**Notes– Ch 7 Sampling and Sampling Distribution**

Statisticians use the word population to refer not only to people but to all items that have been chosen for study. Statisticians use the word sample to describe a portion chosen from the population. Often the cost of collecting information from a sample is substantially less than from a population, especially when personal interviews must be conducted to collect the information.

Several methods can be used to select a sample from a population:

**Probability Sampling technique –** Elements selected from the population have a known probability of being included in the sample. The probability sampling techniques are random sampling, stratified sampling, cluster sampling and systematic sampling

**1) Random sampling:** Simple random sampling selects samples by methods that allow each possible sample to have an equal probability of being picked and each item in the population to have an equal chance of being included in the sample. Random numbers are used to select the samples.

SIMPLE RANDOM SAMPLE (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. There are two ways of sampling:

* Sampling without replacement
* Sampling with replacement

RANDOM SAMPLE (INFINITE POPULATION) A random sample of size n from an infinite population is a sample selected such that the following conditions are satisfied.

* Each element selected comes from the same population.
* Each element is selected independently.

**2) Stratified random sampling**: In 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. 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. After the strata are formed, a simple random sample is taken from each stratum. Formulas are available for combining the results for the individual stratum samples into one estimate of the population parameter of interest. The value of stratified random sampling depends on how homogeneous the elements are within the strata. If elements within strata are alike, the strata will have low variances. Thus, relatively small sample sizes can be used to obtain good estimates of the strata characteristics. If strata are homogeneous, the stratified random sampling procedure provides results just as precise as those of simple random sampling by using a smaller total sample size. Stratified random sampling works best when the variance among elements in each stratum is relatively small.

For example, sampling from different age groups, income groups, employee grades

**3) Cluster Sampling:** A method of random sampling in which the population is divided into groups or clusters of elements, and then a random sample of these clusters is selected. We assume that these individual clusters are representative of the population as a whole. In the ideal case, each cluster is a representative small-scale version of the entire population. The value of cluster sampling depends on how representative each cluster is of the entire population. If all clusters are alike in this regards, sampling a small number of clusters will provide good estimates of the population parameters.

One of the primary applications of cluster sampling is area sampling, where clusters are city blocks or other well-defined areas. Cluster sampling generally requires a larger total sample size than either simple random sampling or stratified random sampling. However, it can result in cost savings because of the fact that when an interviewer is sent to a sampled cluster (e.g., a city-block location), many sample observations can be obtained in a relatively short time. Hence, a larger sample size may be obtainable with a significantly lower total cost.

With both stratified and cluster sampling, the population is divided into well-defined groups. We use stratified sampling when each group has small variations within itself but there is a wide-variations between the groups. We use cluster sampling when each group has considerable variations within itself but the groups are essentially similar to each other.

**4) Systematic sampling:** 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. For example, if a sample size of 50 is desired from a population containing 5000 elements, we will sample one element for every 5000/50 = 100 elements in the population. A systematic sample for this case involves selecting randomly one of the first 100 elements from the population list. Other sample elements are identified by starting with the first sampled element and then selecting every 100th element that follows in the population list. In effect, the sample of 50 is identified by moving systematically through the population and identifying every 100th element after the first randomly selected element. The sample of 50 usually will be easier to identify in this way than it would be if simple random sampling were

used. Because the first element selected is a random choice, a systematic sample is usually

assumed to have the properties of a simple random sample. This assumption is especially applicable when the list of elements in the population is a random ordering of the elements. Eg. Interviewing every fifth visitor, taking feedback from every ninth customer

**Non-Probability Sampling technique-** Sampling methods which do not use probabilities for selection of sample. The non-probability sampling techniques are Convenience sampling and Judgement sampling

