Dr. Saeed A. Dobbah Alghamdi

Department of Statistics
Faculty of Sciences
Building 90, Office 26F41
King Abdulaziz University
http://saalghamdy.kau.edu.sa

Dr. Saeed's Website

Statistics
for
Business & Economics

David R. Anderson
Dennis J. Sweeney
Thomas A. Williams

Sampling and Sampling Probability

Chapter 7

Learning Objectives

LO1 Demonstrate the knowledge of statistical terms used in sampling and sampling probability.

LO2 Differentiate between probability and non-probability sampling.

LO3 Identify some sampling techniques used in probability and non-probability sampling.

LO4 Identify relevant point estimators for a population mean, population standard deviation and population proportion.

LO5 Explain the term sampling distribution.

LO6 Describe the form and characteristics of the sampling distribution of the sample mean and the sample proportion.

LO7 Explain the central limit theorem.

LO8 Apply the central limit theorem to find probabilities of selecting possible sample means from a specified population.

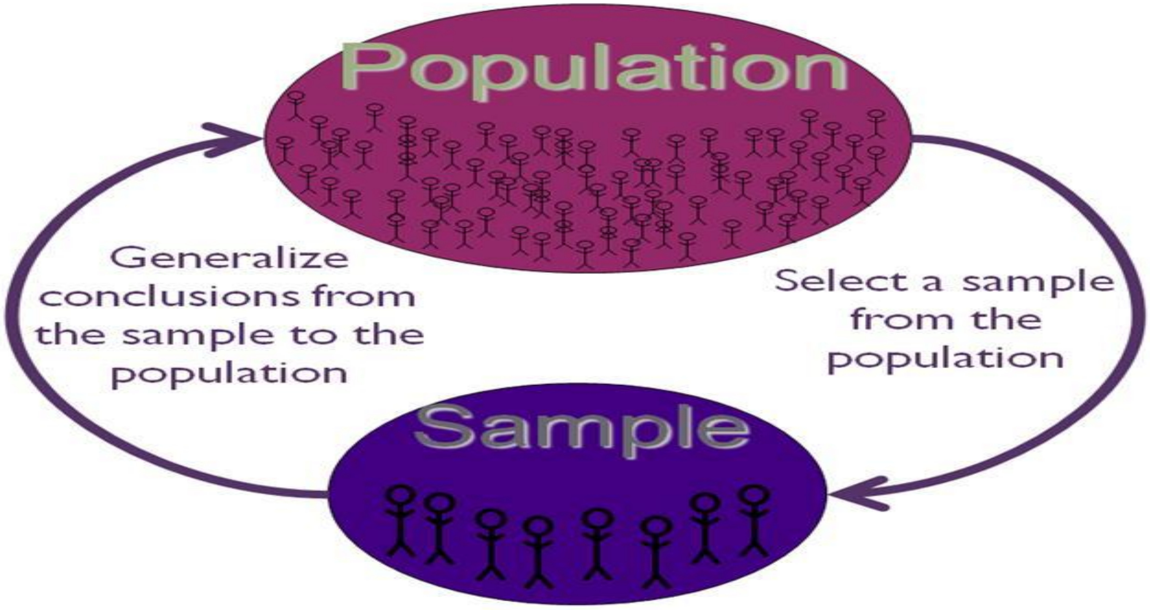
Introduction

- The reason we sample is to collect data to make an inference and answer a research question about a population.
- Numerical characteristics of a population are called parameters such as population mean, population standard deviation.
- Numerical characteristics of a sample are called sample statistics such as sample mean, sample standard deviation.
- Primary purposes of statistical inference are to make estimates and test hypotheses about population parameters using sample statistics.
- It is important to realize that sample results provide only estimates of the values of the population characteristics, because the sample contains only a portion of the population.
- A sample mean provides an estimate of a population mean, and a sample proportion provides an estimate of a population proportion.

Let us define some of the terms used in sampling.

- The sampled population is the population from which the sample is drawn.
- A sampling frame is a list of the elements from which the sample will be selected.

Population vs Sample



1. To contact the whole population would be **time consuming**.
2. The **cost** of studying all the items in a population **may be prohibitive**.
3. The **physical impossibility** of checking all items in the population.
4. The **destructive nature** of some tests.
5. The **sample results** are **adequate**.

