

SAMPLING AND SAMPLING DISTRIBUTIONS

1.1. Sampling Theory

A sample is a tool to infer something about a population. We begin by discussing methods of selecting a sample from a population. Next, we construct a distribution of the samples to understand how the sample means tend to cluster around the population means and that the shape of this distribution tends to follow the normal distribution.

The Need for Sampling

In many cases sampling is the only way to determine something about the population. Some of the major reasons why sampling is necessary are:

1. The destructive nature of certain tests

If wine tasters at a quality control drink all the wine to evaluate the vintage content, they will consume the entire wine produced by **Awash Wine Factory** and none would be available for sale.

2. The physical impossibility of checking all items in the population.

The population of fish, birds, snakes, mosquitoes, and the like are large and are constantly moving, being born, and dying. So taking population census is impossible.

3. The cost of studying all the items in a population is often prohibitive.

Television executives of ETV want to know the proportion of television viewers who watch the “*Fegegta Talk show*”. Since more than 20 million people in the country may be watching ETV programs on a given evening, determining the proportions that are watching “*Fegegta Talk show*” program is prohibitively expensive.

4. The adequacy of sample results

Even if funds were available (for the case of ETV above), it is doubtful the additional accuracy of a 100 percent sample that is studying the entire population is essential in most problems. As you can see from the above example, the sample results are more than enough to know about the population under consideration.

5. To contact the whole population would often be time consuming

A candidate for a national office may wish to determine her chances for election. A sample poll using the regular staff and field interviews of a professional polling firm would take only one or two days. By using the same staff and interviewers and working seven days a week, it would take nearly 200 years to contact all the voting population.

3.1.2 Sampling Methods

In general, there are two types of samples: *a probability sample* and *non probability sample*.

1. Probability Sample

A probability samples is a sample selected in such a way that each item or person in the population has a known (nonzero) likelihood of being included in the sample. If probability sampling is done, each item in the population has a chance of being chosen.

2. Non probability Sampling

If non probability methods are used, not all items or people have a chance of being included in the sample. In such instances the results may be biased, meaning that the sample results may not be representative of the population. Judgmental sampling and convenience sampling are two non probability methods. For example a panel is formed to solicit opinions on a newly developed cat food and there are 2000 cat owners selected for the sample. Selection of the members is based on the judgment of the person conducting the research, and the sample results may therefore not be representative of the entire population of cat owners (since not all cat owners have a chance of being chosen).

The difference between non probability and probability sampling is that non probability sampling does not involve *random* selection but probability sampling does. Does that mean that non probability samples aren't representative of the population? Not necessarily. But it does mean that non probability samples cannot depend upon the rationale of probability theory. At least with a probabilistic sample, we know the odds or probability that we have represented the population well. We are able to estimate confidence intervals for the statistic. With non probability samples, we may or may not represent the population well, and it will often be hard for us to know how well we've done so. In general, researchers prefer probabilistic or random sampling methods over non probabilistic ones, and consider them to be more accurate and rigorous. However, in applied social research there may be circumstances where it is not feasible, practical or theoretically sensible to do random sampling.

The statistical procedures used in this material are based on probability sampling. Therefore, only the methods of probability sampling will be discussed in the following section.

There is no one “best” method of selecting a probability sample from a population of interest. A method used to select a sample of invoices in a file drawer, might not be the most appropriate method for choosing a national sample of voters. However, all probability sampling methods have a similar goal, namely, to allow chance to determine the items or persons to be included in the sample.

Types of probability sampling

Generally there are four probability samplings. Each of these are discussed as follows:

1. Simple random sampling

The most widely used type of sampling is simple random sample. In simple random sampling, a sample selected so that each item or person in the population has the same chance of being included.

Example: Suppose a population consists of 845 employees of Sheraton Addis hotel. A sample of 52 employees is to be selected from that population. One way of ensuring that every employee in the population has a chance of being chosen is to first write the name of each employee on a small slip of paper and deposit all of the slips in a box. After they have been mixed, the first selection is made by drawing a

slip out of the box without looking at it. This process is repeated until the sample size of 52.

2. Systematic random sample

The items or individuals of the population are arranged in some way alphabetically in a file drawer by date received or by some other method. A random starting point is selected, and then every K^{th} member of the population is selected for the sample. In a systematic sample the first item is chosen at random.

3. Stratified Random Sample

A population is divided into subgroups, called strata, and a sample is selected from each stratum. After the population has been divided into strata, either a proportional or a non-proportional sample can be selected. As the name implies, a proportional sampling procedure requires that the number of items in each stratum be in the same proportion as in the population. In a non-proportional stratified sample, the number of items chosen in each stratum is disproportionate to the respective numbers in the population. We then weight the sample results according to the stratum's proportion of the total population. Regardless of whether a proportional or a non-proportional sampling procedure is used every item or person in the population has a chance of being selected for the sample.

4. Cluster sampling

Other common type of sampling is cluster sampling. It is often employed to reduce the cost of sampling a population scattered over a large geographic area. Suppose you want to determine the views of industrialists in a country about Federal and regional government on environmental protection policies. Selecting a simple random of industrialists in the country and personally contacting each one would be time consuming and very expensive. Instead, you could employ cluster sampling by subdividing the country into small units either regions or zones. These are often called **primary units**. Suppose you divided the country into 14 regions, then selected at random four regions -2, 4, 7, and 12 and concentrated your efforts in these primary units you could take a random sample of the industrialists in each of these regions and interview them. (Note that this is a combination of cluster sampling and simple random sampling).

5. Multi-stage sampling

The four methods we have covered so far: simple, stratified, systematic and cluster are the simplest random sampling strategies. In most real applied social research, we would use sampling methods that are considerably more complex than these simple variations. The most important principle here is that we can combine the simple methods described earlier in a variety of useful ways that help us address our sampling needs in the most efficient and effective manner possible. When we combine sampling methods, we call this **multi-stage sampling**.

1.2. Sampling Distributions

A sampling distribution is a probability distribution for the possible values of a sample statistic, such as a sample mean.

NOTE: The normal probability distribution is used to determine probabilities for the normally distributed individual measurements, given the mean and the standard deviation. Symbolically, the variable is the measurement X , with the population mean μ and population standard deviation δ . In contrast to such distributions of individual measurements, a sampling distribution is a probability distribution for the possible values of a sample statistic.

A. Sampling Distribution of the Mean

The sampling distribution of the mean is the probability distributions of the means, $\bar{X}$ of all simple random samples of a given sample size n that can be drawn from the population.

NB: the sampling distribution of the mean is not the sample distribution, which is the distribution of the measured values of X in one random sample. Rather, the sampling distribution of the mean is the probability distribution for $\bar{X}$, the sample mean.

For any given sample size n taken from a population with mean μ and standard deviation δ , the value of the sample mean $\bar{X}$ would vary from sample to sample if several random samples were obtained from the population. This variability serves as the basis for sampling distribution.