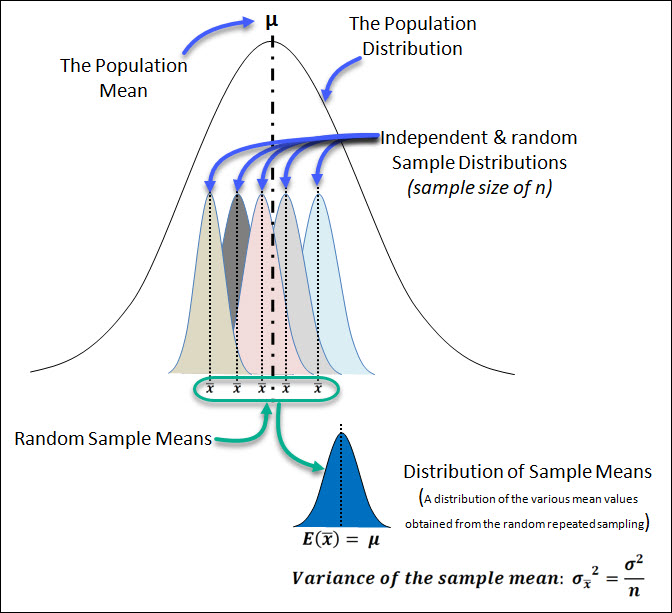
**1) Convenience sampling:** As the name implies, the sample is identified primarily by convenience. Elements are included in the sample without prespecified or known probabilities of being selected. For example, a professor conducting research at a university may use student volunteers to constitute a sample simply because they are readily available and will participate as subjects for little or no cost. Similarly, an inspector may sample a shipment of oranges by selecting oranges haphazardly from among several crates. Labelling each orange and using a probability method of sampling would be impractical. Samples such as wildlife captures and volunteer panels for consumer research are also convenience samples. Convenience samples have the 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. A convenience sample may provide good results or it may not; no statistically justified procedure allows a probability analysis and inference about the quality of the sample results. Sometimes researchers apply statistical methods designed for probability samples to a convenience sample, arguing that the convenience sample can be treated as though it were a probability sample. However, this argument cannot be supported, and we should be cautious in interpreting the results of convenience samples that are used to make inferences about populations.

**2) Judgement sampling:** In this approach, the person most knowledgeable on the subject of the study selects elements of the population that he or she feels are most representative of the population. Often this method is a relatively easy way of selecting a sample. For example, a reporter may sample two or three senators, judging that those senators reflect the general opinion of all senators. However, the quality of the sample results depends on the judgment of the person selecting the sample. Again, great caution is warranted in drawing conclusions based on judgment samples used to make inferences about populations.

**Sampling Distribution:** The Sampling Distribution imagines what would happen if we took repeated samples of the same size from the same (or similar) populations done under the identical conditions. From this hypothetical experiment we “build” a probability mass (or density) function or probability distribution function that is used to determine probabilities for various hypothetical outcomes.

**Point estimator:** The sample statistic, such as , that provides the point estimate of the population parameters .

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



**Sampling Distribution of mean:** A probability distribution of all possible means of samples of a given size n, from a population. In practice, the size and character of most populations prohibit decision makers from taking all the possible samples from a population distribution. In most cases, decision makers take only one sample from the population, calculate statistics for that sample, and from those statistics infer something about the parameters of the entire population. Just as with other probability distributions, the sampling distribution of has an expected value or mean, a standard deviation, and a characteristic shape or form.

**Expected Value:** The expected value of equals the mean of the population from which the

sample is selected.

**Standard Deviation:** The standard deviation of the sampling distribution of depends on whether the population is finite or infinite.

The factor is required for the finite population case but not for the infinite population case. This factor is commonly referred to as the **finite population correction factor**. In many practical sampling situations, we find that the population involved, although finite, is “large,” whereas the sample size is relatively “small.” In such cases the finite population correction factor is close to 1. As a result, the difference between the values of the standard deviation of for the finite and infinite population cases becomes negligible.

