AI & Machine Learning

Population vs Sample: Definitions, Differences and Examples

simplilearn   All Courses

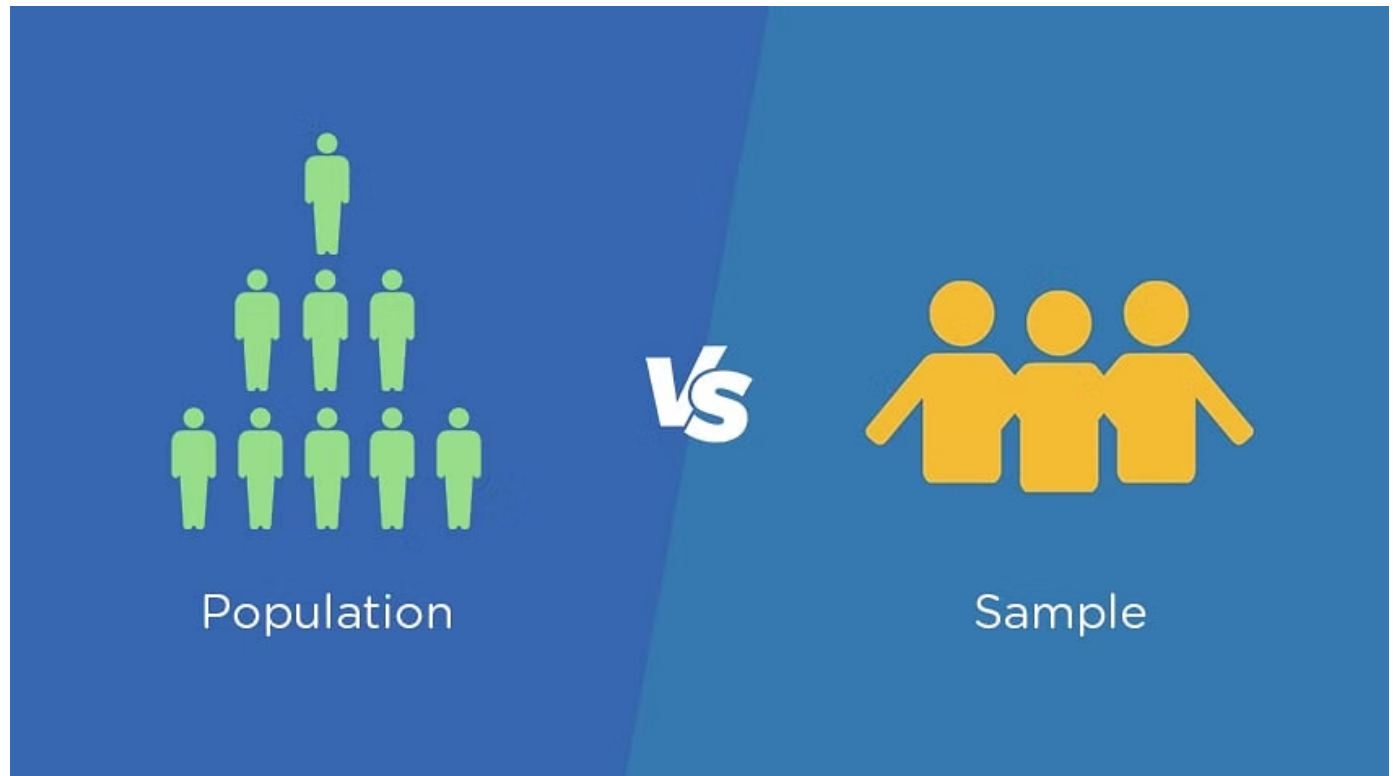By Ravikiran A S

Last updated on Aug 10, 2023   233691

Table of Contents

View More

In statistics, data plays an essential role in deciding the validity of the outcome. The data being used must be relevant, correct, and representative of all classes. While more data is good to get impartial results, it is crucial to make sure that the data collected is suitable for the problem at hand.

You can do this using population vs. sample. In this tutorial, you will learn all you need to know

about population vs. sample.
AI & Machine Learning

Population refers to the entire group of individuals about whom you wish to draw conclusions.
The sample refers to the group of people from which you will be collecting data.

What is Population?

In statistics, population is the entire set of items from which you draw data for a statistical study.
It can be a group of individuals, a set of items, etc. It makes up the data pool for a study.

Generally, population refers to the people who live in a particular area at a specific time. But in
statistics, population refers to data on your study of interest. It can be a group of individuals,
objects, events, organizations, etc. You use populations to draw conclusions.

Figure: Population

An example of a population would be the entire student body at a school. It would contain all the
students who study in that school at the time of data collection. Depending on the problem
statement, data from each of these students is collected. An example is the students who speak
Hindi among the students of a school.