The sampling distribution of the mean is described by two parameters: the expected value ($\bar{X}$) = $\bar{\bar{X}}$, or mean of the sampling distribution of the mean, and the standard deviation of the mean $\delta_{\bar{x}}$, the standard error of the mean.

Properties of the Sampling Distribution of Means

1. The mean of the sampling distribution of the means is equal to the population mean. $\mu = \mu_{\bar{x}} = \bar{\bar{X}}$.
2. the standard deviation of the sampling distribution of the means (standard error) is equal to the population standard deviation divided by the square root of the sample size: $\delta_{\bar{x}} = \delta/\sqrt{n}$. This hold true if and only of $n < 0.05N$

and N is very large. If N is finite and $n \geq 0.05N$, $\delta_{\bar{x}} = \frac{\delta}{\sqrt{n}} * \sqrt{\frac{N-n}{N-1}}$. The

expression $\sqrt{\frac{N-n}{N-1}}$ is called finite population correction factor/finite population multiplier. In the calculation of the standard error of the mean, if the population standard deviation δ is unknown, the standard error of the mean $\delta_{\bar{x}}$, can be estimated by using the sample standard error of the mean

$S_{\bar{x}}$ which is calculated as follows: $S_{\bar{x}} = S/\sqrt{n}$ or $S_{\bar{x}} = \frac{S}{\sqrt{n}} * \sqrt{\frac{N-n}{N-1}}$.

3. The sampling distribution of means is approximately normal for sufficiently large sample sizes ($n \geq 30$).

Example:

A population consists of the following ages: 10, 20, 30, 40, and 50. A random sample of three is to be selected from this population and mean computed. Develop the sampling distribution of the mean.

Solution:

The number of simple random samples of size n that can be drawn without replacement from a population of size N is ${}^N C_n$. With $N = 5$ and $n = 3$, ${}^5 C_3 = 10$ samples can be drawn from the population as:

Sampled items	Sample means ($\bar{X}$)
10, 20, 30	20.00
10, 20, 40,	23.33
10, 20, 50	26.67
10, 30, 40	26.67
10, 30, 50	30.00
10, 40, 50	33.33
20, 30, 40	30.00
20, 30, 50	33.33
20, 40, 50	36.67
30, 40, 50	40.00

A systematic organization of the above figures gives the following:

Sample mean ($\bar{X}$)	Frequency	Prob. (relative freq.) of $\bar{X}$
20.00	1	0.1
23.33	1	0.1
26.67	2	0.2
30.00	2	0.2
33.33	2	0.2
36.67	1	0.1
40.00	1	0.1
	10.00	1.00

- ➡ Columns 1 and 2 show frequency distribution of sample means.
- ➡ Columns 1 and 3 show sampling distribution of the mean.

$$\mu = \frac{\sum X}{N} = \frac{\sum \bar{x}}{n} = 30, \text{ regardless of the sample size } \mu = \bar{\bar{X}}.$$

$$\sigma = \sqrt{\frac{\sum (X_i - \bar{X})^2}{N}} = \sqrt{\frac{1000}{5}} = 14.142.$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} * \sqrt{\frac{N-n}{N-1}} = \frac{14.142}{\sqrt{3}} * \sqrt{\frac{5-3}{5-1}} = 5.774$$

$$= \sqrt{\frac{\sum (\bar{X}_i - \bar{\bar{X}})^2}{N}} = \sqrt{\frac{333.4}{10}} = 5.774$$

➡ Since averaging reduces variability $\sigma_{\bar{X}} < \sigma$ except the cases where $\sigma = 0$ and $n = 1$.

Central Limit Theorem and the Sampling Distribution of the Mean

The Central Limit Theorem (CLT) states that:

1. If the population is normally distributed and population standard deviation is known, the distribution of sample means is normal regardless of the sample size.
2. If the population from which samples are taken is not normal, the distribution of sample means will be approximately normal if the sample size (n) is sufficiently large ($n \geq 30$). The larger the sample size is used, the closer the sampling distribution is to the normal curve.

The relationship between the shape of the population distribution and the shape of the sampling distribution of the mean is called the Central Limit Theorem.

The significance of the Central Limit Theorem is that it permits us to use sample 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. It also permits us to use the normal distribution (curve) for analyzing distributions whose shape is unknown. It creates the potential for applying the normal distribution to many problems when the sample is sufficiently large.

Example:

1. The distribution of annual earnings of all bank tellers with five years of experience is skewed negatively. This distribution has a mean of Birr 15,000 and a standard deviation of Birr 2000. If we draw a random sample of 30 tellers, what is the probability that their earnings will average more than Birr 15,750 annually?

Solution:

Steps:

1. Calculate μ and $\sigma_{\bar{X}}$

$$\mu = \text{Birr } 15,000$$

$$\sigma_{\bar{X}} = \sigma / \sqrt{n} = 2000 / \sqrt{30} = \text{Birr } 365.15$$

2. Calculate Z for $\bar{X}$

$$Z_{\bar{X}} = \frac{\bar{X} - \bar{X}}{\delta_{\bar{X}}} = \frac{\bar{X} - \mu}{\delta_{\bar{X}}}$$

$$Z_{15,750} = \frac{15,750 - 15,000}{365} = +2.05$$

3. Find the area covered by the interval

$$\begin{aligned} P(\bar{X} > 15,750) &= P(Z > +2.05) \\ &= 0.5 - P(0 \text{ to } +2.05) \\ &= 0.5 - 0.47892 \\ &= 0.02018 \end{aligned}$$

4. Interpret the results

There is a 2.02% chance that the average earning being more than Birr 15, 750 annually in a group of 30 tellers.

2. 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 of randomly selecting 49 different shopping hours, counting the shoppers, and having the sample mean fall between 441 and 446 shoppers, inclusive?

Solution:

1. Calculate μ and $\delta_{\bar{x}}$

$$\begin{aligned} \mu &= 448 \text{ shoppers} \\ \delta_{\bar{x}} &= \delta/\sqrt{n} = 21/\sqrt{49} = 3 \end{aligned}$$

2. Calculate Z for $\bar{X}$

$$Z_{\bar{X}} = \frac{\bar{X} - \bar{X}}{\delta_{\bar{X}}} = \frac{\bar{X} - \mu}{\delta_{\bar{X}}}$$

$$Z_{441} = \frac{441 - 448}{3} = -2.33 \qquad Z_{446} = \frac{446 - 448}{3} = -0.67$$

3. Find the area covered by the interval

$$\begin{aligned} P(441 \leq \bar{X} \leq 446) &= P(-2.33 \leq Z \leq -0.67) \\ &= P(0 \text{ to } -2.33) - P(0 \text{ to } -0.67) \\ &= 0.49010 - 0.24857 \\ &= 0.24153 \end{aligned}$$

4. Interpret the results

There is a 24.153% chance of randomly selecting 49 hourly periods for which the sample mean falls between 441 and 446 shoppers.

