N8104: Artificial Neural Networks

Population & Sample

Sachchida Nand Chaurasia

Assistant Professor

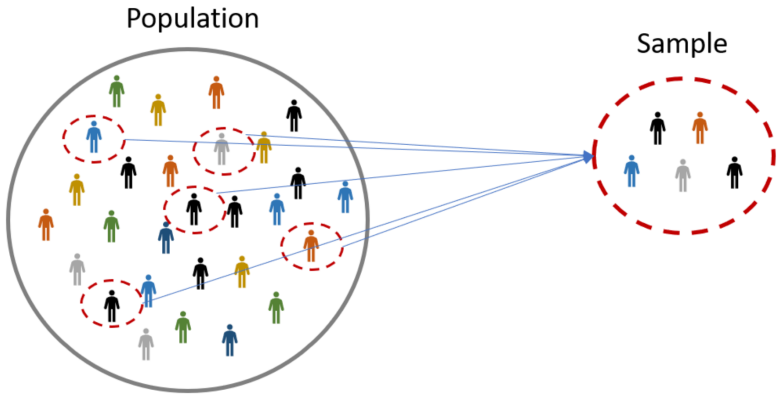
Department of Computer Science

Banaras Hindu University

Varanasi

Email id: snchaurasia@bhu.ac.in, sachchidanand.mca07@gmail.com

Population and Sample I



Population and Sample II

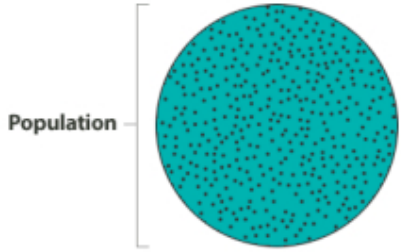
We begin a statistical investigation with a research question. The research question is frequently something we want to know about a population.

✓ The population can be people or other things, such as animals or objects.

✓ For example, we might want to know the answer to questions such as:

- What percentage of U.S. adults supports the death penalty? (Population: U.S. adults)
- Do cell phones affect bees? (Population: bees)
- Do cars get better gas mileage with a new gasoline additive? (Population: cars)

Population and Sample III



✓ In most cases, the population is a large group. Often, the population is so large that we cannot collect information from every individual in the population. So we select a sample from the population.

✓ Then we collect data from this sample.

✓ This is the first step in the statistical investigation. *We call this step producing data*.

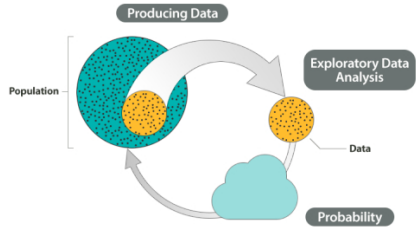
Population and Sample IV

✓ Data is often a long list of information.

✓ To make sense of the data, we explore it and summarize it using graphs and different numerical measures, such as percentages or averages.

✓ *This is called exploratory data analysis.*

Population and Sample V



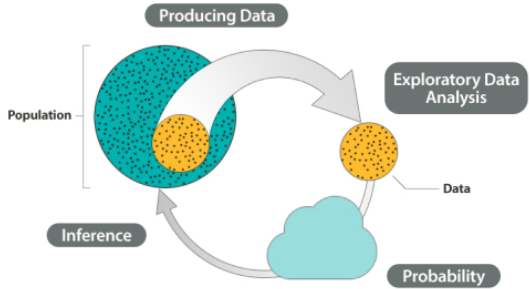
✓ Remember, our goal is to answer a question about a population based on a sample.

✓ Of course, samples will vary due to chance, and we will need to answer our question in spite of this variability.

✓ So we need to understand how sample results will vary and how sample results relate to the population as a whole when chance is involved.

✓ *This is where probability comes in*.

