SAMPLING

THEME 4: SAMPLING

OUTCOMES

7.1.1. *Sampling Distributions.* Outcomes: At the end of this study unit, the student should be able to:

- Understand simple methods on how to simulate/sample from simple continuous and discrete probability distributions
- Implement sampling techniques in Python
- Understand Monte Carlo Integration and simulation
- Perform a simulation study to discover probability laws
- Know how to use samples to make statements about the population
- Implement the probability integral transform method for sampling data.
- Explain and implement the bootstrap method.
- Define and implement the acceptance/rejection algorithm.

WHAT IS SAMPLING?

GENERALLY SPEAKING, A SAMPLING METHOD IS A TOOL CONSISTING IN DRAWING A SUBSET FROM A DATASET/POPULATION AND CALCULATING STATISTICS AND METRICS ON THE SAMPLE IN ORDER TO OBTAIN FURTHER INFORMATION ABOUT SOMETHING:

- IN THE MACHINE LEARNING SETTING, THIS SOMETHING IS THE PERFORMANCE OF A MODEL.
- IN STATISTICAL ANALYSIS THIS COULD BE ADDITIONAL INSIGHT ABOUT THE BEHAVIOR OF SOME PARAMETER OR POPULATION.

SAMPLING VS RESAMPLING

Sampling

- Sampling is a crucial concept in statistics that involves selecting a subset of individuals or items from a larger population for the purpose of making inferences about the population.
- There are many sampling techniques that can be used to gather a data sample depending upon the need and situation.

Resampling

- Resampling methods are an indispensable tool in modern statistics. They involve repeatedly drawing samples from a training set and refitting a model of interest on each sample in order to obtain additional information about the fitted model.
- Understand the effectiveness of the model without resorting to the test set

ADVANTAGES OF SAMPLING – WHY DO WE NEED IT?

- In many cases, it's impractical or impossible to study an entire population due to its size, cost, time constraints, etc.
- Sampling allows us to obtain information about a population by studying a smaller, manageable subset.
- It is more efficient and economical than collecting data from the entire population.
- Accurate and unbiased sampling can provide reliable insights about a population.
- Well-designed samples can yield results that closely mirror the population characteristics, leading to meaningful and generalizable conclusions.
- Where is it used?
  - Sampling is used in various fields such as market research, medical studies, social sciences, quality control, and more.
  - In market research, companies might survey a subset of their customers to infer preferences or attitudes of the entire customer base.
  - In medical studies, clinical trials often use samples to draw conclusions about the effects of treatments on a larger patient population.

WHAT TO CONSIDER WHEN SAMPLING

- Sample Goal: The population property that you wish to estimate using the sample.
- Population: The scope or domain from which observations could theoretically be made.
- Selection Criteria: The methodology that will be used to accept or reject observations in your sample.
- Sample Size: The number of observations that will constitute the sample.
- Margin of error: The amount of uncertainty in our results that comes from only looking at a sample of the population instead of everyone. The smaller the margin of error, the more accurate our results will be.
- Confidence level: How sure we can be that the results of a survey or experiment are accurate and not just due to chance. Usually, confidence level is expressed as a percentage.
- Population proportion: The percentage of people or things in a population that share the same characteristic.

SAMPLING METHODS

THERE ARE TWO MAIN CATEGORIES OF SAMPLING METHODS: PROBABILITY SAMPLING AND NON-PROBABILITY SAMPLING.

Probability

- Probability sampling methods involve random selection, ensuring that each element in the population has a known, nonzero chance of being included in the sample.
- These methods are more likely to produce representative samples.
- Common types include:
- Simple Random Sampling
- Stratified Sampling
- Systematic Sampling
- Cluster Sampling