3. A production company's 350 hourly employees average 37.6 year 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:

1. Calculate μ and $\delta_{\bar{x}}$

$$\mu = 37.6 \text{ years}$$

$$n/N = 45/350 > 5\% \dots \text{FPCF is needed}$$

$$\delta_{\bar{x}} = \frac{\delta}{\sqrt{n}} * \sqrt{\frac{N-n}{N-1}} = \delta_{\bar{x}} = \frac{8.3}{\sqrt{45}} * \sqrt{\frac{350-45}{350-1}} = 1.16$$

2. Calculate Z for $\bar{X}$

$$Z_{\bar{X}} = \frac{\bar{X} - \mu}{\delta_{\bar{X}}} = \frac{\bar{X} - \mu}{\delta_{\bar{X}}}$$

$$Z_{40} = \frac{40 - 37.6}{1.16} = +2.07$$

3. Find the area covered by the interval

$$\begin{aligned} P(\bar{X} < 40) &= P(Z < +2.07) \\ &= 0.5 + P(0 \text{ to } +2.07) \\ &= 0.5 + 0.48077 \\ &= 0.98077 \end{aligned}$$

4. Interpret the results

There is a 98.08% chance of randomly selecting 45 hourly employees and their mean age be less than 40 years.

4. Suppose that a random sample size of 36 is being drawn from a population with a mean of 278. If 86% of the time the sample mean is less than 280, what is the population standard deviation?

Solution:

$$\mu = 278$$

$$n = 36$$

$$P(\bar{X} < 280) = 0.86 \quad \delta = ?$$

$$(Z/P) = +1.08$$

$$Z_{\bar{X}} = \frac{\bar{X} - \mu}{\delta_{\bar{X}}}$$

$$\delta_{\bar{X}} = \frac{\delta}{\sqrt{n}}$$

$$Z_{280} = \frac{280 - 278}{\delta_{\bar{X}}}$$

$$1.85 = \frac{\delta}{\sqrt{36}}$$

$$+1.08 = \frac{280 - 278}{\delta_{\bar{X}}}$$

$$1.85 = \frac{\delta}{6}$$

$$+1.08 = \frac{2}{\delta_{\bar{X}}}$$

$$\begin{aligned} \delta &= 6 * 1.85 \\ &= 11.1 \end{aligned}$$

$$\delta_{\bar{X}} = \frac{2}{1.08} = 1.85$$

5. A teacher gives a test to a class containing several hundred students. It is known that the standard deviation of the scores is about 12 points. A random sample of 36 scores is obtained.

- a) What is the probability that the sample mean will differ from the population mean by more than 6 points?
 b) What is the probability that the sample mean will be within 6 points of the population mean?

Solution:

a)

$$n = 36 \quad \delta = 12 \quad \delta_{\bar{x}} = \frac{\delta}{\sqrt{n}} = \frac{12}{\sqrt{36}} = \frac{12}{6} = 2$$

$$P(\bar{X} > \mu + 6) + P(\bar{X} < \mu - 6) = ?$$

$$Z_{\mu+6} = \frac{\mu + 6 - \mu}{2} = +3$$

$$Z_{\mu-6} = \frac{\mu - 6 - \mu}{2} = -3$$

$$\begin{aligned} P(\bar{X} > \mu + 6) + P(\bar{X} < \mu - 6) &= P(Z > 3) + P(Z < -3) \\ &= [0.5 - P(0 \text{ to } +3)] + [0.5 - P(0 \text{ to } -3)] \\ &= (0.5 - 0.49865) + (0.5 - 0.49865) \\ &= 0.00135(2) = 0.00270 \end{aligned}$$

b)

$$n = 36 \quad \delta = 12 \quad \delta_{\bar{x}} = \frac{\delta}{\sqrt{n}} = \frac{12}{\sqrt{36}} = \frac{12}{6} = 2$$

$$\begin{aligned} P(\mu - 6 \leq \bar{X} \leq \mu + 6) &= P(-3 \leq Z \leq 3) \\ &= P(0 \text{ to } 3) * 2 \\ &= 0.49865 * 2 \\ &= 0.9973 \end{aligned}$$

If the population standard deviation is 12, in a random sample of 36 scores there is a 99.73% chance of getting a sample mean score to lie within 6 points of the population mean.

B. Sampling Distribution of Proportions ($\bar{P}$)

Some times in statistics it is important to know the proportion of a certain characteristic in a population. That is, there are numerous problems in business where we want to know the proportion of items in a population that possess a certain characteristic. For example,

- A quality control engineer might want to know what proportions of products of an assembly line are defective.
- A labor economist might want to know what proportion of the labor force is unemployed.

Whereas the mean is computed by averaging a set of values, the sample proportion is computed by dividing the frequency that a given characteristic occurs in a sample by the number of items in the sample.

$$\bar{P} = \frac{X}{n}$$

Where $\bar{P}$ = sample proportions

X = number of items in a sample that possess the characteristic

n = number of items in the sample

Like other probability distribution, sampling distribution of the proportion is described by two parameters: the mean of the sample proportions, $E(\bar{P})$ and the standard deviation of the proportions, $\delta_{\bar{P}}$ which is called the standard error of the proportion.

Properties of Sampling distribution of $\bar{P}$

1. As the sampling distribution of the mean does, the population proportion, P , is always equal to the mean of the sample proportion, i.e., $P = E(\bar{P})$.
2. The standard error of the proportion is equal to: $\delta_{\bar{P}} = \sqrt{\frac{Pq}{n}}$, where P = population proportion
 $q = 1 - P$
 n = sample size.

Or

$$\delta_{\bar{P}} = \sqrt{\frac{Pq}{n}} * \sqrt{\frac{N-n}{N-1}}, \text{ where } \sqrt{\frac{N-n}{N-1}} = \text{finite population correction factor.}$$

The finite population correction factor is not needed if $n < 0.05N$.

Central Limit Theorem (CLT) and Sampling distribution of $\bar{P}$

How does a researcher use the sample proportion in analysis?

Answer: By applying the Central Limit Theorem. The CLT states that normal distribution approximates the shape of the distribution of sample proportions if **np and nq are greater than 5**. Consequently we solve problems involving sample proportions by using a normal distribution whose mean and standard deviation are:

$$\mu_{\bar{P}} = P, \quad \delta_{\bar{P}} = \sqrt{\frac{Pq}{n}} \quad \text{and} \quad Z_{\bar{P}} = \frac{\bar{P} - P}{\delta_{\bar{P}}}$$

NB: The sampling distribution of $\bar{p}$ can be approximated by a normal distribution whenever the sample size is large i.e., np and nq > 5.

Example:

1. Suppose that 60% of the electrical contractors in a region use a particular brand of wire. What is the probability of taking a random sample of size 120 from these electrical contractors and finding that 0.5 or less use that brand of wire?