Population and Sample VI

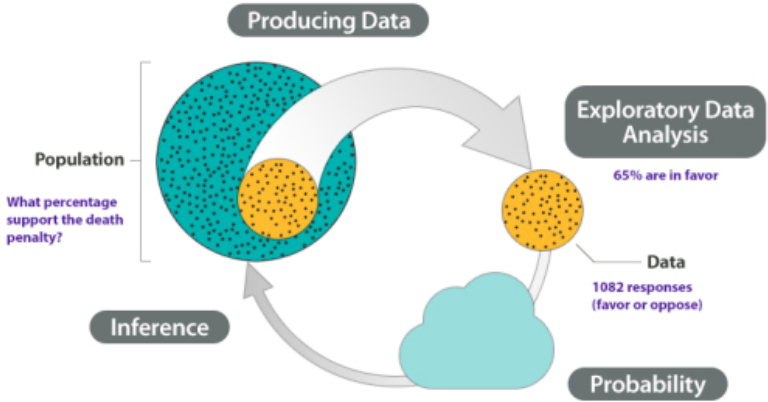


✓ Probability is the "machinery" behind the last step in the process called **inference**.

✓ We infer something about a population based on a sample.

✓ This inference is the conclusion we reach from our sample data that answers our original question about the population.

Population and Sample VII



Population and Sample VIII

1. Produce Data: Determine what to measure, then collect the data.
   ✓ The poll selected 1,082 U.S. adults at random. Each adult answered this question: "Do you favor or oppose the death penalty for a person convicted of murder?"
2. Explore the Data: Analyze and summarize the data.
   ✓ In the sample, 65% favored the death penalty.
3. Draw a Conclusion: Use the data, probability, and statistical inference to draw a conclusion about the population.

Population and Sample IX

✓ Our goal is to determine the percentage of the U.S. adult population that supports the death penalty.

✓ We know that different samples will give different results.

✓ What are the chances that our sample reflects the opinions of the population within 3%?

✓ Probability describes the likelihood that our sample is this accurate.

✓ So we can say with 95% confidence that between 62% and 68% of the population favor the death penalty.

Population and Sample X

Type of Research Question	Examples
Make an estimate about the population (often an estimate about an average value or a proportion with a given characteristic)	What is the average number of hours that community college students work each week?
	What proportion of all U.S. college students are enrolled at a community college?
Test a claim about the population (often a claim about an average value or a proportion with a given characteristic)	Is the average course load for a community college student greater than 12 units? Do the majority of community college students qualify for federal student loans?

Population and Sample XI

Type of Research Question	Examples
Compare two populations (often a comparison of population averages or proportions with a given characteristic)	In community colleges, do female students have a higher GPA than male students?
	Are college athletes more likely than non athletes to receive academic advising?
Investigate a relationship between two variables in the population	Is there a relationship between the number of hours high school students spend each week on Facebook and their GPA?
	Is academic counseling associated with quicker completion of a college degree?

Population and Sample XII

✓ An observational study can answer questions about a population. But populations are generally large groups, so we cannot gather data from every individual in the population.

✓ Instead, we select a sample and gather data from the sample. We use the data from the sample to make statements about the population.

Here are two examples:

- A political scientist wants to know what percentage of college students consider themselves conservatives. The population is college students. It would be too time consuming and expensive to poll every college student, so the political scientist selects a sample of college students. Of course, the sample must be carefully selected to represent the political perspectives that are present in the population.

Population and Sample XIII

- A government agency plans to test airbags from Honda to determine if the airbags work properly. Testing an airbag means it has to be inflated and punctured, which ruins the airbag, so the researchers certainly cannot test every airbag.
  ✓ Instead, they test a sample of airbags and draw a conclusion about the quality of airbags from Honda.

Important Point

✓ A goal is to use a sample to make valid conclusions about a population. Therefore, the sample must be representative of the population.

✓ A representative sample is a subset of the population that reflects the characteristics of the population.

✓ A sampling plan describes exactly how we will choose the sample. A sampling plan is biased if it systematically favors certain outcomes.

How biased sampling in a survey leads to misleading conclusions about the population:

Example

In 1936, Democrat Franklin Roosevelt and Republican Alf Landon were running for president. Before the election, the magazine Literary Digest sent a survey to 10 million Americans to determine how they would vote. More than 2 million people responded to the poll; 60% supported Landon.