Thus, becomes a good approximation to the standard deviation of whenever

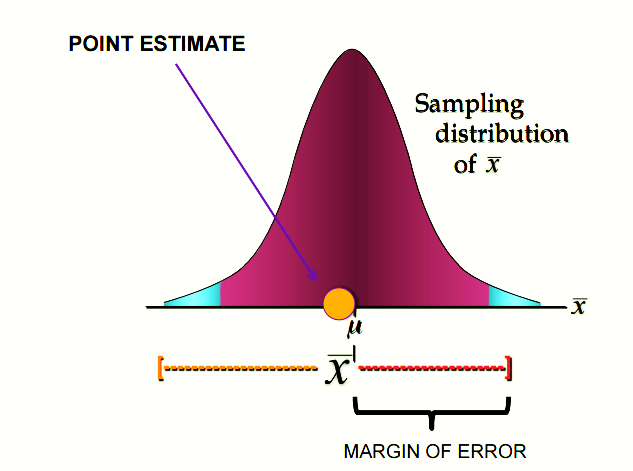
1. The population is infinite; or

2. The population is finite and the sample size is less than or equal to 5% of the

population size; that is,

**Standard error ():** Rather than say “standard deviation of the distribution of sample

mean” to describe a distribution of sample means, statisticians refer to the standard error of the mean. The term standard error is used throughout statistical inference to refer to the standard deviation of a point estimator. Standard error is denoted by where as standard deviation of the population is denoted by . The term standard error is used because it conveys specific meaning. While taking samples, it is highly unlikely that all the sample means would be the same. We expect to see some variability in our observed means. This variability in the sample statistics results from the sampling error due to chance, that is, there are differences between each sample and the population, and among the several samples, owing solely to the elements we happened to choose for the samples.



Standard error

The standard deviation of the distribution of sample means measures the extent to which we expect the means from the different samples to vary because of this chance error in the sampling process. Thus, the standard deviation of the point estimate is known as the standard error of the statistics.

The standard error indicates not only the size of the chance error that has been made, but also the accuracy we are likely to get if we use a sample statistics to estimate a population parameter. A distribution of sample means that is less spread out is a better estimator of the population mean than a distribution of sample means that is widely dispersed and has a large standard error.

**Form of the Sampling Distribution:** To determine the form or shape of the sampling distribution of , we will consider two cases:

(1) Population has a normal distribution: In many situations it is reasonable to assume that the population from which we are selecting a random sample has a normal, or nearly normal distribution. When the population has a normal distribution, the sampling distribution of is normally distributed for any sample size.

(2) Population does not have a normal distribution: When the population from which we are selecting a random sample does not have a normal distribution, the central limit theorem is helpful in identifying the shape of the sampling distribution of . A statement of the central limit theorem as it applies to the sampling distribution of follows.

**Central Limit Theorem**: In selecting random samples of size n from a population, the sampling distribution of the sample mean can be approximated by a normal distribution as the sample size become large. The significance of the central limit theorem is that it permits us to use samples statistics to make inference about population parameters without knowing anything about the shape of the frequency distribution of that population other than what we can get from the sample. For most applications, the sampling distribution of can be approximated by a normal distribution whenever the sample is size 30 or more. In cases where the population is highly skewed or outliers are present, samples of size 50 may be needed. Finally, if the population is discrete, the sample size needed for a normal approximation often depends on the population proportion.

**Relationship Between the Sample Size and the Sampling Distribution of**

regardless of the sample size. Thus, the mean of all possible values of is equal to the population mean μ regardless of the sample size n. However, note that the standard error of the mean, , is related to the square root of the sample size. Whenever the sample size is increased, the standard error of the mean, decreases. as the sample size is increased, the standard error of the mean decreases. As a result, the larger sample size provides a higher probability that the sample mean is within a specified distance of the population mean.

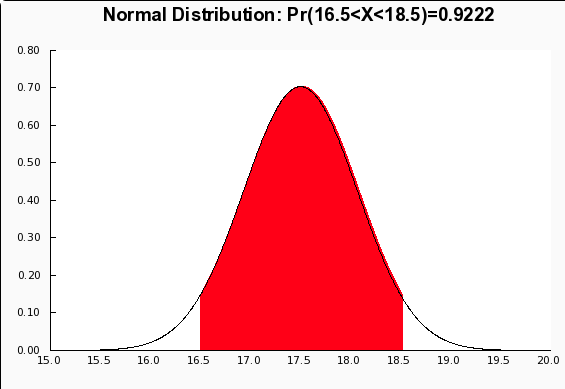
**Example:** Barron’s reported that the average number of weeks an individual is unemployed is 17.5 weeks and that the population standard deviation is 4 weeks. Suppose you would like to select a random sample of 50 unemployed individuals for a follow-up study.

a. Show the sampling distribution of the sample mean average for a sample of 50 unemployed individuals.

b. What is the probability that a simple random sample of 50 unemployed individuals

will provide a sample mean within 1 week of the population mean?