Solution:

$$n = 120 \quad P = 0.6 \quad q = 0.4 \quad P(\bar{p} < 0.5) = ?$$

Steps:

1. Check that np and $nq > 5$
 $120 * 0.6 = 120$, and $120 * 0.4 = 48$. Both are greater than 5.

2. Calculate $\delta_{\bar{p}}$

$$\delta_{\bar{p}} = \sqrt{\frac{Pq}{n}} = \sqrt{\frac{0.6 * 0.4}{120}} = 0.0477$$

3. Calculate Z for $\bar{p}$

$$Z_{\bar{p}} = \frac{\bar{P} - P}{\delta_{\bar{p}}}$$

$$Z_{0.5} = \frac{0.5 - 0.6}{0.0477} = -2.24$$

4. Find the area covered by the interval

$$\begin{aligned} P(\bar{p} < 0.5) &= P(Z < -2.24) \\ &= 0.5 - P(0 \text{ to } -2.24) \\ &= 0.5 - 0.48745 \\ &= 0.01255 \end{aligned}$$

5. Interpret the results

The probability of finding 50% or less of the contractors to use this particular brand of wire is very low (1.255%) if we take a random sample of 120 contractors.

2. If 10% of a population of parts is defective, what is the probability of randomly selecting 80 parts and finding that 12 or more are defective?

Solution:

$$n = 80$$

$$P = 0.1$$

$$X = 12$$

$$\bar{p} = X/n = 12/80 = 0.15$$

$$P(\bar{p} > 0.15) = ?$$

$$Z_{0.15} = \frac{0.15 - 0.1}{0.0335} = +1.49$$

$$\begin{aligned} P(\bar{p} > 0.15) &= P(Z > +1.49) = 0.5 - P(0 \text{ to } +1.49) = 0.5 - P(0 \text{ to } +1.49) = \\ &0.5 - 0.43189 = 0.06811 \end{aligned}$$

About 6.81% of the time, twelve or more defective parts would appear in a random sample of eighty parts when the population proportion is 0.10.

3. Suppose that a population proportion is .40 and that 80% of the time you draw a random sample from this population, you get a sample proportion of 0.35 or more. How large a sample were you taking?

Solution:

$$P = 0.4$$

$$P(\bar{p} > 0.35) = 0.80$$

$$n = ?$$

$$(Z/P = 0.30) - 0.84$$

$$\delta_{\bar{p}} = \sqrt{\frac{Pq}{n}}; \text{ squaring both sides}$$

$$Z_{0.35} = \frac{0.35 - 0.4}{\delta_{\bar{p}}}$$

$$0.0595 = \left(\sqrt{\frac{.4 * .6}{n}} \right)$$

$$-0.84 = -0.05 / \delta_{\bar{p}}$$

$$(0.0595)^2 = \left(\sqrt{\frac{.4 * .6}{n}} \right)^2 \quad 0.0035 = 0.24/n$$

$$0.84 \delta_{\bar{p}} = 0.05$$

$$0.0035 = 0.24/n$$

$$\delta_{\bar{p}} = 0.05/0.84$$

$$n = 0.24/0.0035$$

$$= 0.0595$$

$$n = 68$$

4. If a population proportion is 0.28 and if the sample size is 140, 30% of the time the sample proportion will be less than what value if you are taking random samples?

Solution:

$$P = 0.28 \quad n = 140$$

$$P(\bar{p} < X) = 0.30 \quad X = ?$$

$$(Z/P = 0.2) = -0.52$$

$$Z_{\bar{p}} = \frac{\bar{P} - P}{\delta_{\bar{p}}}$$

$$\delta_{\bar{p}} = \sqrt{\frac{Pq}{n}}$$

$$-0.52 = \frac{\bar{P} - 0.28}{0.0379}$$

$$= \sqrt{\frac{0.28 * 0.72}{140}}$$

$$-0.0197 = \bar{P} - 0.28$$

$$= 0.0379$$

$$\bar{P} = 0.26$$

STATISTICAL ESTIMATION

Introduction

After developing sampling distributions of different population parameters, it is very important to estimate where these population parameters might be located. That means the population parameters must be estimated relative the samples in hand.

As its name suggests, the objective of estimation is to determine the approximate value of a population parameter on the bases of a sample statistic. An estimator of a population parameter is a random variable that is a function of the sample data. An estimate is the calculation of a specific value of this random variable.

The sampling distribution of the mean shows how far sample means could be from a known population mean. Similarly, the sampling distribution of the

proportion shows how far sample proportions could be from a known population proportion. In estimation, our aim is to determine how far an unknown population mean could be from the mean of a simple random sample selected from that population; or how far an unknown population proportion could be from a sample proportion. Those are the concerns of statistical inference, in which a statement about an unknown population parameter is derived from information contained in a random sample selected from the population.

Basic concepts:

- ▶ Estimation: is the process of using statistics as estimates of parameters. It is any procedure where sample information is used to estimate/ predict the numerical value of some population measure (called a parameter).
- ▶ Estimator- refers to any sample statistic that is used to estimate a population parameter. E.g. $\bar{x}$ for μ , $\bar{p}$ for p .
- ▶ Estimate- is a specific numerical value of our estimator. E.g. $\bar{x} = 9, 2, 5$

$$\begin{cases} \bar{x}, \bar{p}, s^2, s \dots\dots\dots \text{Estimators} \\ \mu, p, \sigma^2, \sigma \dots\dots\dots \text{items being estimated} \\ 1, 0.5, 9, 3 \dots\dots\dots \text{Estimates} \end{cases}$$

Four Important Properties of Estimators

A number of different estimators are possible for the same population parameter, but some estimators are better than others. To understand how, we need to look at four important properties of estimators: **unbiasedness, efficiency, consistency, and sufficiency**.

Unbiasedness: An estimator exhibits **unbiasedness** when the mean of the sampling estimator is equal to the population parameter: $E(\theta) = \Theta$.

In general, unbiasedness is a desirable property for an estimator. The sample mean is an unbiased estimator of the population mean. Similarly, the sample variance is an unbiased point estimator the population variance because the mean of the sampling distribution of the sample variance is equal to the population variance. And the sample proportion is an unbiased estimator of the population proportion. However, because standard deviation is a nonlinear function of variance, the sample standard deviation is not an unbiased estimator of population standard deviation. The bias of a point estimator is: $\text{Bias} = E(\theta) - \Theta$. If there are a number of unbiased estimators to choose from, there are **three** other criteria that could be used to select an estimator.

Efficiency: Efficiency is another standard that can be used to evaluate estimators. **Efficiency** refers to the size of the standard error of the statistics. The most efficient estimator is the one with the smallest variance. Thus, if there are two estimators for Θ with $\text{var}(\theta_1)$ and $\text{var}(\theta_2)$, then the first estimator θ_1 is said to be more efficient than the second estimator θ_2 , if $\text{var}(\theta_1) < \text{var}(\theta_2)$ although $E(\theta_1) = E(\theta_2) = \Theta$.