Non-probability

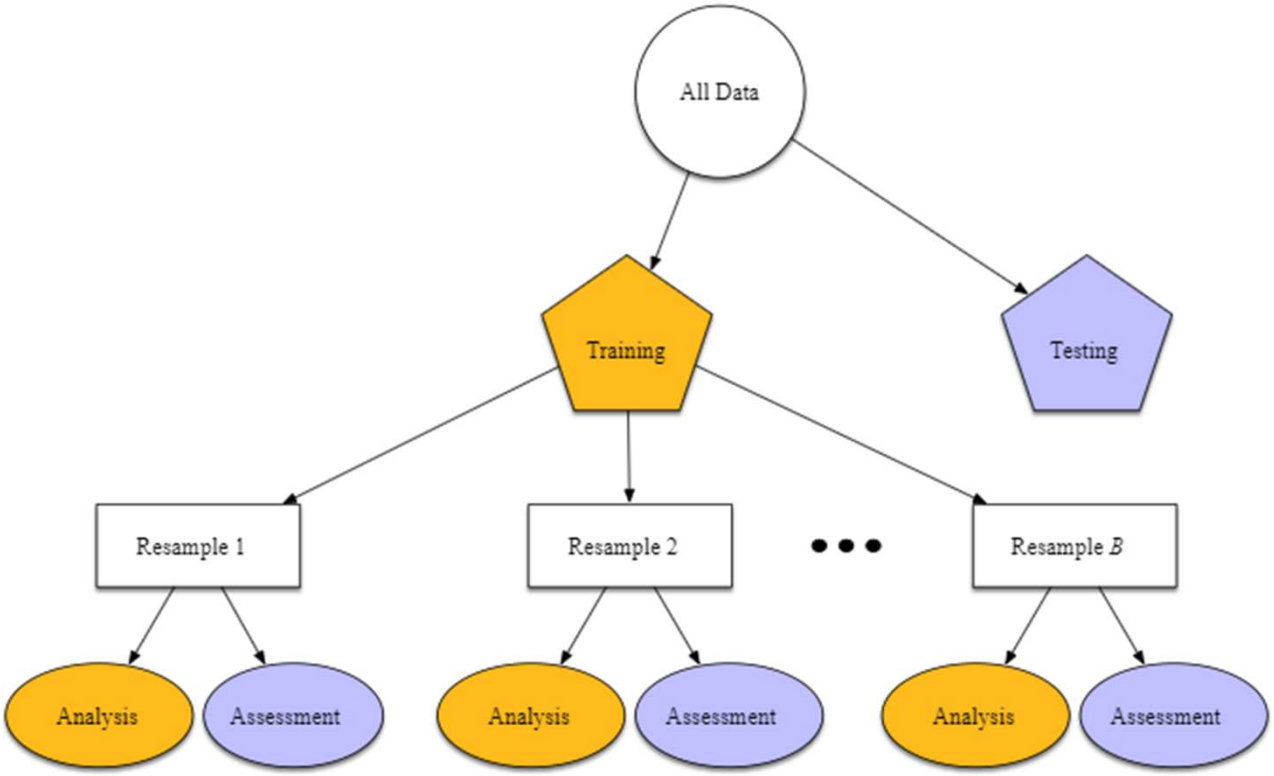
- Non-probability sampling methods do not involve random selection and may lead to biased samples.
- They are commonly used when random sampling is difficult or impractical.
- Examples include:
- Convenience Sampling
- Judgmental (or Purposive) Sampling
- Snowball Sampling

Different sampling methods have their own strengths and limitations.
The choice of sampling method depends on the research objectives, available resources, and the nature of the population.

RESAMPLING METHODS

- Once we have a data sample, it can be used to estimate the population parameter. The problem is that we only have a single estimate of the population parameter, with little idea of the variability or uncertainty in the estimate.
- One way to address this is by estimating the population parameter multiple times from our data sample. This is called resampling.
- A downside of the methods is that they can be computationally very expensive, requiring tens, hundreds, or even thousands of resamples in order to develop a robust estimate of the population parameter.
- Two commonly used resampling methods:
  - **Bootstrap** - Samples are drawn from the dataset with replacement (allowing the same sample to appear more than once in the sample), where those instances not drawn into the data sample may be used for the test set.
  - **k-fold Cross-Validation** - A dataset is partitioned into k groups, where each group is given the opportunity of being used as a held out test set leaving the remaining groups as the training set.

RESAMPLING PICTURE



SAMPLING TECHNIQUES AND ALGORITHMS

- In this course, we will focus on specific sampling techniques.
- Next we are going to define and implement the following techniques:
  - Simulation study
  - Monte Carlo Integration
  - Probability Integral Transformation
  - Bootstrap Method
  - Acceptance/Rejection Algorithm

SIMULATION

- Simulation is the process of using a computer to mimic a physical experiment. In this class, those experiments will almost invariably involve chance. For example we can simulate the results of a coin toss experiment.
- Simulation and sampling are powerful tools that cater to different analytical needs.
- Simulation helps us understand complex systems, predict outcomes, and explore scenarios.
- Sampling aids in making inferences about populations, estimating parameters, and optimizing resources.
- Examples of simulating data you have encountered thus far = generating random values from a known probability distribution.

- Homework … write an algorithm to generate a sample of random values from discrete and continuous distributions in Python. Consider the following distributions: Binomial, Poisson, Uniform, Normal, Gamma, Weibull, Pareto and Exponential.