**Ans.** a. Here,

****The sampling distribution of is given by

E( = 17.5 and = 0.567

b. For = 3

For = 5

There is 0.9222 probability that the random sample with provide a sample mean with 1 week of the population mean.

**Sampling Distribution of proportion:** A probability distribution of all possible proportions of samples of a given size n, from a population is called sampling distribution of .

where x is number of elements in the sample that possess the characteristic of interest and n is the sample size

**Expected Value:** The expected value of equals the mean of all possible values of is equal to the population proportion p.

**Standard Deviation:** The standard deviation of the sampling distribution of depends on whether the population is finite or infinite.

The factor is required for the finite population case but not for the infinite population case. This factor is commonly referred to as the **finite population correction factor**. In many practical sampling situations, we find that the population involved, although finite, is “large,” whereas the sample size is relatively “small.” In such cases the finite population correction factor is close to 1. As a result, the difference between the values of the standard deviation of for the finite and infinite population cases becomes negligible.

Thus, becomes a good approximation to the standard deviation of whenever

1. The population is infinite; or

2. The population is finite and the sample size is less than or equal to 5% of the

population size; that is,

Similarly, “standard deviation of the distribution of sample proportion” is shortened to the standard error of the proportion.

**Form of the Sampling Distribution of :**

Now that we know the mean and standard deviation of the sampling distribution of , the

final step is to determine the form or shape of the sampling distribution. The sample proportion is . For a simple random sample from a large population, the value of x is a binomial random variable indicating the number of elements in the sample with the characteristic of interest. Because n is a constant, the probability of x/n is the same as the

binomial probability of x, which means that the sampling distribution of is also a discrete

probability distribution and that the probability for each value of x/n is the same as the

probability of x.

A a binomial distribution can be approximated by a normal distribution whenever the sample size is large enough to satisfy the following two conditions:

Assuming these two conditions are satisfied, the probability distribution of x in the sample

proportion, , can be approximated by a normal distribution. And because n is a constant, the sampling distribution of can also be approximated by a normal distribution. This approximation is stated as follows:

The sampling distribution of can be approximated by a normal distribution whenever

In practical applications, when an estimate of a population proportion is desired, we find that sample sizes are almost always large enough to permit the use of a normal approximation

for the sampling distribution of .

**Example:** The president of Doerman Distributors, Inc., believes that 30% of the firm’s orders come from first-time customers. A random sample of 100 orders will be used to estimate the proportion of first-time customers.

a. Assume that the president is correct and p = 0.30. What is the sampling distribution of

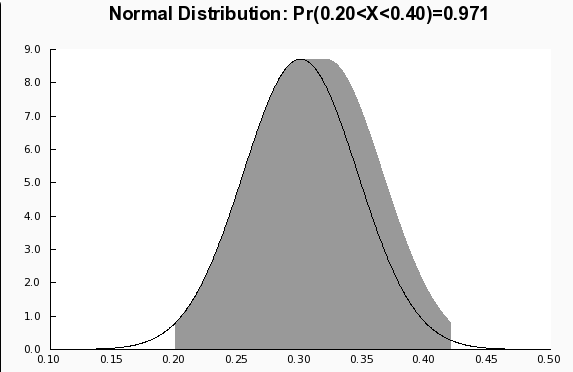
for this study?

b. What is the probability that the sample proportion will be between 0.20 and 0.40?