The magazine published the findings and predicted that Landon would win the election. However, Roosevelt defeated Landon in one of the largest landslide presidential elections ever.

What happened?

The magazine used a biased sampling plan. They selected the sample using magazine subscriptions, lists of registered car owners, and telephone directories. The sample was not representative of the American public. In the 1930s, Democrats were much less likely to own a car or have a telephone. The sample therefore systematically underrepresented Democrats. The poll results did not represent the way people in the general population voted.

Common survey plans that produce unreliable and potentially biased results

How to Sample Badly

Online polls: *The American Family Association (AFA) is a conservative Christian group that opposes same-sex marriage. In 2004, the AFA began a campaign in support of a constitutional amendment to define marriage as strictly between a man and a woman. The group posted a poll on its website asking AFA members to voice their opinion about same-sex marriage. The AFA planned to forward the results to Congress as evidence of America's opposition to same-sex marriage. Almost 850,000 people responded to the poll. In the poll, 60% favored legalizing same-sex marriage.*

What happened?

Population and Sample XVII

Against the wishes of the AFA, the link to the poll appeared in blogs, social-networking sites, and a variety of email lists connected to gay/lesbian/bisexual groups. The AFA claimed that gay rights groups had skewed its poll. Of course, the results of the poll would have been skewed in the other direction had only AFA members been allowed to participate.

This is an example of a voluntary response sample. **The people in a voluntary response sample are self-selected, not chosen**. *For this reason, a voluntary response sample is biased because only people with strong opinions make the effort to participate.*

Mall surveys: *Have you ever noticed someone surveying people at a mall? People shopping at a mall are more likely to be teenagers, retired people, or people who have more money than the typical Indian. In addition, unless interviewers are carefully trained, they tend to interview people with whom they are comfortable talking. For these reasons, mall surveys*

Population and Sample XVIII

frequently over-represent the opinions of rich middle-class or retired people. Mall surveys are an example of a convenience sample.

How to Eliminate Bias in Sampling

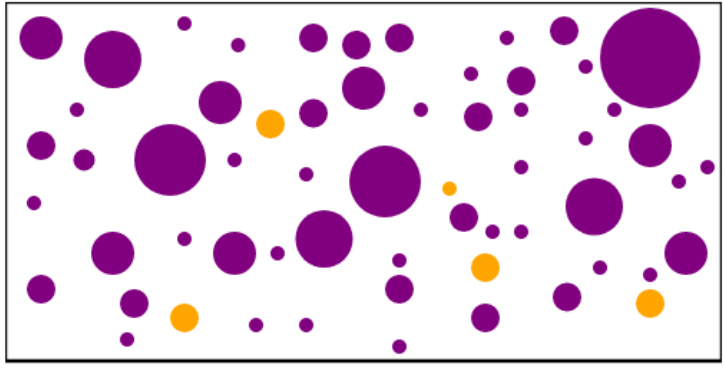
✓ In a voluntary response sample, people choose whether to respond. In a convenience sample, the interviewer chooses who will be part of the sample. In both cases, personal choice produces a biased sample. Random sampling is the best way to eliminate bias.

✓ Collecting a random sample is like pulling names from a hat (assuming every individual in the population has a name in the hat!).

✓ In a simple random sample everyone in the population has an equal chance of being chosen.

✓ Reputable polling firms use techniques that are more complicated than pulling names out of a hat. But the goal is the same: eliminate bias by using random chance to decide who is in the sample.

Random Samples



Population and Sample XXI

✓ Random selection also guarantees that the sample results do not change unsystematic from sample to sample.

✓ When we use random selection, the variability we see in sample results is due to chance.

✓ The results obey the mathematical laws of probability.

✓ Probability is the machinery for drawing conclusions about a population on the basis of samples. To use this machinery, the sample must be chosen by random chance.

The main points about sampling:

- We draw a conclusion about the population on the basis of the sample.
- To draw a valid conclusion, the sample must be representative of the population.
- A representative sample is a subset of the population that reflects the characteristics of the population.

Population and Sample XXII

- A sample is biased if it systematically favors a certain outcome.
- Random selection eliminates bias.