Consistency: A third property of estimators, **consistency**, is related to their behavior as the sample gets large. A statistic is a consistent estimator of a population parameter if, as the sample size increases, it becomes almost certain that the value of the statistic comes very close to the value of the population parameter.

An unbiased estimator is a consistent estimator if the variance approaches 0 as n increases. For example, the sample mean is an unbiased and a consistent estimator of population mean. Although the sample standard deviation is not an unbiased estimator of population standard deviation, it is a consistent estimator of population standard deviation.

Sufficiency: The last property of a good estimator is sufficiency. A **sufficient statistic** is an estimator that utilizes all the information a sample contains about the parameter to be estimated. For example, the sample mean is a sufficient estimator of the population mean. This means that no other estimator of the population mean from the same sample data, such as the sample median, can add any further information about the parameter (population mean) that is being estimated.

Types of Estimates:

We can make two types of estimates about a population: a point estimate and an interval estimate.

A point estimate: - is a single number that is used to estimate an unknown population parameter. It is a single value that is measured from a sample and used as an estimate of the corresponding population parameter.

The most important point estimates (given that they are single values) are:

- Sample mean ($\bar{x}$) for population mean (μ);
- Sample proportion ($\bar{p}$) for population proportion (p);
- Sample variance (s^2) for population variance (σ^2) and
- Sample standard deviation (s) for population standard deviation (σ)

An interval estimate - is a range of values used to estimate a population parameter. It describes the range of values within which a parameter might lie.

Stated differently, an interval estimate is a range of values within which the analyst can declare with some confidence that the population parameter will fall.

Example:

Suppose we have the sample 10,20,30,40 and 50 selected randomly from a population whose mean μ is unknown.

The sample mean, $\bar{x}, \frac{\sum xi}{n} = \frac{10 + 20 + 30 + 40 + 50}{5} = 30$ is a point estimate of μ .

On the other hand, if we state that the mean, μ , is between $\bar{x} \pm 10$, the range of values from 20 (30-10) to 40 (30+10) is an interval estimate.

Interval Estimation

Point estimators of population parameters, while useful, do not convey as much information as interval estimators. Point estimation produces a single value as an estimate of the unknown population parameter. The estimate may or may not be close to the parameter value; in other words, the estimate may be incorrect. An interval estimate, on the other hand, is a range of values that conveys the fact that estimation is an uncertain process. The standard error of the point estimator is used in creating a range of values; thus a measure of variability is incorporated into interval estimation. Further, a measure of confidence in the interval estimator is provided; consequently, interval estimates are also called confidence intervals. For these reasons, interval estimators are considered more desirable than point estimators.

Interval estimation for population mean, μ

As a result of the Central Limit Theorem (discussed in Chapter III) the following z formula for sample means can be used when sample sizes are large, regardless of the shape of the population distribution or for smaller sizes if the population is normally distributed.

$$Z = \frac{\bar{X} - \mu}{\sigma / n}$$

Rearranging the formula:

$$\mu = \bar{X} - Z \sigma / n$$

Because the sample mean can be greater than or less than the population mean, z can be positive or negative. Thus, the preceding expression takes the form:

$$\mu = \bar{X} \pm Z \sigma / n$$

The value of the population mean, μ , lies somewhere within this range. Rewriting this expression yields the confidence interval for population mean:

$$\bar{X} - Z \sigma / n \leq \mu \leq \bar{X} + Z \sigma / n$$

The confidence interval for population mean is affected by:

1. The population distribution, i.e., whether the population is normally distributed or not
2. The standard deviation, i.e., whether σ is known or not.
3. The sample size, i.e., whether the sample size, n , is large or not.

Confidence interval estimate of μ - Normal population, σ known

A confidence interval estimate for μ is an interval estimate together with a statement of how confident we are that the interval estimate is correct.

When the population distribution is normal and at the same time σ is known, we can estimate μ (regardless of the sample size) using the following formula¹.

$$\mu = \bar{X} \pm Z_{\alpha/2} \sigma / n$$

Where: $\bar{X}$ = sample mean; Z = value from the standard normal table reflecting confidence level; σ = population standard deviation; n = sample size; α = the proportion of incorrect statements ($\alpha = 1 - C$); and μ = unknown population mean

From the above formula we can learn that an interval estimate is constructed by adding and subtracting the error term to and from the point estimate. That is, the point estimate is found at the center of the confidence interval.

To find the interval estimate of population mean, μ we have the following steps.

1. Compute the standard error of the mean ($\sigma_{\bar{x}}$)
2. Compute $\alpha/2$ from the confidence coefficient.
3. Find the Z value for the $\alpha/2$ from the table
4. Construct the confidence interval
5. Interpret the results

Examples:

1. The vice president of operations for Ethiopian Tele Communication Corporation (ETC) is in the process of developing a strategic management plan. He believes that the ability to estimate the length of the average phone call on the system is important. He takes a random sample of 60 calls from the company records and finds that the mean sample length for a call is 4.26 minutes. Past history for these types of calls has shown that the population standard deviation for call length is

¹ This formula works also for problems which involve large sample size ($n > 30$) even though the population is not normally distributed. And if $n > .05N$, finite population correction factor may be used.

about 1.1 minutes. Assuming that the population is normally distributed and he wants to have a 95% confidence, help him in estimating the population mean.

Solution:

$n = 60$ calls $\bar{X} = 4.26$ minutes $\sigma = 1.1$ minutes $C = 0.95$

$$\begin{aligned} \text{i. } \sigma_{\bar{X}} &= \frac{\sigma}{\sqrt{n}} = \frac{1.1}{\sqrt{60}} = 0.142 & \text{iv. } \mu &= \bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \\ \text{ii. } \alpha &= 1 - C = 1 - 0.95 = 0.05 & &= 4.26 \pm 1.96(0.142) \\ \alpha/2 &= 0.05/2 = 0.025 & &= \underline{4.26 \pm 0.28} \\ \text{iii. } Z_{\alpha/2} &= Z_{0.025} = 1.96 \end{aligned}$$

$$\Rightarrow 3.98 \leq \mu \leq 4.54$$

The vice-president of ETC can be 95% confident that the average length of a call for the population is between 3.98 and 4.54 minutes.

2. A survey conducted by "Addis Zemen Gazette" found that the sample mean age of men was 44 years and the sample mean age of women was 47 years. All together, 454 people from Addis were included in the reader poll –340 women and 114 men. Assume that the population standard deviation of age for both men and women is 8 years.
 - a. Develop a 95% confidence interval estimate for the mean age of the population men who read the gazetta.
 - b. Develop a 95% confidence interval estimate for the mean age of the population women who read the gazetta.
 - c. Compare the widths of the two interval estimates form part (a) & (b) which one has a better precision? Why?

