqrt{\frac{(0.16)(0.84)}{212}} \leq p \leq 0.16 + 1.646\sqrt{\frac{(0.16)(0.84)}{212}}$$

$$=(0.16\text{-}0.04) \leq p \leq (0.16+0.04)$$

$$=0.12 \leq p \leq 0.20$$

Thank You