We have not mentioned the size of the sample

Are larger samples more accurate? Well, the answer is *Yes* and *No*

For *No:* Recall the 1936 presidential election. A sample of over 2 million people did not correctly identify the winner of the election. Two million people is a huge sample, yet the results were completely wrong. So a large sample does not guarantee reliable results.

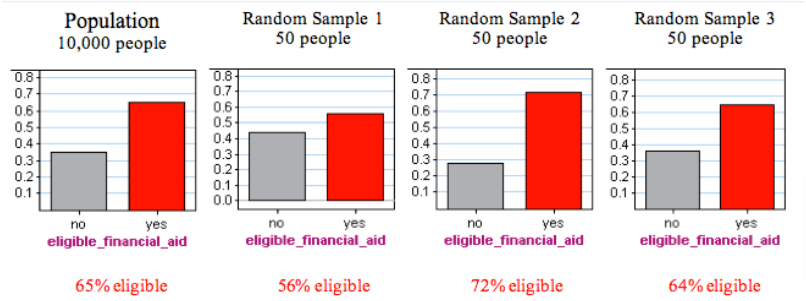
For Random Samples, Size Matters:

Let's compare the accuracy of random samples of different sizes.

✓ Suppose there are 10,000 students at your college. Also suppose that 65% of these students are eligible for financial aid. How accurate are random samples at predicting this population value?

Population and Sample XXIV

✓ To answer this question, we randomly select 50 students and determine the proportion who are eligible for financial aid. We repeat this several times. Here are the results for three random samples:



Population and Sample XXV

✓ Notice that each random sample has a different result. Some results are larger than the true population value of 65%; some results are smaller than the true population value.

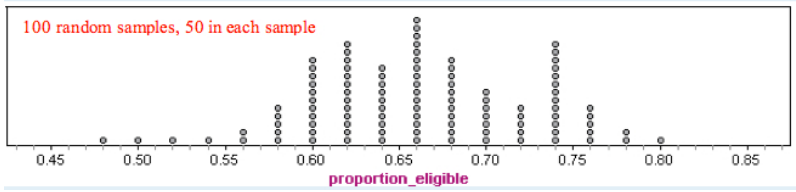
✓ Because there is no bias in random samples, we expect results above and below the true value to occur with similar frequency.

✓ A simulation to take many more random samples.

✓ Again, each sample is composed of 50 randomly selected people.

✓ Here is a dotplot of the proportion who are eligible for financial aid in 100 samples. Each dot is a random sample.

Population and Sample XXVI

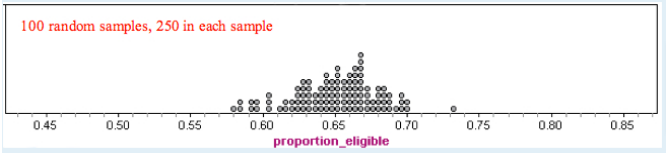


✓ We see that the results from random samples vary from 0.48 to 0.80. Typical values range from about 0.58 to 0.74.

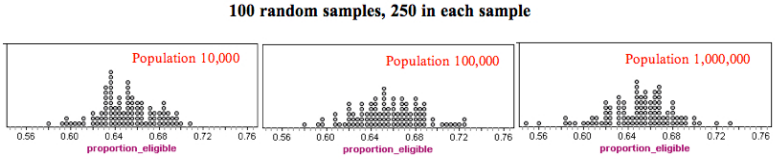
Many samples have results below the true population value of 0.65, and many have results above 0.65. This shows that random samples are not biased. For the question Are you eligible for financial aid?, there is no systematic favoring of one response over another. The samples are representative of the population.

What happens when we increase the number of people in the random sample?

Population and Sample XXVII



✓ The precision of the sample results depends on the size of the sample, not the size of the population.



Population and Sample XXVIII

✓ Dotplots illustrate this point. The selected samples with 250 people in each sample, but we varied the size of the population. Each dotplot contains 100 samples.

✓ Notice that the sample results look very similar. For each population, the sample results fall between about 0.58 and 0.73. In each graph, it is common for sample results to fall between about 0.62 and 0.68.