For the above situation, it is easy to collect data. The population is small and willing to provide
data and can be contacted. The data collected will be complete and reliable.

data and can be contacted. The data collected will be complete and reliable.

AI & Machine Learning

If you had to collect the same data from a larger population, say the entire country of India, it would be impossible to draw reliable conclusions because of geographical and accessibility constraints, not to mention time and resource constraints. A lot of data would be missing or might be unreliable. Furthermore, due to accessibility issues, marginalized tribes or villages might not provide data at all, making the data biased towards certain regions or groups.

Your AI/ML Career is Just Around The Corner!

AI Engineer Master's Program

EXPLORE PROGRAM

What is a Sample?

A sample is defined as a smaller and more manageable representation of a larger group. A subset of a larger population that contains characteristics of that population. A sample is used in statistical testing when the population size is too large for all members or observations to be included in the test.

The sample is an unbiased subset of the population that best represents the whole data.

To overcome the restraints of a population, you can sometimes collect data from a subset of your population and then consider it as the general norm. You collect the subset information from the groups who have taken part in the study, making the data reliable. The results obtained for different groups who took part in the study can be extrapolated to generalize for the population.
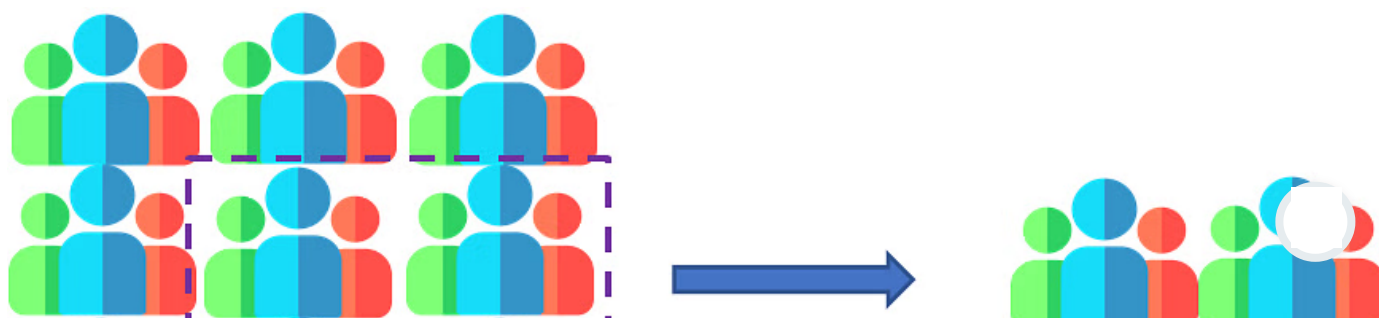
Figure: Sample

The process of collecting data from a small subsection of the population and then using it to generalize over the entire set is called Sampling.

Samples are used when :

- The population is too large to collect data.

- The data collected is not reliable.

- The population is hypothetical and is unlimited in size. Take the example of a study that documents the results of a new medical procedure. It is unknown how the procedure will affect people across the globe, so a test group is used to find out how people react to it.

A sample should generally :

- Satisfy all different variations present in the population as well as a well-defined selection criterion.

- Be utterly unbiased on the properties of the objects being selected.

- Be random to choose the objects of study fairly.

Say you are looking for a job in the IT sector, so you search online for IT jobs. The first search result would be for jobs all around the world. But you want to work in India, so you search for IT jobs in India. This would be your population. It would be impossible to go through and apply for all positions in the listing. So you consider the top 30 jobs you are qualified for and satisfied with and apply for those. This is your sample.

Population and Sample Formulas

Population Parameter:

Mean: μ = (ΣX) / N, where ΣX is the sum of all values in the population and N is the size of population

AI & Machine Learning
Standard Deviation: σ = √[(Σ(X-μ)²) / N], where X is a value in the population, μ is the population mean, and N is the size of the population

simpli·learn   All Courses

Sample Statistic:

Mean: x̄ = (Σx) / n, where Σx is the sum of all values in the sample and n is the size of the sample

Standard Deviation: s = √[(Σ(x-x̄)²) / (n-1)], where x is a value in the sample and x̄ is the sample mean

Note that the formulas for the population parameter and sample statistic are similar, but they use different notation and have slightly different calculations. The population parameter uses the entire population, while the sample statistic uses a subset (i.e., sample) of the population.

Population Parameter vs. Sample Statistic

Population parameter and sample statistic are two important concepts in statistics that are used to describe a population or a sample.

A population parameter is a numerical value that describes a characteristic of a population, such as the mean or standard deviation. It is usually unknown and is estimated from sample data. For example, the population mean height of all students in a school is a population parameter.

A sample statistic, on the other hand, is a numerical value that describes a characteristic of a sample, such as the sample mean or sample standard deviation. It is calculated from sample data and used to make inferences about the population. For example, the sample mean height of a group of randomly selected students is a sample statistic.