Probability and Non-Probability Sampling

- Probability Sampling

A sample selected such that each item or person in the population being studied has a known likelihood of being included in the sample.

- Four Most Commonly Used Probability Sampling Methods
  1. Simple Random Sample
  2. Systematic Random Sampling
  3. Stratified Random Sampling
  4. Cluster Sampling
- Non-Probability sampling

Any method of sampling for which the probability of selecting a sample of any given configuration cannot be computed.

- Commonly Used Non-Probability Sampling Methods
  1. Convenience Sampling
  2. Judgmental Sampling

Probability Sampling Methods

Several methods can be used to select a sample from a population. One important method is **simple random sampling**.

The definition of a simple random sample and the process of selecting such a sample depend on whether the population is finite or infinite.

- **Sampling from a finite population**

A simple random sample of size **n** from a finite population of size **N** is a sample selected such that each possible sample of size **n** has the same probability of being selected.

1- Sampling with replacement: the sample size will be N^n

2- Sampling without replacement: the sample size will be $\frac{N!}{n!(N-n)!}$

Example: Suppose Population= 1,2,3,4 (So, N=4) and sample size n=2 is needed.

When sampling with replacement: $N^n = 4^2 = 4*4 = 16$ possible samples

When sampling without replacement: $\frac{N!}{n!(N-n)!} = \frac{4!}{2!(4-2)!} = \frac{4!}{2! * 2!} = 6$ samples

- **Sampling from an infinite population**

A simple random sample from an infinite population is a sample selected such that each element selected comes from the population and each element is selected independently.

Probability Sampling Methods

In some sampling situations, especially those with large populations, it is time-consuming to select a simple random sample by first finding a random number and then counting or searching through the list of the population until the corresponding element is found. An alternative to simple random sampling is systematic sampling.

- **Systematic sampling**

A method of choosing a sample by selecting every *k*th element after randomly selecting the first element from 1 through k, where k=N/n

Example: Suppose we have 30 car and a sample of 6 is needed. Since 30/6=5, then k=5, thus every 5th car would be selected. However, the first car numbered between 1 and 5 would be selected at random. Suppose car 3 were the first car selected, then the sample would be

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
16	17	18	19	20	21	22	23	24	25	26	27	28	29	30

3	8
13	18
23	28

Probability Sampling Methods

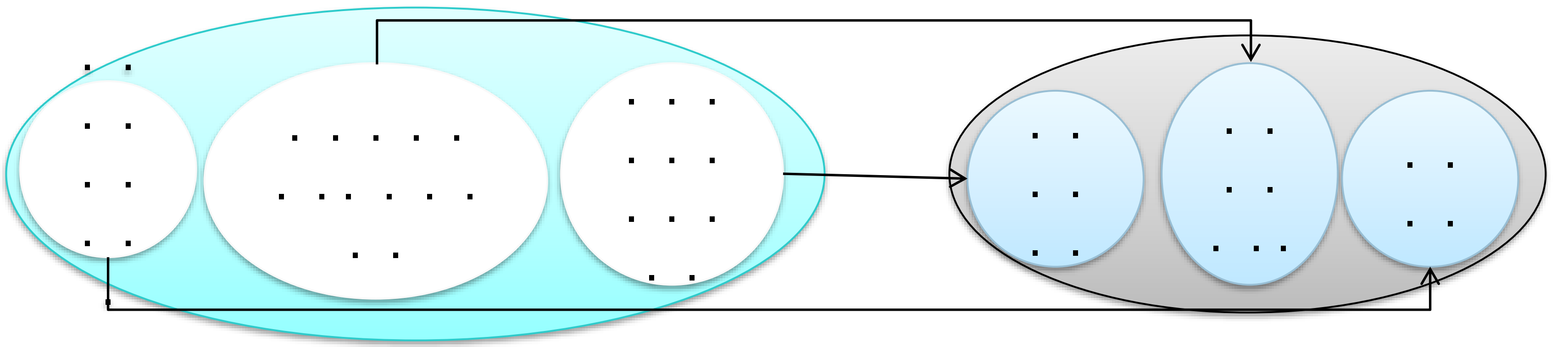
When you need to get a sample that is representative for every group in the population, you should use stratified random sampling.