Example:

In California in 2000, the ballot included an initiative to add "none of the above" to the list of options in all candidate races. Prior to the election, a Field Poll indicated that support for the measure was 10% below opposition for the measure.

The poll was a telephone survey of 1000 registered voters in California. According to the Field Poll, "The survey was completed by telephone in English and Spanish using random digit dialing methods."

A spokesperson for the initiative was critical of the poll because only 1000 people were surveyed. He pointed out that there are 23.5 million registered voters in California.

Is his criticism valid?

Population and Sample XXX

1. No, because he was a spokesperson for the initiative, so he is biased.
   It is always a good idea to consider the source of a critical statement. But his criticism is not valid because he is saying that the size of the population affects the accuracy of the sample. This is not correct when the sample is randomly chosen.
2. Yes, because it is impossible for 1,000 people to be representative of 23,500,000 people.
   His criticism is not valid because he is saying that size of the population affects the accuracy of the sample. This is not correct when the sample is randomly chosen.

3. No, because the sample was randomly selected.
   The poll used a random sampling method. So the sample is representative of the population. The size of the population does not affect the accuracy of a random sample.
4. Yes, because a telephone survey will not represent the opinions of those without telephones.
   It is true that a telephone survey will not represent those without telephones. But there is no reason to assume that those without telephones will have a different opinion about this initiative. So the sampling method will not systematically skew the results. This situation is different from the 1936 presidential election. During that time period, many people did not have phones. Those without phones were

Population and Sample XXXII

Democrats. So sampling from telephone lists produced a biased sample in that situation.

Population and Sample XXXIII

Would his criticism be valid if this was a national initiative and only 1000 people were randomly selected from all registered voters in the nation? Why or why not?

Population and Sample XXXIV

Answer: Random sampling should produce a representative sample regardless of the population size. A random sample of 1,000 registered voters has the same level of accuracy representing California as the nation. Of course, we must make sure that the sample is randomly selected from the population of interest.

Population and Sample XXXV

Comment: *If an attempt is made to include every individual from a population in a sample, then the investigation is called a census. Every 10 years, the U.S. Census Bureau conducts a population census. It attempts to collect information about every person living in the United States. However, the population census misses between 1% and 3% of the U.S. population and accidentally counts some people more than once. A full census is possible only for small populations.*

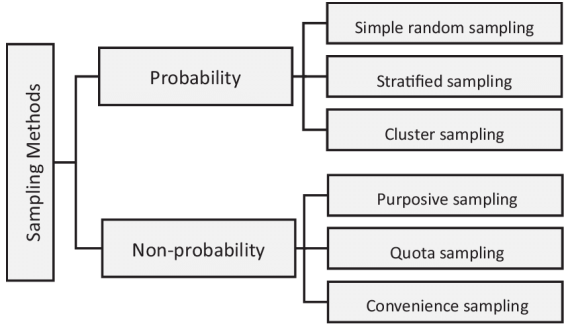
Population and Sample XXXVI

	Population	Sample
Distinction	The measurable quality is called a parameter	The measurable quality is called statistic
	The population is a complete set	The sample is subset of the population
	It contains all members of a specified group	It is a subset that represents the entire population
Symbols	N=population	n=sample size
	μ=population mean	x=sample mean
	σ=population standard deviation	s=sample standard deviation
	π=population percentage	p=sample percentage

According to E.R.Babble: A sample is a special subset of population that is observed for purposes of making inference about the nature of the total population itself.

Type of sampling:

In psychological research and other types of social research, experiments typically rely on a few different sampling methods.



Population and Sample XXXVIII

1. **Probability Sampling:** probability sampling means that every individual in a population stands a chance of being selected.

   ✓ It ensures that every subset of the population has an equal chance of being represented in the sample.

   ✓ This makes probability samples more representative, and researchers are better able to generalize their results to the group as a whole.

Different types of probability sampling

- **Simple random sampling:** Researchers take every individual in a population and randomly select their sample, often using type of compute program or random number generator.