The key difference between a population parameter and a sample statistic is that the former describes the entire population, while the latter describes only a sample from the population. In general, population parameters are more precise and accurate, as they are calculated using all available data. However, they are usually unknown and can only be estimated from sample data, which is where sample statistics come into play.

Parameter	Population Parameter	Sample Statistic

AI & Machine Learning

All Courses

Definition	Numerical value that describes a characteristic of a population	Numerical value that describes characteristic of a sample
Calculation	Calculated using data from the entire population	Calculated using data from sample of the population
Purpose	Used to describe the entire population	Used to estimate the popula parameter
Precision and Accuracy	Usually more precise and accurate than sample statistic	Usually less precise and le accurate than population parameter
Notation	Greek letters (e.g., μ for population mean)	Roman letters (e.g., $\bar{x}$ for sam mean)
Example	The population mean income of all households in a country	The sample proportion of pe who own a car in a random selected group of househol

Your AI/ML Career is Just Around The Corner!

AI Engineer Master's Program

EXPLORE PROGRAM

Similarities Between Population and Sample

Population and sample are both concepts used in statistical analysis. Here are some similarities between population and sample:

Data: Both population and sample involve data. Population refers to the entire group or set of individuals, objects, or events being studied, while a sample is a subset of the population that is used for analysis.

Descriptive Statistics: Descriptive statistics can be used to analyze both populations and samples. For example, measures of central tendency, such as mean and median, and measures of variability, such as standard deviation and range, can be calculated for both populations and samples.

Probability: Probability theory can be used to analyze both populations and samples. For example, the probability of an event occurring can be calculated for both populations and samples.

Inferential Statistics: Inferential statistics can be used to draw conclusions about the population based on the sample. By using probability theory, inferential statistics can estimate population parameters, such as mean and variance, from the sample statistics.

Sampling Error: Sampling error is a potential source of error in both populations and samples. Sampling error refers to the difference between the sample statistics and the population parameters that they are meant to estimate.

How to Collect Data From a Sample?

Samples are used when the population is large, scattered, or if it's hard to collect data on individual instances within it. You can then use a small sample of the population to make overall hypotheses.

Samples should be randomly selected and should represent the entire population and every class within it. To ensure this, statistical methods such as probability sampling, are used to collect random samples from every class within the population. This will reduce sampling bias and increase validity.

AI & Machine L

simpl

urses

Figure: Collecting random samples

Consider the polls conducted during election season to gauge the public support for various political parties all over the nation. It is impossible to ask millions of voters who their preferred candidate is, so they collect the opinions of a few hundred or thousand people from different sectors of the voting population.

That was all about population vs. sample.

Future-Proof Your AI/ML Career: Top Dos and Don'ts

Free Webinar | 5 Dec, Tuesday | 7 PM IST

REGISTER NOW

Importance of Accurate Population Definition and Measurement (Under Population)

Accurate population definition and measurement are crucial in many fields, including public health, social sciences, and business, among others. Here are some reasons why:

AI & Machine Learning if the population is not defined and measured accurately, the results obtained may not be valid. For example, if a study on the prevalence of a disease only includes certain subgroups of the population, the results may not be representative of the entire population.

Generalizability: Accurate population definition and measurement are essential to ensure that the results obtained from a study can be generalized to the entire population. If the population is not well-defined or measured accurately, the findings may not be applicable to other groups or contexts.

Resource Allocation: In many cases, accurate population definition and measurement are necessary for resource allocation decisions. For example, government agencies need accurate population data to determine funding priorities for various programs and services.

Planning and Policy Development: Accurate population data are necessary for effective planning and policy development. For example, in urban planning, accurate population data can help identify areas of high population density and inform decisions about where to build new infrastructure.

Ethical Considerations: Accurate population definition and measurement are also important for ethical reasons. Inaccurate or biased population data can lead to unfair treatment of certain groups or populations.

Importance of Accurate Sampling and Sample Size Determination (Under Sample)

Accurate sampling and sample size determination are essential in many fields, including research, market analysis, and quality control, among others. Here are some reasons why:

Representative Results: Sampling is used when it is impractical or impossible to study an entire population. By using a representative sample, the results obtained can be generalizable to the entire population. An accurate sample ensures that the results are representative of the population and are not biased or misleading.

Resource Efficiency: Sampling is often more efficient and cost-effective than studying an entire population. Accurate sampling techniques can reduce the number of participants needed, saving time and resources.

Precision of Results: Sample size determination is crucial to ensure that the results obtain e precise and reliable. A sample that is too small can lead to imprecise results, while a sample that

is too large can be unnecessarily costly.
AI & Machine Learning

Generalizability: Similar to accurate population definition and measurement, accurate sampling and sample size determination are necessary to ensure that the results obtained from a study can be generalized to the entire population.

