Disclaimer: Please note that any topics that are not covered in today's lecture will be covered in the next lecture.

Content

- **Population vs Sample**
- **Sample Statistics**
  - Sample Mean
  - Sample variance
  - Sample Standard Deviation
- **Point Estimates**
- **Standard Error**
- **Sampling techniques**
  - Random Sampling
- **Uniform Distribution**

Population vs Sample

Population:

- Think of a population as the entire set of items under study.
- This could be the entire group of interest, whether it's people, objects, data points, or any other relevant entities.
- We use populations to draw conclusions.

For example:

If we were conducting a survey to understand the average income of all residents in a city,

- the population in this case would be every single resident in that city.

Sample:

- A sample is a smaller, manageable subset of the population.

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

To overcome the restraints of a population, we can sometimes collect data from a subset of our population and then consider it as the general norm.

Example:

- Suppose from an entire city we have selected and collected data from 500 residents from that city, this group of 500 individuals would represent our sample.

The idea here is that the characteristics of the sample should resemble the characteristics of the entire population, allowing us to make **inferences** or **estimates** about the population using the sample data.

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**.

Conclusion

> **Why should we use samples instead of studying entire populations?**

- Practicality:
- Reduced Error:
- In some cases, testing or studying an entire population is destructive or impractical.

Sample Statistics

A **parameter** is a measure that describes the **whole population**

Population data has a few population parameters like:

1. Population Mean
2. Population Variance
3. Population SD

A **statistic** is a measure that describes the **sample**.

Sample statistics are descriptive values calculated from sample data.

Sample statistics include measures like:

1) Sample Mean $(\bar{x})$

The sample mean, denoted as $\bar{x}$, is the average value of a set of data points within a sample.

Formula:

$$\bar{x} = \frac{\sum x_i}{n}$$

Where,

- $\bar{x}$ = Sample mean
- $\sum x_i$ = sum of each value in the sample

- n = number of values in the sample

Difference from Population Mean (μ):

- The sample mean is an estimate of the population mean and is calculated from sample data.
- It may vary from one sample to another but is expected to be close to the population mean when using a sufficiently large, random sample.

Example:

Suppose we want to estimate the average income of people in a city.

Instead of surveying the entire population, we randomly select 100 individuals and calculate their average income, which is the sample mean.

2) Sample Variance (s^2)

The sample variance, denoted as s^2, measures the spread or dispersion of data within a sample.

The sample variance is used to make estimates or inferences about the population variance.

Formula:

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

Where,

- s^2 = Sample variance
- x_i = represents individual data points.
- $\bar{x}$ = sample mean
- n = number of data points in the sample

How it is different from population variance.

- The population variance is a parameter that describes the variability in the entire population.
- The sample variance is a statistic used to estimate the population variance based on a sample.

Sample Standard Deviation (s)

The sample standard deviation, denoted as s, is a measure of how spread out the data in a sample is.

Formula

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

How it is different from Population Standard Deviation (σ):

- The sample standard deviation estimates the population standard deviation but may not match it exactly due to the finite size of the sample.

How it is different from Population Standard Deviation (σ):

- The sample standard deviation estimates the population standard deviation but may not match it exactly due to the finite size of the sample.

Point Estimates

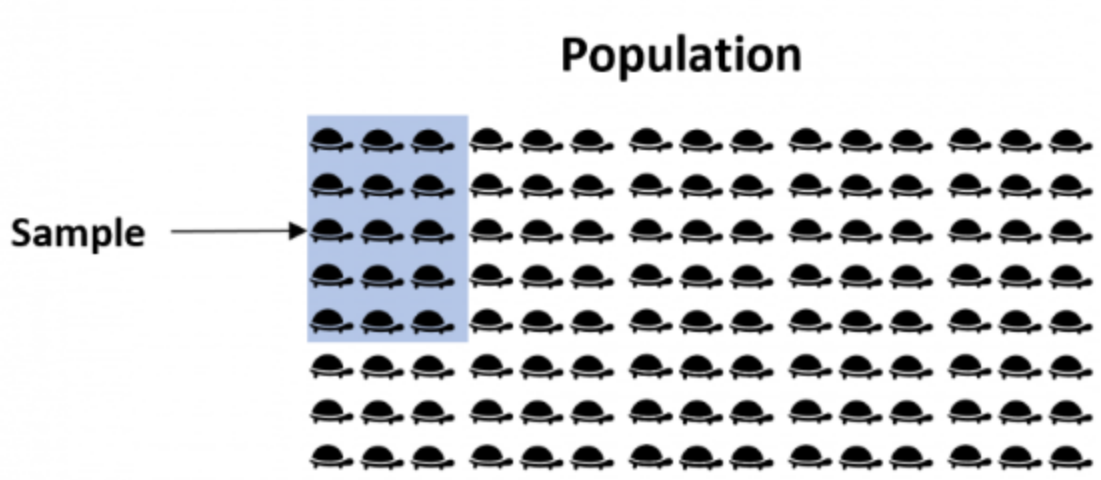
Now, we will use these sample statistics as a point estimate to estimate population parameters.

- It serves as the best guess for the true, but often unknown, population parameters.
- For instance, the sample mean estimates the population mean, and the sample variance estimates the population variance.