- **Stratified random sampling**

The elements in the population are first divided into groups called *strata*, such that each element in the population belongs to one and only one stratum. After the strata are formed, a simple random sample is taken from each stratum

The basis for forming the strata, such as department, location, age, industry type, and so on, is at the discretion of the designer of the sample. However, the best results are obtained when the elements within each stratum are as much alike as possible.

Example: Suppose we want to get a sample form a high school but we need to be sure that the sample contains subjects from each school's level.



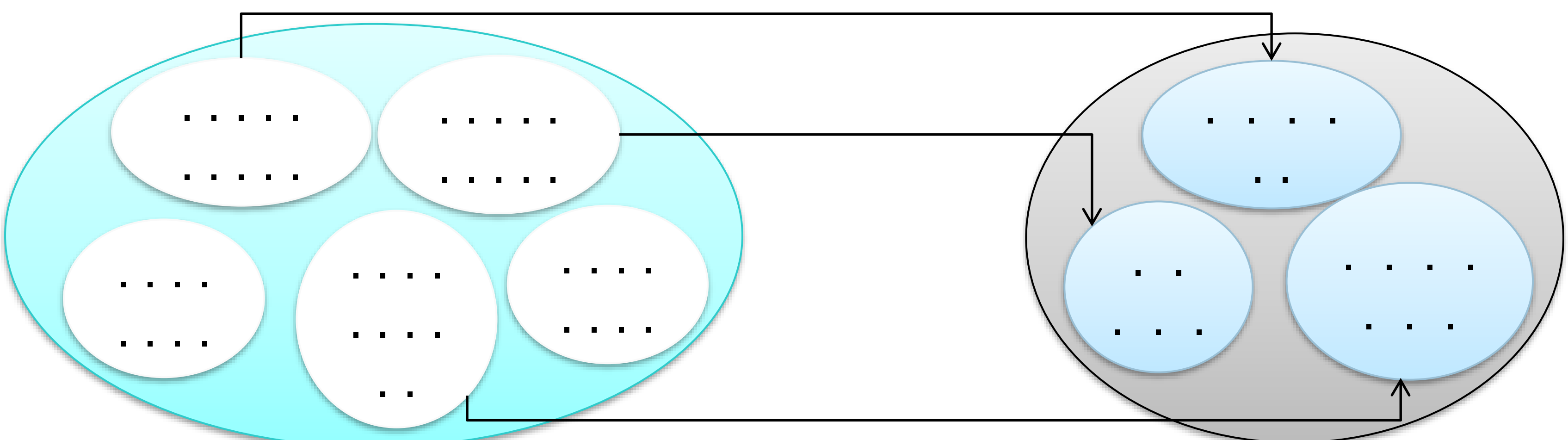
Probability Sampling Methods

When you need to get a sample from some groups in the population, you should use cluster sampling.

- **Cluster sampling**

A probabilistic method of sampling in which the population is first divided into clusters and then one or more clusters are selected for sampling.

- In single stage cluster sampling, every element in each selected cluster is sampled;
- in two-stage cluster sampling, a sample of the elements in each selected cluster is collected.



Non-probability Sampling Methods

- **Convenience Sampling**

Elements are selected for the sample on the basis of convenience of the researcher. It has advantage of relatively easy sample selection and data collection; however, it is impossible to evaluate the “goodness” of the sample in terms of its representativeness of the population.

Example: A professor conducting research at a university may use student volunteers to constitute a sample simply because they are readily available and will participate as subjects for little or no cost.

