IE6400 Foundations for Data Analytics Engineering

Fall 2023

Module 2: Data and Sampling Distributions

- FULL READING -

Data Distribution:

A data distribution refers to the way in which a set of data points is spread or distributed. When you collect data and plot it on a graph, the shape that data takes can be referred to as its distribution.

Sampling Distribution:

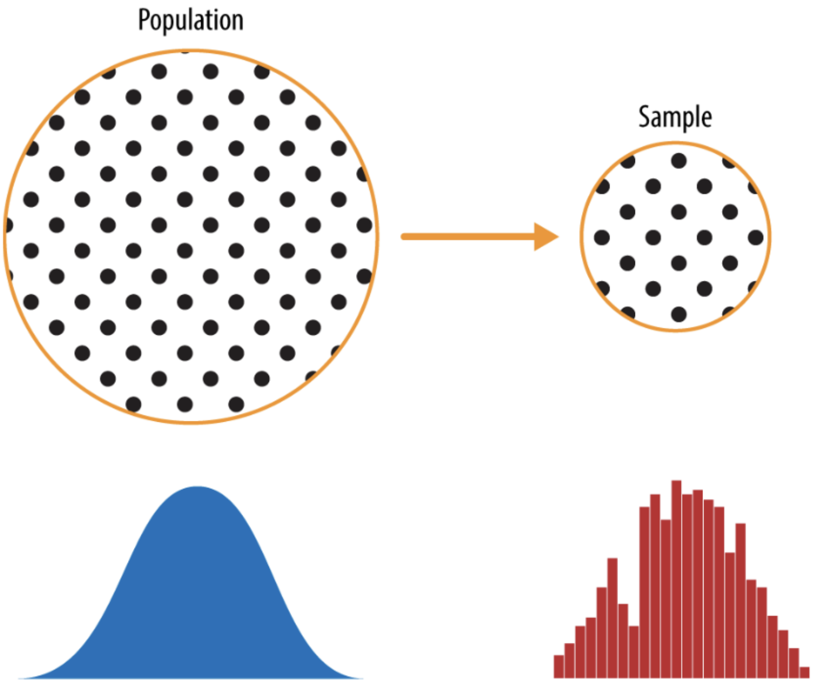
A sampling distribution is a probability distribution of a statistic obtained through a large number of samples drawn from a specific population. The sampling distribution of a given population is the distribution of frequencies of a range of different outcomes that could possibly occur for a statistic of a population.

Population and Sample

A **population** is a complete set of people with a specialized set of characteristics, and a **sample** is a subset of the population.

A **population** is the entire group that you want to draw conclusions about.

A **sample** is the specific group that you will collect data from. The size of the sample is always less than the total size of the population. To get sample, a *sampling procedure* is used.



In general, data scientists need not worry about the theoretical nature of the population, and instead should focus on the sampling procedures and the data at hand.