Population and Sample XXXIX

- **Stratified random sampling:** It involves separating the population into subgroups and then taking a simple random sample from each of these subgroups.

  ✓ *For example: research might divide the population up into subgroups based on race, gender, or age and then take a simple random sample of each of these groups*.

  ✓ *Stratified random sampling often provides greater statistical accuracy than simple random sampling and helps ensure that certain groups are accurately represented in the sample.*

Population and Sample XL

- **Cluster sampling:** involves dividing a population into smaller clusters, often based upon geographic location or boundaries.
  A random sample of these clusters is then selected, and all of the subjects within the cluster are measured.
  ✓ *For example, imagine that you are trying to do a study on school principals in your state. Collecting data from every single school principal would be cost-prohibitive and time-consuming. Using a cluster sampling method, you randomly select five counties from your state and then collect data from every subject in each of those five counties.*

Population and Sample XLI

2. **Nonprobability Sampling:** It involves selecting participants using methods that do not give every subset of a population an equal chance of being represented.
   - **Convenience sampling:** involves using participants in a study who are convenient and easily available.
   - **Purposive sampling:** involves seeking out individuals that meet certain criteria.
     For example, marketers might be interested in learning how their products are perceived by women between the ages of 18 and 35. They might hire a market research firm to conduct telephone interviews that intentionally seek out and interview women that meet their age criteria.

Population and Sample XLII

- **Quota sampling:** involves intentionally sampling specific proportions of each subgroup within a population.

  ✓ *For example, political pollsters might be interested in researching the opinions of a population on a certain political issue. If they use simple random sampling, they might miss certain subsets of the population by chance. Instead, they establish criteria to assign each subgroup a certain percentage of the sample. Unlike stratified sampling, researchers use non-random methods to fill the quotas for each subgroup.*

Sampling Errors

✓ Sampling naturally cannot include every single individual in a population, therefore error can occur.

✓ **Differences between what is present in a population and what is present in a sample are known as sampling errors.**

✓ *In political polls, for example, you might often hear of the margin of errors expressed by certain confidence levels.*

✓ In general, the larger the sample size, the smaller the level of error.

✓ This is simply because as the sample becomes closer to reaching the size of the total population, the more likely it is to accurately capture all of the characteristics of the population.

Population and Sample XLIV

✓ The only way to completely eliminate sampling error is to collect data from the entire population, which is often simply too cost-prohibitive and time-consuming.

✓ Sampling errors can be minimized, however, by using randomized probability testing and a large sample size.

What Is Sample Size Formula?

✓ The sample size formula helps us find the accurate sample size through th difference between the population and sample.

✓ Since it is not possible to survey the whole population, we take a sample from the whole population and then conduct a survey or research.

The sample size formula is determined in two steps.

1. Calculate the sample size for the infinite population, and
2. Adjust the sample size to the required population.

Formula 1: Sample size for infinite population

$S = Z^2 \times P \times \frac{(1-P)}{M^2}$,

- S= sample size of infinite population
- Z = Z score
- P= population proportion (Assume as 50% or 0.50)
- M = Margin of error

Formula 2: Adjusted sample size

Adjusted Sample Size = $\frac{(S)}{1+\frac{(S-1)}{Population}}$

✓ Z score is determined based on the confidence level.

Population and Sample XLVII

✓ Confidence Level: Probability that the value of a parameter falls within a specified range of values. *For example: for 95% confidence level Z score is 1.960.*

✓ The margin of error: It is defined as a small amount that is allowed for in case of miscalculation or change of circumstances. Generally, the margin of error is taken as 5% or 0.05.

How to Apply Sample Size Formula?

✓ In order to calculate the required sample size, we need to find several other sets of values and then substitute them into an appropriate formula.

1. **Step 1: Determining Key Value:** One of the key values to be determined is the population size which refers to the total number of people within the required demographic.
2. **Step 2: Determining the margin of error or confidence interval:** The margin of error is considered to be the amount of error that can be allowed in the study. The margin of error is actually a percentage that shows how close the sample results will be with respect to the true value of the overall population that is considered in the study.