Ethical Considerations: Accurate sampling and sample size determination are also important for ethical reasons. If a sample is not representative of the population, it can lead to unfair treatment of certain groups or populations.

Inference

Inference is a statistical technique used to draw conclusions or make predictions about a population based on data from a sample. It involves using probability theory and statistical methods to estimate population parameters, such as mean or variance, from the sample statistics. Inference can help researchers make informed decisions, identify patterns and relationships in the data, and determine whether the results are significant or due to chance. It is widely used in research, marketing, quality control, and many other fields where data analysis is necessary.

How Population and Sample are Used in Statistical Inference

In statistical inference, population and sample are used to estimate population parameters using sample statistics. The sample is used as a representation of the population, and probability theory and statistical methods are applied to draw conclusions or make predictions about the population based on the sample data.

The use of population and sample in statistical inference is essential because it is often impractical or impossible to study the entire population. Instead, a representative sample is selected, and statistical inference is used to estimate the population parameters. If the sample is selected correctly and is representative of the population, statistical inference can provide accurate and reliable estimates of the population parameters. However, incorrect sampling techniques or failure to define the population accurately can lead to biased or inaccurate estimates. Thus, careful consideration of population and sample is crucial for accurate statistical inference.

Your AI/ML Career is Just Around The Corner!

AI Engineer Master's Program

EXPLORE PROGRAM

Examples of Statistical Inference Using Population and Sample Data

Statistical inference using population and sample data can be applied in various fields. Here are some examples:

Medical Research: In medical research, clinical trials are conducted on a sample of the population to estimate the effects of a drug or treatment. Statistical inference is used to estimate the effect size and determine the probability that the results are due to chance.

Market Research: In market research, a sample of customers is surveyed to estimate the demand for a product or service. Statistical inference is used to estimate the proportion of the population that would be interested in the product or service.

Quality Control: In quality control, a sample of products is tested to estimate the proportion of defective items in the population. Statistical inference is used to determine whether the proportion of defects in the sample is significantly different from the population.

Political Polling: In political polling, a sample of voters is surveyed to estimate the proportion of voters who support a candidate or party. Statistical inference is used to estimate the margin of error and determine the probability of a candidate winning the election.

In all these examples, statistical inference using population and sample data is used to draw conclusions or make predictions about the population of interest. By using probability theory and statistical methods, researchers can estimate population parameters, such as proportions or means, and determine the likelihood that the results are due to chance.

Choose the Right Program

Supercharge your career in AI and ML with Simplilearn's comprehensive courses. Gain the skills and knowledge to transform industries and unleash your true potential. Enroll now and unlock limitless possibilities!

AI & Machine Learning

simplilearn

All Courses

Program Name	AI Engineer	Post Graduate Program In Intelligence
Geo	All Geos	All Geos
University	Simplilearn	Purdue
Course Duration	11 Months	11 Months
Coding Experience Required	Basic	Basic
Skills You Will Learn	10+ skills including data structure, data manipulation, NumPy, Scikit-Learn, Tableau and more.	16+ skills includin chatbots, NLP, Python, K more.
Additional Benefits	Get access to exclusive Hackathons, Masterclasses and Ask-Me-Anything sessions by IBM Applied learning via 3 Capstone and 12 Industry-relevant Projects	Purdue Alumni Assoc Membership Free IIMJc Membership of 6 mo Resume Building Assis
Cost	$$	$$$$
	Explore Program	Explore Program

Conclusion

In this tutorial titled 'population vs. sample,' you look at what population and sample mean in
AI & Machine Learning
statistics with the help of examples, some of the differences between population vs. sample You
then looked at how data is collected from a population and a sample.

simplilearn   All Courses

We hope this helped you understand what population and sample mean in statistics. To learn
more about statistics and machine learning, check out Simplilearn's Caltech Post Graduate
Program in AI and Machine Learning. If you have any questions or doubts, mention them in this
tutorial's comments section, and we'll have our experts answer them for you at the earliest!

Happy learning!

Find our Caltech Post Graduate Program In AI And Machine Learning Online Bootcamp in top cities:

Name	Date	Place
Caltech Post Graduate Program In AI And Machine Learning	Cohort starts on 8th Feb 2024, Weekend batch	Your City
Caltech Post Graduate Program In AI And Machine Learning	Cohort starts on 12th Feb 2024, Weekend batch	Your City
Caltech Post Graduate Program In AI And Machine Learning	Cohort starts on 5th Mar 2024, Weekend batch	Your City

About the Author

Ravikiran A S

Ravikiran A S works with Simplilearn as a Research Analyst. He an enthusiastic geek always in
the hunt to learn the latest technologies. He is proficient with Java Programming Language, Big
Data, and pow...

View More