- **Judgement Sampling**

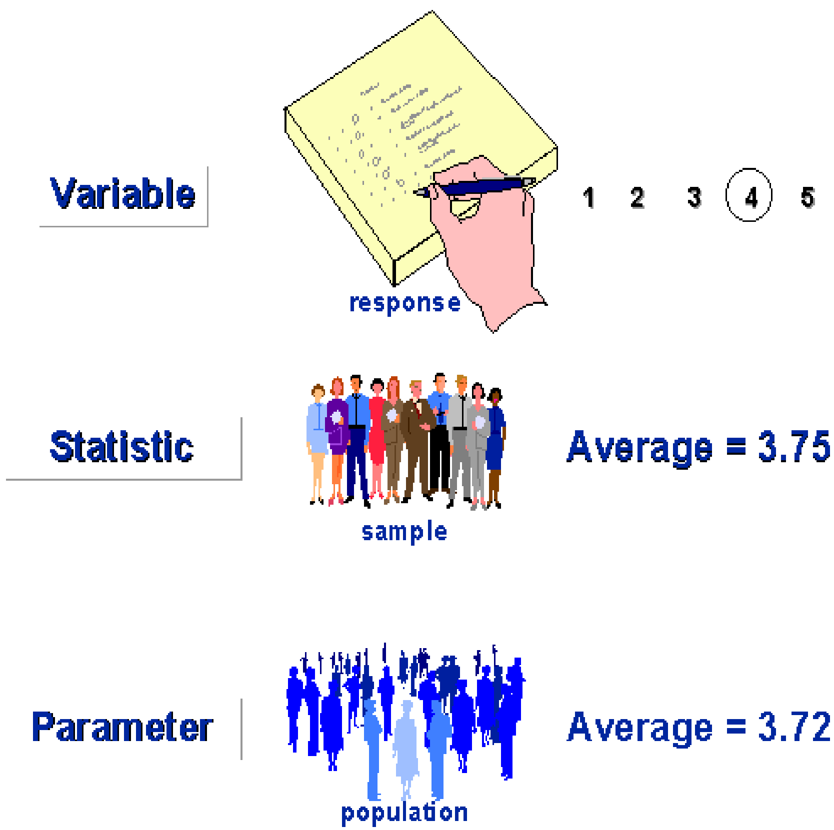
Elements are selected for the sample on the judgement of the person conducting the study. Quality of the sample results depends on the judgment of the person selecting the sample.

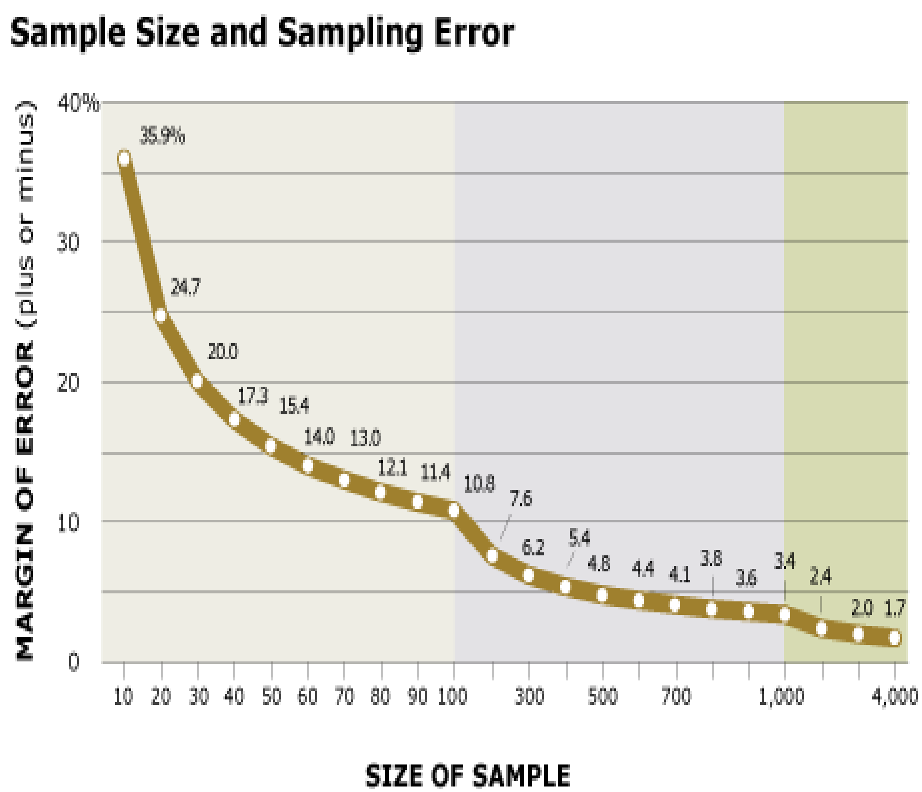
Great caution is warranted in drawing conclusions based on judgment samples used to make inferences about populations.

Example: A reporter may sample two or three senators, judging that those senators reflect the general opinion of all senators.