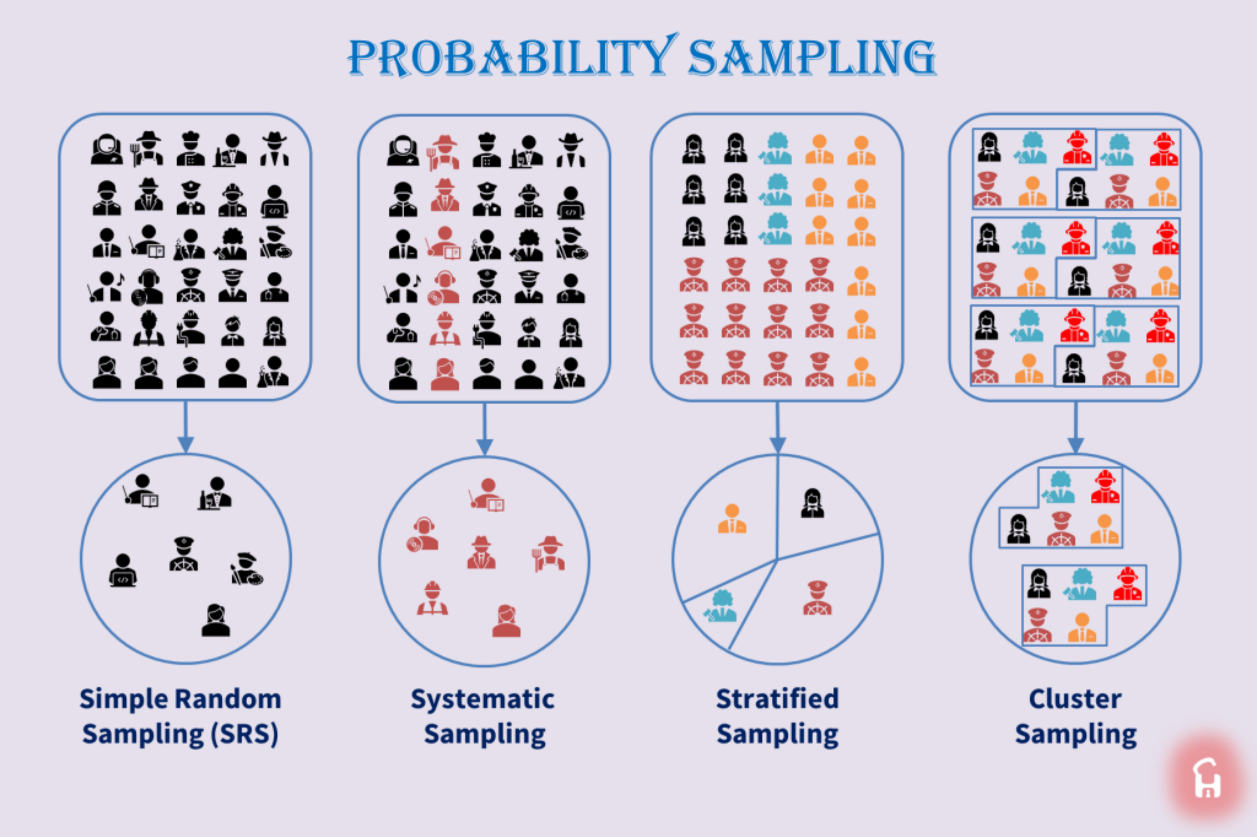
Sampling Procedure

To draw valid conclusions from your results, you have to carefully decide how you will select a sample that is representative of the group as a whole. This is called a sampling method. There are two primary types of sampling methods that you can use in your research:

- Probability sampling involves random selection, allowing you to make strong statistical inferences about the whole group.
- Non-probability sampling involves non-random selection based on convenience or other criteria, allowing you to easily collect data.

Probabilistic Sampling Methods

Probability sampling means that every member of the population has a chance of being selected. If you want to produce results that are representative of the whole population, probability sampling techniques are the most valid choice.



1. Simple Random Sample

In a simple random sample, every member of the population has an equal chance of being selected. Your sampling frame should include the whole population.

To conduct this type of sampling, you can use tools like random number generators or other techniques that are based entirely on chance.

When you select records randomly from a larger data set, you can achieve the sampling in a few different ways:

- sampling without replacement, in which a subset of the observations are selected randomly, and once an observation is selected it cannot be selected again.
- sampling with replacement, in which a subset of observations are selected randomly, and an observation may be selected more than once.

2. Biased Sampling

Sampling bias occurs when some members of a population are systematically more likely to be selected in a sample than others.

Sampling bias limits the generalizability of findings because it is a threat to external validity, specifically population validity. In other words, findings from biased samples can only be generalized to populations that share characteristics with the sample.

3. Stratified Sampling 分层采样

Stratified Sampling is a sampling technique used to obtain samples that best represent the population. It reduces bias in selecting samples by dividing the population into homogeneous subgroups called strata, and randomly sampling data from each stratum(singular form of strata).

4. Systematic Sampling

Systematic sampling is a method that involves selecting every nth element from the population, where "n" is a constant determined by the sampling interval. This method is useful when you want to ensure a random and representative sample without going through the entire population.