Population and Sample XLIX

- Usually, you can obtain more accurate answers with a smaller margin of error, but if a small margin of error is chosen, then you may require a larger sample.
- The margin of error usually is represented with a minus or a plus percentage when the results of a survey are presented.

✓ *For instance, 35% of people choose option B, with a margin of error of +/- 5%". In this particular example, the margin of error actually indicates that, if the question was asked to the entire population, then you are confident that between 30% (35 - 5) and 40% (35 + 5) of the people will agree with option B.*

3. **Step 3: Setting the confidence level:** The confidence level is pretty closely related to the margin of error or confidence interval. This value is used to measure the degree of certainty about how well a sample actually represents the entire population within the margin of error chosen for the study.
   - When the confidence level is chosen as 95%, then this means that you can be 95% certain that the results will accurately fall within the margin of error chosen by you.
   - When a larger confidence level is chosen, it shows a greater degree of accuracy provided that the sample size is larger. Some of the most common confidence levels used in studies are 99% confident, 90% confident, and 95% confident.
   - When the confidence level is set to 95%, then it shows that you are 95% confident that 30% to 40% of the total chosen population would definitely agree with option B of the survey.

Population and Sample LI

4. **Step 4: Specifying the standard of deviation:** The standard of deviation shows how much variation can be expected from the responses of the study.
   - Compared to the moderate results, you can expect extreme answers to be more accurate.
   - Consider an example where 1% of the survey responses says "No", and then 99% answer "Yes", then it means that the sample actually represents the overall population in an accurate manner.
   - In another case, if 55% answer "No" and 45% answer "Yes," then this means that there could be a greater chance of error.

✓ *Since this value is difficult to be calculated in an actual survey, most people choose to use 0.5 (50%) as the value which is actually the worst-case scenario percentage. Thus, using this value will actually guarantee that the calculated sample size is huge enough to show the overall population within the confidence level and the confidence interval in an accurate manner.*

5. **Step 5: Finding the Z-score:** The Z-score can be considered as a constant value that is set automatically depending on the confidence level.

✓ Z-score shows the number of standard deviations or the standard normal score between the average/mean of the population and any selected value.

✓ Due to the fact that the confidence levels are all standardized, most researchers actually memorize the required z-score for most of the commonly used confidence levels:

Population and Sample LIII

Confidence Level	Z-score
80%	1.28
85%	1.44
90%	1.65
95%	1.96
99%	2.58

Using the Standard Formula:

✓ If the size of the population is small to moderate, then it is easier to know all the key values and thus the standard formula can be used. The standard formula for calculating the sample size is:

Formula 2: Adjusted sample size

Sample size = $\frac{[Z^2 \times P(1-P)]}{e^2 \times N}$,

- N is the population size
- Z is the Z-score
- e is the margin of error
- P is the standard of deviation

Example 1: Calculate the sample size for a population of 100000. Take confidence level as 95% and margin of error as 5%.

Population and Sample LV

Solution

Give: Z=1.960, P=0.5, M=0.05

$$S = (1.960)^2 \times 0.5 \times \frac{(1-0.5)}{(0.05)^2}$$

$$= \frac{3.8416 \times 0.25}{0.0025}$$

S = 384.16 The sample size of infinite population is 3.84.16

Example 2 : Using the sample size formula, adjust the sample size for the required population in solved example 1.

Solution

Given: Z = 1.960, P =0.5, M = 0.05

Using sample size formula for adjusted sample size,

$$\text{Adjusted Sample Size} = \frac{384.16}{1+\frac{384.16-1}{100000}}$$

Required sample size = 382.69 or 383

Example 3: Using the Sample Size Formula, find the sample size for a survey where confidence level = 95%, standard deviation = .5, and margin of error = +/- 5%.

Solution

$$S = \frac{(Z-score)^2 \times SD \times (1-SD)}{(margin\ of\ error)^2}$$

$$S = \frac{(1.96)^2 \times 0.5 \times (0.5)}{(0.05)^2}$$

$$= \frac{3.8416}{0.0025}$$

$$= \frac{0.9604}{0.0025}$$

$$= 384.16$$