Sampling Error

The sampling error is the difference between a sample statistic and its corresponding population parameter such as $\bar{x}$ - μ.





Point Estimation

Point estimation is the statistical procedure used to calculate point estimate.

- To estimate the value of a population parameter, we compute a corresponding characteristic of the sample.
- Point estimate

The numerical value of a point estimator used in a particular instance as an estimate of a population parameter.

- Point estimator

The formula of a sample statistic such as $\bar{x}$, S *and* $\bar{p}$ that provides the point estimate of the population parameter.

Population Parameter	Parameter Value	Point Estimator	Point Estimate
μ = Population mean annual salary	\$51,800	$\bar{x}$ = Sample mean annual salary	\$51,814
σ = Population standard deviation for annual salary	\$4000	s = Sample standard deviation for annual salary	\$3348
p = Population proportion having completed the management training program	.60	$\bar{p}$ = Sample proportion having completed the management training program	.63

Properties of Point Estimation

Before using any sample statistic as a point estimator, statisticians check to see whether the sample statistic demonstrates certain properties associated with good point estimators.

Properties of a good point estimators

- Unbiased

If the expected value of the sample statistic is equal to the population parameter being estimated, $E(\hat{\theta}) = \theta$ *where* θ *is* the population parameter of interest and $\hat{\theta}$ is the sample statistic or point estimator of θ

- Efficiency

The point estimator with the smaller standard error is said to have greater **relative efficiency** than the other because it tends to provide estimates closer to the population parameter.

- Consistency

A point estimator is consistent if the values of the point estimator tend to become closer to the population parameter as the sample size becomes larger.

Sampling Distribution

- Sampling distribution A probability distribution consisting of all possible values of a sample statistic such as sample mean, sample standard deviation, sample proportion and so on.

Example: Suppose a population consists of the numbers 2,4,6,8 (N=4) and we are interested in selecting a simple random sample of size (n=2) using sampling *with replacement*.

Formula for sampling with replacement = $N^n = 4^2$=4*=16

Population mean = μ=(2+4+6+8)/4=20/4=5

Population Standard Deviation= σ =2.236

Now we select 16 samples of size 2 from this population which are all possible samples of size 2 that can be taken from this population

S.No	Samples	S.No	Samples	S.No	Samples	S.No	Samples
1	(2,2)	5	(4,2)	9	(6,2)	13	(8,2)
2	(2,4)	6	(4,4)	10	(6,4)	14	(8,4)
3	(2,6)	7	(4,6)	11	(6,6)	15	(8,6)
4	(2,8)	8	(4,8)	12	(6,8)	16	(8,8)

Sampling Distribution of $\bar{x}$

- Sampling distribution of $\bar{x}$

The sampling distribution of $\bar{x}$ is the probability distribution of all possible values of the sample mean.

Proprieties of Sampling distribution of $\bar{x}$

1. Expected value of $\bar{x}$

Consider the $\bar{x}$ values generated by the various possible simple random samples. The mean of all these values is known as the expected value $E(\bar{x})$. It can be shown that with simple random sampling, $E(\bar{x}) = \mu$. Thus, the mean of the sample means is unbiased point estimator for the population mean.

1. Standard deviation of $\bar{x}$

With simple random sampling, the standard deviation of $\bar{x}$ depends on whether the population is finite or infinite.

Finite population	Infinite population
$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\left(\frac{\sigma}{\sqrt{n}}\right)$	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

1. The shape of the sampling distribution of $\bar{x}$

- When the population has a normal distribution, the sampling distribution of $\bar{x}$ is normally distributed for any sample size.

- As the sample size ***n*** increases, the shape of the distribution of the sample means will approach a normal distribution regardless of the original shape (Central Limit Theorem).