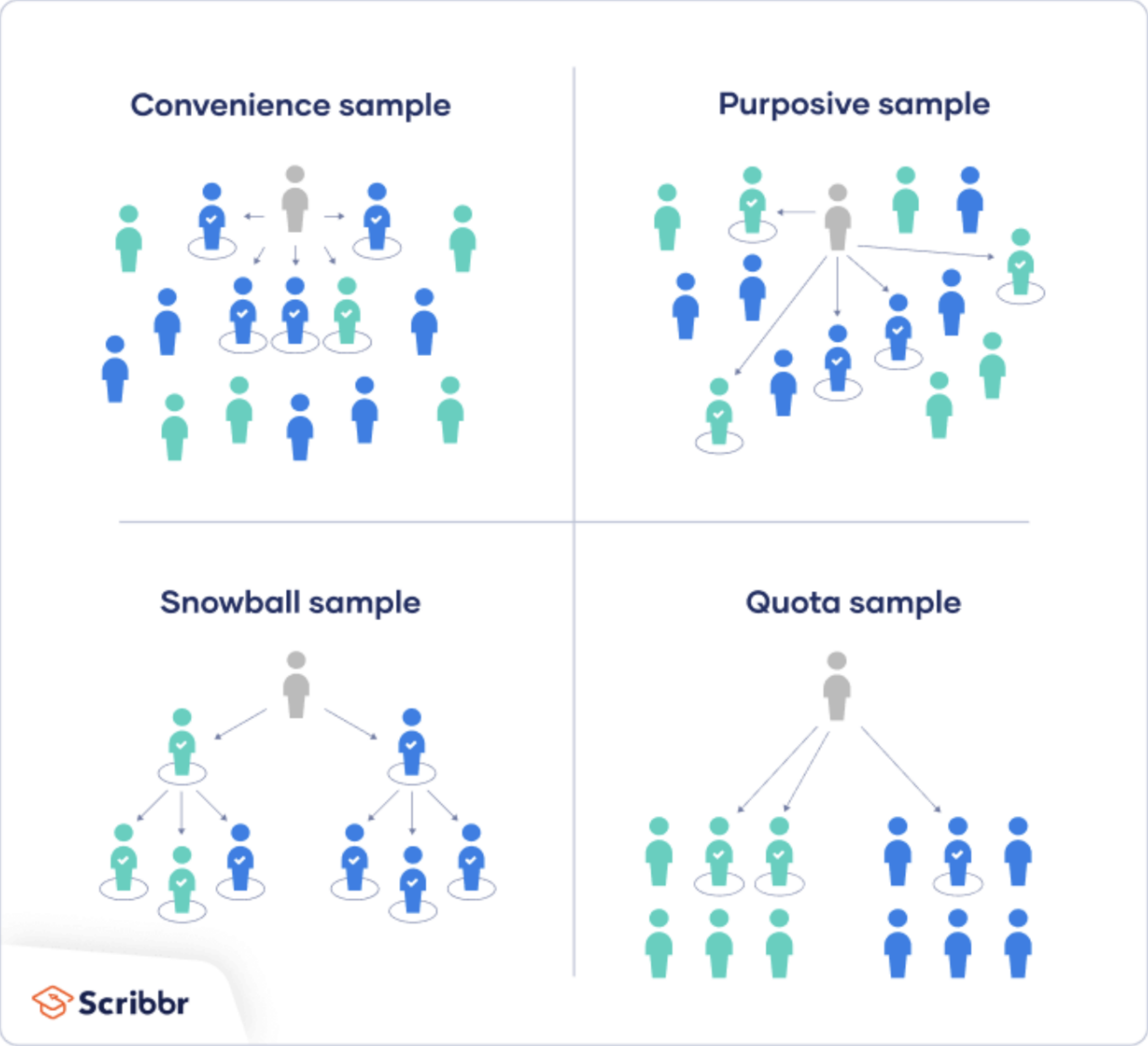
5. Cluster Sampling 集群采样

Cluster sampling is a sampling technique used in statistics and research to select a sample from a large population that is divided into clusters or groups. Unlike simple random sampling, where individual elements are selected randomly from the entire population, cluster sampling involves randomly selecting entire clusters and then sampling all the elements within those selected clusters.

Non-Probabilistic Sampling Methods

In a non-probability sample, individuals are selected based on non-random criteria, and not every individual has a chance of being included.

This type of sample is easier and cheaper to access, but it has a higher risk of sampling bias.



1. Convenience Sampling

Convenience sampling is a non-probability sampling technique where you select a sample that is easily accessible or convenient to collect, rather than using random or structured methods.

2. Purposive Sampling

Purposive sampling, also known as judgmental or selective sampling, is a non-probability sampling technique used in research to select a sample based on specific criteria or the researcher's judgment. Unlike random sampling methods, purposive sampling involves deliberately choosing individuals or elements from the population that possess certain characteristics or knowledge relevant to the research question.

3. Snowball Sampling

Snowball sampling is a non-probability sampling technique commonly used in social research, especially when studying hard-to-reach or hidden populations. It is a method of sampling where initial participants are selected, and then they help identify and recruit additional participants who meet the research criteria. The process continues like a snowball rolling down a hill, gradually accumulating more participants.

4. Quota Sampling

Quota sampling is a non-probability sampling technique used in research to create a sample that is representative of a larger population with specific characteristics or attributes. In quota sampling, the researcher divides the population into subgroups or strata based on certain criteria, such as age, gender, income, or education. The researcher then sets quotas for each subgroup based on the desired proportions in the final sample. Individuals are then selected non-randomly to fill these quotas until they are met.

Descriptive Statistics

Descriptive statistics is about describing and summarizing data. It uses two main approaches:

- The quantitative approach describes and summarizes data numerically.
- The visual approach illustrates data with charts, plots, histograms, and other graphs.