Solution:

a.

$$\begin{aligned} n &= 114 \text{ men} & \bar{X} &= 44 \text{ years} & \sigma &= 8 \text{ years} & C &= 0.95 \\ \text{i. } \sigma_{\bar{X}} &= \frac{\sigma}{\sqrt{n}} = \frac{8}{\sqrt{114}} = 0.75 & \text{iv. } \mu &= \bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \\ \text{ii. } \alpha &= 1 - C = 1 - 0.95 = 0.05 & &= 44 \pm 1.96(0.75) \\ \alpha/2 &= 0.05/2 = 0.025 & &= \underline{4.26 \pm 1.47} \\ \text{iii. } Z_{\alpha/2} &= Z_{0.025} = 1.96 \end{aligned}$$

$$\Rightarrow 42.53 \leq \mu \leq 45.47$$

b.

$$\begin{aligned} n &= 340 \text{ women} & \bar{X} &= 47 \text{ years} & \sigma &= 8 \text{ years} & C &= 0.95 \\ \text{i. } \sigma_{\bar{X}} &= \frac{\sigma}{\sqrt{n}} = \frac{8}{\sqrt{340}} = 0.434 & \text{iv. } \mu &= \bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \\ \text{ii. } \alpha &= 1 - C = 1 - 0.95 = 0.05 & &= 47 \pm 1.96(0.434) \end{aligned}$$

$$\alpha/2 = 0.05/2 = 0.025 \qquad \qquad \qquad = \underline{47 \pm 0.85}$$

$$\text{iii. } Z_{\alpha/2} = Z_{0.025} = 1.96$$

$$\blacktriangleright 46.15 \leq \mu \leq 47.85$$

- c. Part b has a better precision because the sample size is larger as compared with part a.
3. Time magazine reports information on the time required for caffeine from products such as coffee and soft drinks to leave the body after consumption. Assume that the 99% confidence interval estimate of the population mean time for adults is 5.6 hrs to 6.4 hrs.
- What is the point estimate of the mean time for caffeine to leave the body after consumption?
 - If the population standard deviation is 2 hrs, how large a sample was used to provide the interval estimate?

Solution:

$$C = 0.99 \qquad \qquad \qquad \text{Confidence interval: } 5.6 \leq \mu \leq 6.4$$

$$\text{a. point estimate} = \frac{5.6 + 6.4}{2} = 6 \text{ hours}$$

Or;

$$+ \begin{cases} 5.6 = \bar{X} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \\ 6.4 = \bar{X} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \end{cases}$$

$$12 = 2 \bar{X}$$

$$\underline{\underline{\bar{X} = 6 \text{ hours}}}$$

$$\text{b. } 0.99 \qquad \sigma = 2 \text{ hours} \qquad \text{Confidence interval: } 5.6 \leq \mu \leq 6.4 \qquad n=?$$

$$\alpha = 1 - C = 1 - 0.99 = 0.01 \qquad \alpha/2 = 0.005 \qquad Z_{\alpha/2} = Z_{0.005} = 2.57$$

$$6.4 = \bar{X} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$6.4 = 6 + 2.57 \frac{2}{\sqrt{n}}$$

$$0.4 = 5.14 \frac{1}{\sqrt{n}} ; \text{ rearranging the expression}$$

$$\sqrt{n} = \frac{5.14}{0.4}$$

$$\sqrt{n} = 12.85 ; \text{ squaring both sides}$$

$$\underline{\underline{n = 165}}$$

We state with 99% confidence that the mean time required for caffeine to leave the body after consumption lies between 5.6 and 6.4 hrs.

Confidence interval estimate of μ - Normal population, σ unknown, n large

If we know that the population is normal, and we know the population standard deviation, σ the confidence interval for μ should be constructed in the manner already shown, i.e., $\mu = \bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$. If the population standard deviation is unknown, it has to be estimated from the sample; i.e., when σ is

unknown, we use sample standard deviation: $S = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}$. Then, the standard error of the mean, $\sigma_{\bar{X}}$, is estimated by the sample standard error of the mean: $S_{\bar{X}} = \frac{S}{\sqrt{n}}$.

Therefore, the confidence interval to estimate μ when population standard deviation is unknown, population normal and n is large is²

$$\mu = \bar{X} \pm Z_{\alpha/2} \frac{S}{\sqrt{n}}.$$

Examples:

1. Suppose that a car rental firm in Addis wants to estimate the average number of miles traveled by each of its cars rented. A random sample of 110 cars rented reveals that the sample means travel distance per day is 85.5 miles, with a sample standard deviation of 19.3 miles. Compute a 99% confidence interval to estimate μ .

Solution:

n= 110 rented cars $\bar{X} = 85.5$ miles s = 19.3 miles C= 0.99

$$\begin{aligned} \text{i. } S_{\bar{X}} &= \frac{S}{\sqrt{n}} = \frac{19.3}{\sqrt{110}} = 1.84 & \text{iv. } \mu &= \bar{X} \pm Z_{\alpha/2} \frac{s}{\sqrt{n}} \\ \text{ii. } \alpha &= 1 - C = 1 - 0.99 = 0.01 & &= 85.5 \pm 2.57(1.84) \\ \alpha/2 &= 0.01/2 = 0.005 & &= \underline{85.5 \pm 4.73} \\ \text{iii. } Z_{\alpha/2} &= Z_{0.005} = 2.57 \end{aligned}$$

➡ $80.77 \leq \mu \leq 90.23$

² This formula also works for large sample size even though the parent population is not normally distributed.

We state with 99% confidence that the average distance traveled by rented cars lies between 80.77 and 90.23 miles.

2. A study is being conducted in a company that has 800 engineers. A random sample of 50 of these engineers reveals that the average sample age is 34.3 years, and the sample standard deviation is 8 years. Assuming normality, construct a 98% confidence interval to estimate the average age of all engineers in this company.

Solution:

$n = 50$ engineers $N = 800$ engineers $\bar{X} = 34.3$ years $s = 8$ years $C = 0.98$

$$\text{i. } S_{\bar{x}} = \frac{S}{\sqrt{n}} * \sqrt{\frac{N-n}{N-1}} = \frac{8}{\sqrt{50}} * \sqrt{\frac{800-50}{800-1}} = 1.10$$

$$\text{ii. } \alpha = 1 - C = 1 - 0.98 = 0.02$$

$$\alpha/2 = 0.02/2 = 0.01$$

$$\text{iii. } Z_{\alpha/2} = Z_{0.01} = 2.33$$

$$\text{iv. } \mu = \bar{X} \pm Z_{\alpha/2} \frac{s}{\sqrt{n}} * \sqrt{\frac{N-n}{N-1}}$$

$$= 34.3 \pm 2.33(1.10)$$

$$= \underline{\underline{34.3 \pm 2.56}}$$

$$\Rightarrow 31.74 \leq \mu \leq 36.86$$

We state with 98% confidence that the mean age of engineers lies between 31.74 and 36.86 years.