Sampling Distribution of $\bar{x}$

Example : In the previous example by calculating the mean and standard deviation (std) of all the possible samples and then calculating the mean and std of samples mean, we found

Mean of the samples mean = $E(\bar{x}) = \Sigma\bar{x}/n$ = 80/16 = 5 = μ

Std of the sample mean = $\sigma_{\bar{x}} = \sqrt{\frac{\Sigma\bar{x}^2 - (\Sigma\bar{x})^2/n}{n}} = \sqrt{\frac{440 - 80^2/16}{16}} = 1.581$

Remember that

Population mean = μ = (2+4+6+8)/4 = 20/4 = 5

Population standard deviation = σ = 2.236

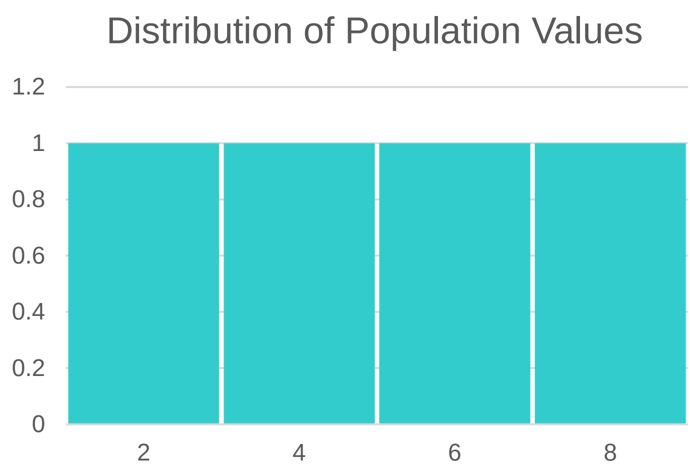
Note that $E(\bar{x}) = \mu = 5$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.236}{\sqrt{2}} = 1.581$

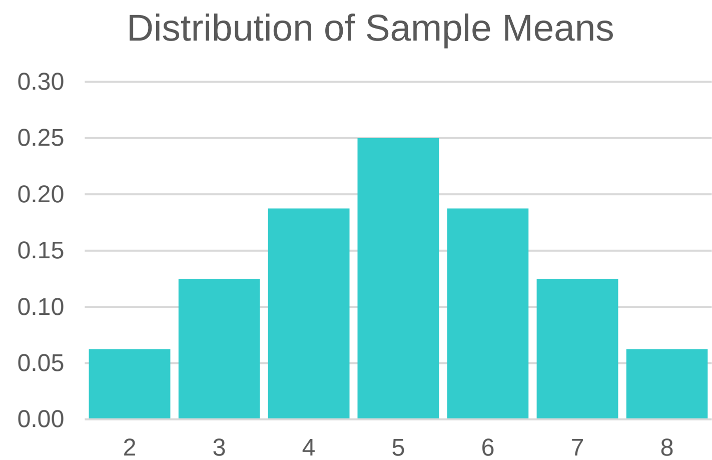
Thus, the first and second properties of $\bar{x}$ sampling distribution are satisfied.

S.No	Samples	Sample means	S.No	Samples	Sample means	S.No	Samples	Sample means	S.No	Samples	Sample means
1	(2,2)	2	5	(4,2)	3	9	(6,2)	4	13	(8,2)	5
2	(2,4)	3	6	(4,4)	4	10	(6,4)	5	14	(8,4)	6
3	(2,6)	4	7	(4,6)	5	11	(6,6)	6	15	(8,6)	7
4	(2,8)	5	8	(4,8)	6	12	(6,8)	7	16	(8,8)	8

Sampling Distribution of $\bar{x}$

Example : In the previous example, the plot of the population values and the sample means data follows:





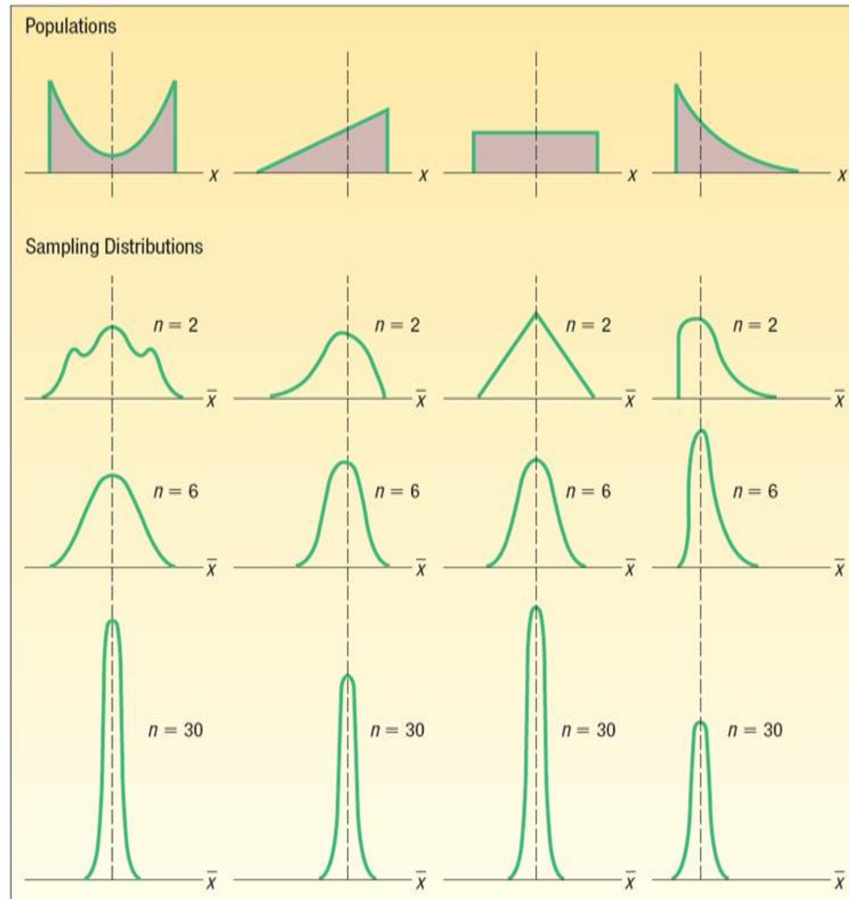
Shape of the population values is *uniform* but the shape of the sample means is approximately *normal*. Thus, the third property of $\bar{x}$ sampling distribution are satisfied

S.No	Samples	Sample means	S.No	Samples	Sample means	S.No	Samples	Sample means	S.No	Samples	Sample means
1	(2,2)	2	5	(4,2)	3	9	(6,2)	4	13	(8,2)	5
2	(2,4)	3	6	(4,4)	4	10	(6,4)	5	14	(8,4)	6
3	(2,6)	4	7	(4,6)	5	11	(6,6)	6	15	(8,6)	7
4	(2,8)	5	8	(4,8)	6	12	(6,8)	7	16	(8,8)	8

Central Limit Theorem

CENTRAL LIMIT THEOREM If all samples of a particular size are selected from any population, the sampling distribution of the sample mean is approximately a normal distribution. This approximation improves with larger samples.

- If the population follows a normal probability distribution, then for any sample size the sampling distribution of the sample mean will also be normal.
- If the population distribution is symmetrical (but not normal), shape of the distribution of the sample mean will emerge as normal with samples as small as 10.
- If a distribution that is skewed or has thick tails, it may require samples of 30 or more to observe the normality feature.



Sampling Distribution of *P*

The sample proportion *P* is a point estimator of the population proportion π.

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

where ***m*** the number of elements in the sample that possess a characteristic
n the sample size.

- Sampling distribution of *P*

The sampling distribution of *P* is the probability distribution of all possible values of the sample proportion *P*.

Proprieties of Sampling distribution of *P*

1. Expected value of *P*

The expected value of *P*, the mean of all possible values of *P*, is equal to the population proportion π. Thus, *P* is an unbiased estimator of π; ***E(P) = π***

1. Standard deviation of *P*

The standard deviation of *P* depends on whether the population is finite or infinite.

Finite population	Infinite population
$\sigma_P = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{\pi(1-\pi)}{n}}$	$\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}}$

1. The shape of the sampling distribution of ***P***

The sampling distribution of *P* can be approximated by a normal distribution whenever $n\pi \geq 5$ and $n(1-\pi) \geq 5$.

Standard Error

The standard deviation of a point estimator is called standard error such as standard error of the mean and standard error of the proportion.