The are several types of measures in descriptive statistics:

- **Central tendency** tells you about the centers of the data. Useful measures include the mean, median, and mode.
- **Variability** tells you about the spread of the data. Useful measures include variance and standard deviation.
- **Correlation** or joint variability tells you about the relation between a pair of variables in a dataset. Useful measures include covariance and the correlation coefficient.

Central Tendency

1. Mean

The sample mean, also called the sample arithmetic mean or simply the average, is the arithmetic average of all the items in a dataset. In other words, it's the sum of all the elements in the dataset divided by the number of items in the dataset.

The harmonic mean is the reciprocal of the mean of the reciprocals of all items in the dataset: n / $\Sigma_i(1/x_i)$, where i = 1, 2, ..., n and n is the number of items in the dataset x.

A trimmed mean is a statistical measure that calculates the mean of a dataset after a certain percentage of the smallest and largest values have been removed. It is a way to reduce the influence of outliers or extreme values on the mean, making it more robust to extreme data points.

2. Median

The sample median is the middle element of a sorted dataset. The dataset can be sorted in increasing or decreasing order. If the number of elements n of the dataset is odd, then the median is the value at the middle position: 0.5(n + 1). If n is even, then the median is the arithmetic mean of the two values in the middle, that is, the items at the positions 0.5n and 0.5n + 1.

3. Mode

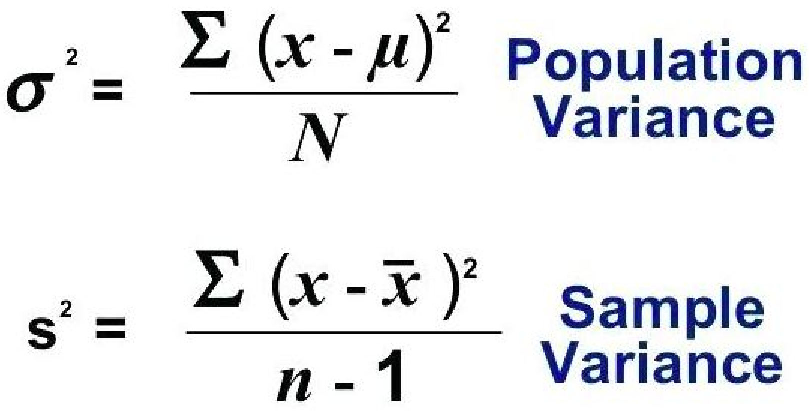
The mode is a statistical measure that represents the most frequently occurring value in a dataset.

Variability

1. Variance

Variance is a statistical measure that quantifies the spread or dispersion of a dataset. It provides information about how much individual data points deviate from the mean (average) of the dataset.

Population variance and sample variance both measure the dispersion or spread of data. However, they differ in the data they analyze. Population variance is calculated using the entire population data, while sample variance is calculated using a subset of the population known as a sample.



$$\sigma^2 = \frac{\Sigma (x - \mu)^2}{N}$$

Population Variance

$$s^2 = \frac{\Sigma (x - \bar{x})^2}{n - 1}$$

Sample Variance

2. Standard Deviation

The standard deviation is a statistical measure of the amount of variation or dispersion in a dataset. It quantifies how individual data points deviate from the mean (average) of the dataset.

Standard deviation is a squared root of variance.

3. Range

The range is a simple statistical measure that represents the spread or dispersion of data in a dataset. It is the difference between the maximum and minimum values in the dataset. In other words, it provides a measure of how much the data values vary from each other.

4. Interquartile Range

The interquartile range (IQR) is a statistical measure of the spread or dispersion of a dataset. It represents the range between the first quartile (25th percentile) and the third quartile (75th percentile) of the data when it is ordered in ascending or descending order. The IQR is useful because it is less sensitive to outliers compared to the range, which is based on the minimum and maximum values.

Summarizing Data

Summarizing data is an essential step in data analysis and reporting. It involves presenting key statistics and insights about a dataset to help understand its characteristics.

Visualizations

1. Stem and Leaf Plots in Python

A Stem and Leaf Plot is a special table where each data value is split into a "stem" (the first digit or digits) and a "leaf" (usually the last digit).

2. Frequency plot

A frequency plot is a graph that shows the pattern in a set of data by plotting how often particular values of a measure occur.

3. Box Plot

A box plot, also known as a box-and-whisker plot, is a data visualization that provides a graphical summary of the distribution of a dataset. It displays the median, quartiles, and potential outliers of the data. Box plots are particularly useful for comparing the distribution of data across different categories or groups.