Confidence interval for $\mu - \sigma$ unknown, n-small, population normal

If the sample size is small ($n < 30$), we can develop an interval estimate of a population mean only if the population has a normal probability distribution. If the sample standard deviation s is used as an estimator of the population standard deviation σ and if the population has a normal distribution, interval estimation of the population mean can be based up on a probability distribution known as t-distribution.

Characteristics of t-distribution

1. The t-distribution is symmetric about its mean (0) and ranges from $-\infty$ to ∞ .
2. The t-distribution is bell-shaped (unimodal) and has approximately the same appearance as the standard normal distribution (Z- distribution).
3. The t-distribution depends on a parameter v (Greek Nu)⁴, called the degrees of freedom of the distribution. $N = n - 1$, where n is sample size. The degree of freedom, v , refers to the number of values we can choose freely.

³ Since the sample size is greater than 5% of the population size, finite population multiplier is used to calculate the sample standard error of the mean.

4. The variance of the t-distribution is $v/(v-2)$ for $v > 2$.
5. The variance of the t-distribution always exceeds 1.
6. As v increases, the variance of the t-distribution approaches 1 and the shape approaches that of the standard normal distribution.
7. Because the variance of the t-distribution exceeds 1.0 while the variance of the Z-distribution equals 1, the t-distribution is slightly flatter in the middle than the Z-distribution and has thicker tails.
8. The t-distribution is a family of distributions with a different density function corresponding to each different value of the parameter v . That is, there is a separate t-distribution for each sample size. In proper statistical language, we would say, "There is a different t-distribution for each of the possible degrees of freedom".
9. The t formula for sample when σ is unknown, the sample size is small, and the population is normally distributed is: $t = \frac{\bar{X} - \mu}{S_{\bar{X}}} = \frac{\bar{X} - \mu}{s/\sqrt{n}}$ This formula is essentially the same as the z-formula, but the distribution table values are not.

The confidence interval to estimate μ becomes:

$$\mu = \bar{X} \pm t_{\alpha/2, v} \frac{s}{\sqrt{n}}$$

Where: $\bar{X}$ = sample mean

$\alpha = 1 - C$

$v = n - 1$ (degrees of freedom)

s = sample standard deviation

n = sample size

μ = unknown population mean

Steps:

1. Calculate degrees of freedom ($v=n-1$) and sample standard error of the mean.
2. Compute $\alpha/2$

⁴ What are degrees of freedom? We can define them as the number of values we can choose freely. In general, the degrees of freedom for a t statistic are the degrees of freedom associated with the sum of squares used to obtain an estimate of the variance. The variance estimate depends on not only on the sample size but also on how many parameters must be estimated with the sample:

$$\text{Degrees of freedom} = \text{Number of Observations} - \text{Number of parameters that must be estimated beforehand}$$

Here we calculate sample variance by using n observations and estimating one parameter (the mean). Thus, there are $(n - 1)$ degrees of freedom.

3. Look up $t_{\alpha/2, v}$
4. Construct the confidence interval
5. Interpret results

Examples:

1. If a random sample of 27 items produces $\bar{x} = 128.4$ and $s = 20.6$. What is the 98% confidence interval for μ ? Assume that x is normally distributed for the population. What is the point estimate?

Solution:

The point estimate of the population mean is the sample mean, in this case 128.4 is the point estimate.

$$\begin{aligned}
 n &= 27 & \bar{X} &= 128.4 & s &= 20.6 & C &= 0.98 \\
 \text{i. } S_{\bar{X}} &= \frac{S}{\sqrt{n}} = \frac{20.6}{\sqrt{27}} = 3.96 & v &= n - 1 = 27 - 1 = 26 \\
 \text{ii. } \alpha &= 1 - C = 1 - 0.98 = 0.02 \\
 \alpha/2 &= 0.02/2 = 0.01 \\
 \text{iii. } t_{\alpha/2, v} &= t_{0.01, 26} = 2.479 \\
 \text{iv. } \mu &= \bar{X} \pm t_{\alpha/2, v} \frac{s}{\sqrt{n}} \\
 &= 128.4 \pm 2.479(3.96) \\
 &= \underline{128.4 \pm 9.82} \\
 &\quad \rightarrow 118.56 \leq \mu \leq 138.22
 \end{aligned}$$

We state with 98% confidence that the population mean lies between 118.56 and 138.23.

2. A sample of 20 cab fares in Bahir Dar city shows a sample mean of Br 2.50 and a sample standard deviation of Br. 0.50. Develop a 90% confidence interval estimate of the mean cab fares in Bahir Dar city. Assume the population of cab fares has a normal distribution.

$$\begin{aligned}
 n &= 20 & \bar{X} &= \text{Birr } 2.50 & s &= \text{Birr } 0.50 & C &= 0.90 \\
 \text{i. } S_{\bar{X}} &= \frac{S}{\sqrt{n}} = \frac{0.5}{\sqrt{20}} = 0.112 & v &= n - 1 = 20 - 1 = 19 \\
 \text{ii. } \alpha &= 1 - C = 1 - 0.90 = 0.10 \\
 \alpha/2 &= 0.10/2 = 0.05 \\
 \text{iii. } t_{\alpha/2, v} &= t_{0.05, 19} = 1.729 \\
 \text{iv. } \mu &= \bar{X} \pm t_{\alpha/2, v} \frac{s}{\sqrt{n}} \\
 &= 2.50 \pm 1.729(0.112) \\
 &= \underline{2.50 \pm 0.194}
 \end{aligned}$$

$$\Rightarrow 2.31 \leq \mu \leq 2.69$$

We state with 90% confidence that the mean of cab fares in Bahir Dar city lies between Birr 2.31 and 2.69.

3. Sales personnel for X Company are required to submit weekly reports listing customer contacts made during the week. A sample of 61 weekly contact reports showed a mean of 22.4 customer contacts per week for the sales personnel. The sample standard deviation was 5 contacts.
 - a. Develop a 95% confidence interval estimate for the mean number of weekly customer contacts for the population of sales personnel.
 - b. Assume that the population of weekly contact data has a normal distribution. Use the t distribution to develop a 95% confidence interval for the mean number of weekly customer contacts.
 - c. Compare your answer for parts (a) and (b). What do you conclude from your results?