Standard Deviation of $\bar{x}$ (Standard Error)

Finite Population

$$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\left(\frac{\sigma}{\sqrt{n}}\right)$$

Infinite Population

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

Standard Deviation of $\bar{p}$ (Standard Error)

Finite Population

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

Infinite Population

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

Using the Sampling Distribution of the Sample Mean-Example

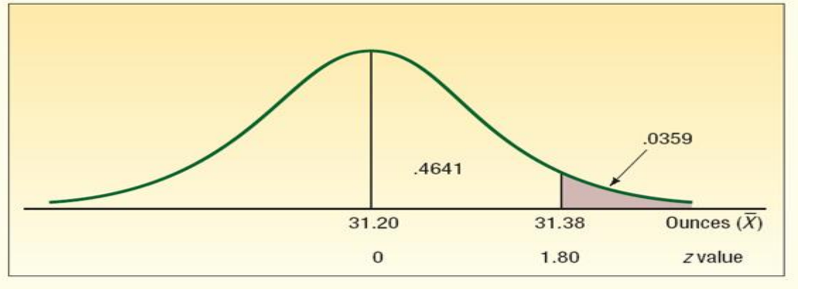
The Quality Assurance Department for Cola, Inc., maintains records regarding the amount of cola in its Jumbo bottle. The actual amount of cola in each bottle is critical, but varies a small amount from a bottle to another. Cola, Inc., does not wish to underfill the bottles or overfill each bottle. Its records indicate that the amount of cola follows the normal probability distribution. The mean amount per bottle is 31.2 ounces and the population standard deviation is 0.4 ounces.

At 8 A.M. today the quality technician randomly selected 16 bottles from the filling line. The mean amount of cola contained in the bottles is 31.38 ounces.

Is this an unlikely result? Is it likely the process is putting too much soda in the bottles? To put it another way, is the sampling error of 0.18 ounces unusual?

Thus, we need to find P($\bar{x}$ > 31.38)

P(lower)	P(upper)	z	X	mean	std.dev
.9641	**.0359**	1.80	31.38	31.20	0.10



Conclusion: It is unlikely, less than a 4 percent chance, we could select a sample of 16 observations from a normal population with a mean of 31.2 ounces and a population standard deviation of 0.4 ounces and find the sample mean equal to or greater than 31.38 ounces. The process is putting too much cola in the bottles.

Summary

Sampled population The population from which the sample is taken.

Frame A listing of the elements the sample will be selected from.

Simple random sample A simple random sample of size n from a finite population of size N is a sample selected such that each possible sample of size n has the same probability of being selected.

Random sample A random sample from an infinite population is a sample selected such that each element selected comes from the same population and is independently selected.

Sampling without replacement Once an element has been included in the sample, it is removed from the population and cannot be selected a second time.

Sampling with replacement Once an element has been included in the sample, it is returned to the population. A previously selected element can be selected again and therefore may appear in the sample more than once.

Stratified random sampling A probability sampling method in which the population is first divided into strata and a simple random sample is then taken from each stratum.

Cluster sampling A probability sampling method in which the population is first divided into clusters and then a simple random sample of the clusters is taken.

Summary

Systematic sampling A probability sampling method in which we randomly select one of the first k elements and then select every kth element thereafter.

Convenience sampling A nonprobability method of sampling whereby elements are selected for the sample on the basis of convenience.

Judgment sampling A nonprobability method of sampling whereby elements are selected for the sample based on the judgment of the person doing the study.

Parameter A numerical characteristic of a population, such as a population mean μ, a population standard deviation σ, a population proportion p, and so on.

Sample statistic A sample characteristic, such as sample mean , sample standard deviation, sample proportion and so on. The value of the sample statistic is used to estimate the value of the corresponding population parameter.

Point estimator The sample statistic, such as μ, s, or P that provides the point estimate of the population parameter.

Point estimate The value of a point estimator used in a particular instance as an estimate of a population parameter.

Target population The population for which statistical inference such as point estimates are made. It is important for the target population to correspond as closely as possible to the sampled population.

Summary

Unbiased A property of a point estimator that is present when the expected value of the point estimator is equal to the population parameter it estimates.

Relative efficiency Given two unbiased point estimators of the same population parameter, the point estimator with the smaller standard error is more efficient.

Consistency A property of a point estimator that is present whenever larger sample sizes tend to provide point estimates closer to the population parameter.

Sampling distribution A probability distribution consisting of all possible values of a sample statistic.

Finite population correction factor The term that is used in the formulas for and whenever a finite population, rather than an infinite population, is being sampled. The generally accepted rule of thumb is to ignore the finite population correction factor whenever $n/N \geq 0.05$.

Central limit theorem A theorem that enables one to use the normal probability distribution to approximate the sampling distribution of whenever the sample size is large.

Standard error The standard deviation of a point estimator.