**Ans.** a. Here,

The sampling distribution of is given by

E( = 0.3 and = 0.0458

b. For = 0.20

For = 0.40

There is 0.9709 probability that the sample proportion will be between 0.20 and 0.30

**Properties of Point Estimators**: Sample statistics such as a sample mean , a sample standard deviation s, and a sample proportion can be used as point estimators of their corresponding population parameters μ, σ, and p. It is intuitively appealing that each of these sample statistics is the point estimator of its corresponding population parameter. However, before using a sample statistic as a point estimator, statisticians check to see whether the sample statistic demonstrates certain properties associated with good point estimators.

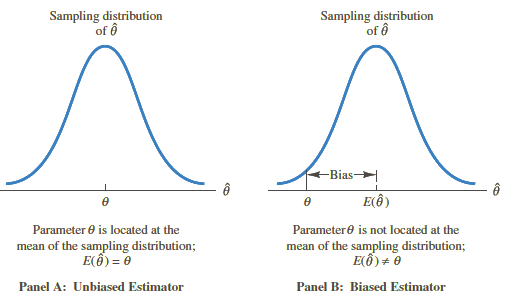
The three properties of good point estimators: unbiased, efficiency, and consistency.

Because several different sample statistics can be used as point estimators of different

population parameters, we use the following general notation.

**1) Unbiased:** If the expected value of the sample statistic is equal to the population parameter being estimated,the sample statistic is said to be an unbiased estimator of the population parameter.

Hence, the expected value, or mean, of all possible values of an unbiased sample statistic is equal to the population parameter being estimated.

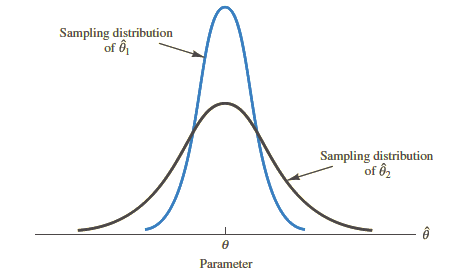
Figure shows the cases of unbiased and biased point estimators. In the illustration showing the unbiased estimator, the mean of the sampling distribution is equal to the value of

the population parameter. The estimation errors balance out in this case, because sometimes the value of the point estimator may be less than θ and other times it may be greater than θ.

In the case of a biased estimator, the mean of the sampling distribution is less than or greater than the value of the population parameter. In the illustration in Panel B, E( ) is greater than θ; thus, the sample statistic has a high probability of overestimating the value of the population parameter. The amount of the bias is shown in the figure.

In discussing the sampling distributions of the sample mean and the sample proportion, we stated that and . Thus, both and are unbiased estimators of their corresponding population parameters μ and p.

In the case of the sample standard deviation s and the sample variance s2, it can be shown that E(s2) = σ2. Thus, we conclude that the sample variance s2 is an unbiased estimator of the population variance σ2. In fact, when we first presented the formulas for the sample variance and the sample standard deviation in Chapter 3, n - 1 rather than n was used in the denominator. The reason for using n - 1 rather than n is to make the sample variance an unbiased estimator of the population variance.

**2) Efficiency:** Assume that a simple random sample of n elements can be used to provide two unbiasedpoint estimators of the same population parameter. In this situation, we would prefer to usethe point estimator with the smaller standard error, because it tends to provide estimates

closer to the population parameter. The point estimator with the smaller standard error is said to have greater relative efficiency than the other.

Figure shows the sampling distributions of two unbiased point estimators, Note that the standard error of is less than the standard error of ; thus, values of have a greater chance of being close to the parameter θ than do values of . Because the standard error of point estimator is less than the standard error of point estimator , is relatively more efficient than and is the preferred point estimator.

**3) Consistency:** A third property associated with good point estimators is consistency. Loosely speaking,a point estimator is consistent if the values of the point estimator tend to becomecloser to the population parameter as the sample size becomes larger. In other words, a large sample size tends to provide a better point estimate than a small sample size. Note that for the sample mean , we showed that the standard error of is given by ,.

Because is related to the sample size such that larger sample sizes provide smaller values for , we conclude that a larger sample size tends to provide point estimates closer to the population mean μ. In this sense, we can say that the sample mean is a consistent estimator of the population mean μ. Using a similar rationale, we can also conclude that the sample proportion , is a consistent estimator of the population proportion p.