Solution:

a. $n = 61$ weekly contact reports⁵ $\bar{X} = 22.4$ contacts $s = 5$ contacts $C = 0.95$

$$\begin{aligned} \text{i. } S_{\bar{X}} &= \frac{S}{\sqrt{n}} = \frac{5}{\sqrt{61}} = 0.64 \\ \text{ii. } \alpha &= 1 - C = 1 - 0.95 = 0.05 \\ \alpha/2 &= 0.05/2 = 0.025 \\ \text{iii. } Z_{\alpha/2} &= Z_{0.025} = 1.96 \\ \text{iv. } \mu &= \bar{X} \pm Z_{\alpha/2} \frac{s}{\sqrt{n}} \\ &= 22.4 \pm 1.96(0.64) \\ &= \underline{22.4 \pm 1.25} \end{aligned}$$

$$\Rightarrow 21.15 \leq \mu \leq 23.65$$

I state with 95% confidence that the mean weekly contact lies between 21.15 and 23.65 contacts.

b. $n = 61$ weekly contact reports $\bar{X} = 22.4$ contacts $s = 5$ contacts $C = 0.95$

$$\begin{aligned} \text{i. } S_{\bar{X}} &= \frac{S}{\sqrt{n}} = \frac{5}{\sqrt{61}} = 0.64 & v = n - 1 = 61 - 1 = 60 \\ \text{ii. } \alpha &= 1 - C = 1 - 0.95 = 0.05 \\ \alpha/2 &= 0.05/2 = 0.025 \\ \text{iii. } t_{\alpha/2, v} &= t_{0.025, 60} = 2.00 \\ \text{iv. } \mu &= \bar{X} \pm t_{\alpha/2, v} \frac{s}{\sqrt{n}} \\ &= 22.4 \pm 2.00 (0.64) \end{aligned}$$

⁵ Since the sample size is large, we use the Z-distribution to construct the confidence interval.

$$= 22.4 \pm 1.28$$

$$\Rightarrow 21.12 \leq \mu \leq 23.68$$

I state with 95% confidence that the mean weekly contact lies between 21.12 and 23.68 contacts.

c. As the sample size increases, the t-distribution and z (normal) distribution approximate to be equal.

Confidence interval for μ - n small, σ unknown, population not normal

This is solved by non-parametric tests, which do not require assumption about the underlying form of the population data.

Interval Estimation of the Population Proportion

We know that a sample proportion, $\bar{p}$, is an unbiased estimator of a population proportion P and if the sample size is large then, the sampling distribution of $\bar{p}$ is

$$\text{normal with } Z = \frac{\bar{P} - P}{\sigma_{\bar{p}}} = \frac{\bar{P} - P}{\sqrt{Pq/n}}.$$

However, here p is unknown and we want to estimate p by $\bar{p}$ and hence z

$$\text{becomes } Z = \frac{\bar{P} - P}{\sqrt{\bar{p}q/n}}. \text{ That is, } \sigma_{\bar{p}} \text{ is substituted by } S_{\bar{p}} = \sqrt{\bar{p}q/n}$$

Solving for P results in $P = \bar{p} + Z\sqrt{\bar{p}q/n}$ and since Z can assume both positive and negative values, it becomes $P = \bar{p} \pm Z\sqrt{\bar{p}q/n}$.

$$\begin{aligned} \text{Since Z represents the confidence level we write it as } P &= \bar{p} \pm Z_{\alpha/2} \sqrt{\bar{p}q/n} \\ &= \bar{p} \pm Z_{\alpha/2} S_{\bar{p}} \end{aligned}$$

Where: $\bar{p}$ = sample proportion

$$\bar{q} = 1 - \bar{p}$$

$$\alpha = 1 - C$$

n = sample size

P = unknown population proportion

Examples:

1. Recently, a study of 87 randomly selected companies with telemarketing operation was completed. The study revealed that 39% of the sampled companies had used telemarketing to assist them in order processing. Using this

information estimate the population proportion of telemarketing companies who use their telemarketing operation to assist them in order processing taking a 95% confidence level.

Solution:

$$n = 87 \quad \bar{p} = 0.39 \quad \bar{q} = 0.61 \quad C = 0.95$$

$$\text{i. } S_{\bar{p}} = \sqrt{\frac{\bar{p}\bar{q}}{n}} = \sqrt{\frac{0.61 * 0.39}{87}} = 0.0523$$

$$\text{ii. } \alpha = 1 - C = 1 - 0.95 = 0.05$$

$$\alpha/2 = 0.05/2 = 0.025$$

$$\text{iii. } Z_{\alpha/2} = Z_{0.025} = 1.96$$

$$\text{iv. } P = \bar{p} \pm Z_{\alpha/2} S_{\bar{p}}$$

$$= 0.39 \pm 1.96(0.0523)$$

$$= \underline{0.39 \pm 0.1025}$$

$$\Rightarrow 0.2875 \leq P \leq 0.4925$$

We state with 95% confidence that the proportion of companies which use telemarketing to assist order processing lies between 0.2875 and

2. A fast food restaurant took a random sample of 400 customers to determine the proportion of customers who are female. A confidence interval of .73 to .87 was reported.

- Find the number of females and the sample proportion
- Find the level of confidence of this interval

Solution:

$$\text{a. } n = 400 \quad 0.73 \leq P \leq 0.87 \quad \bar{p} = ? \quad \text{Number of females} = ?$$

$$\text{Point estimate} = \frac{0.73 + 0.87}{2} = 0.80$$

Or;

$$+ \begin{cases} 0.73 = \bar{p} - Z_{\alpha/2} S_{\bar{p}} \\ 0.87 = \bar{p} + Z_{\alpha/2} S_{\bar{p}} \end{cases}$$

$$1.60 = 2 \bar{p}$$

$$\underline{\underline{\bar{p} = 0.8}}$$

$$\text{Number of females (X)} = n * \bar{p} = 400 * 0.8 = 320$$

$$\text{b. } P = \bar{p} \pm Z_{\alpha/2} S_{\bar{p}}$$

$$0.87 = 0.8 + Z_{\alpha/2} S_{\bar{p}}$$

$$0.07 = Z_{\alpha/2} \sqrt{0.8 * 0.2 / 400}$$

$$0.07 = Z_{\alpha/2} * 0.02$$

$$3.50 = Z_{\alpha/2}$$

$$(P/Z=3.5) = 0.49977$$

$$C = 0.49977 * 2$$

$$= \underline{99.954\%}$$

3. A random sample of 400 faculty members at AAU contained 120 people who believed that the University should improve its library service. On the basis of this sample information, an analyst calculated the confidence interval (.25, .35) for the population proportion of faculty members favoring improvement. What is the level of confidence of this interval?

Solution:

$$n = 400 \quad X = 120 \quad \bar{p} = 0.30 \quad \text{Interval estimate } 0.25 \leq P \leq 0.30 \quad C = ?$$

$$P = \bar{p} \pm Z_{\alpha/2} S_{\bar{p}}$$

$$0.25 = 0.30 - Z_{\alpha/2} S_{\bar{p}}$$

$$0.05 = Z_{\alpha/2} \sqrt{0.7 * 0.3 / 400}$$

$$0.05 = Z_{\alpha/2} * 0.023$$

$$2.17 = Z_{\alpha/2}$$

$$(P/Z=2.17) = 0.485$$

$$C = 0.485 * 2$$

$$= \underline{97\%}$$