> **Imagine a scenario:**

```
We want to estimate the mean weight of a certain species of turtle in Florida.
```

We opt to select a **random sample of 50 turtles** and use the mean weight of this sample to estimate the true population mean



Assume the sample mean is **150.4 pounds**,

Then our point estimate for the true population mean of the entire species would be 150.4 pounds.

> **Importance of representative samples**

To get accurate insights about a whole group, we need a sample that reflects the group's main traits.

If our sample closely resembles the population, we can trust that estimates drawn from it are reliable and unbiased reflections of the entire population.

Sampling Techniques

Sampling methods/techniques refer to how we select members from the population to be in the study.

There are two primary types of sampling methods that we can use in our research:

1. **Probability sampling**
    - In this every member of the population has a chance of being selected, allowing us to make strong statistical inferences about the whole group.
2. **Non-probability sampling**

- In this method, 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.

Ideally, a sample should be randomly selected and representative of the population.

In this module, we are going to study about probability sampling and that too Random Sampling and in the ML module we will see other sampling techniques.

Simple Random Sampling

A simple random sample is a randomly selected subset of a population.

- In this sampling method, each member of the population has an exactly equal chance of being selected which tends to produce representative, unbiased samples.

For instance,

If we want to pick 1000 individuals from a town with 100,000 residents, each person has a 0.01 probability of being selected.

This straightforward calculation doesn't require in-depth knowledge of the population's composition, Hence, simple random sampling.

> **Q1. What are some prerequisites before using this method?**

- We have a complete list of every member of the population.
- We can contact or access each member of the population if they are selected.
- We have the time and resources to collect data from the necessary sample size

Simple random sampling works best if we have a lot of time and resources to conduct our study,

Or if we are studying a limited population that can easily be sampled.

> **Q2. How to perform simple random sampling?**

There are 4 key steps to select a simple random sample.

> **Step 1: Define the population**

Imagine we're conducting a survey to study the eating habits of people in a particular city.

- Our defined population is all the residents of that city who are aged 18 to 60.
- Let the size of the population be 100,000 in this case

> **Step 2: Determine the Sample Size**

Now, we need to decide on the size of our sample.

Larger samples enhance statistical confidence but they also come with increased costs and effort.

- The sample size is based on the factors like,
  - The city's population size (let say, 100,000),
  - How sure do we want to be in our results?
    - Let's say we're aiming for 95% **certainty**.
  - How precise do we need to be?
    - Let's say we're cool with a 5% **margin of error**.
  - An estimated standard deviation.
- After crunching those numbers, it turns out we'll need to survey about 5000 people.

> **Step 3: Randomly Select our Sample**

We have two options for random selection:

1. **The lottery method**

In the lottery method, think of it like picking names from a wheel. Imagine we have a wheel/bowl with everyone's name in the group written on separate pieces of paper.

To choose a random sample, mix the papers well and roll the wheel, and get a few names.

It's like randomly selecting people without any specific order, just like how a lottery randomly picks winners.

2. **The random number method.**

We assign a unique number to each potential participant.

Using a random number generator, we then select 500 numbers at random from the list of assigned numbers.

> **Step 4: Collect Data from our Sample**

We proceed to collect data from the 500 individuals in our sample.

Through this range of methods, we can understand the eating habits of city residents efficiently and reliably.

Standard Error

The standard error (SE) quantifies this variability, indicating how much the sample mean is expected to deviate from the population mean when different samples are drawn.

When we want to assess the accuracy/reliability of the estimates, the standard error helps us assess the reliability of sample-based estimates.

- A **smaller standard error** suggests that the sample statistic is **likely to be close to the population parameter**
- While a **larger standard error** indicates **more variability** and less precision in the estimate.

Standard Error Formula:

The standard error of the mean is calculated using the standard deviation and the sample size.

When the population standard deviation is known

we can use it in the below formula to calculate standard error precisely.

The standard error of an estimate can be calculated as the standard deviation divided by the square root of the sample size:

$\Large SE = \frac{σ}{\sqrt n}$

where:

$σ$ = The population standard deviation

$√n$ =The square root of the sample size

When the population standard deviation is unknown

Here we can use sample standard deviation as a point estimate for the population standard deviation.

$\Large SE = \frac{s}{\sqrt n}$$\Large SE = \frac{s}{\sqrt n}$

where:

$s$$s$ = The sample standard deviation

$√n$$√n$ =The square root of the sample size

We can observe from the formula that SE is inversely proportional to the sample size.

- The Larger the sample size, the Samller the sample error will be. As the sample size grows, the sample statistic will approach the actual value of the population

> **Q. What is sample mean distributions?**

When we **collect samples** from the population (let say **5 samples**), calculate **it's mean and iterate this process numerous times** then it'll **form a distribution of sample means**.

- This is also known as **sampling distributions**.

Let's take an example:

Example

- Let's say there are 300 million people in the USA. To determine this population's average age, the statistician takes a sample of 1000 people.
- He determined the **average age of the sample was 37.5 years**.
    - The **actual average age** of the entire population is **36.9** which is different from the determined average age of sample.

The standard error is an estimate of the accuracy of the sample average.

Here,

If the statistician had taken the sample of 5000 people instead of 1000, the standard error would have been smaller and the average age of the sample would have been closer to the actual population's average age.

Law of Large Numbers

As we saw if the statistician would have taken the sample of 5000 people instead of 1000, the sample mean age would have been closer to the true population average.

This concept is known as Law of Large Numbers.

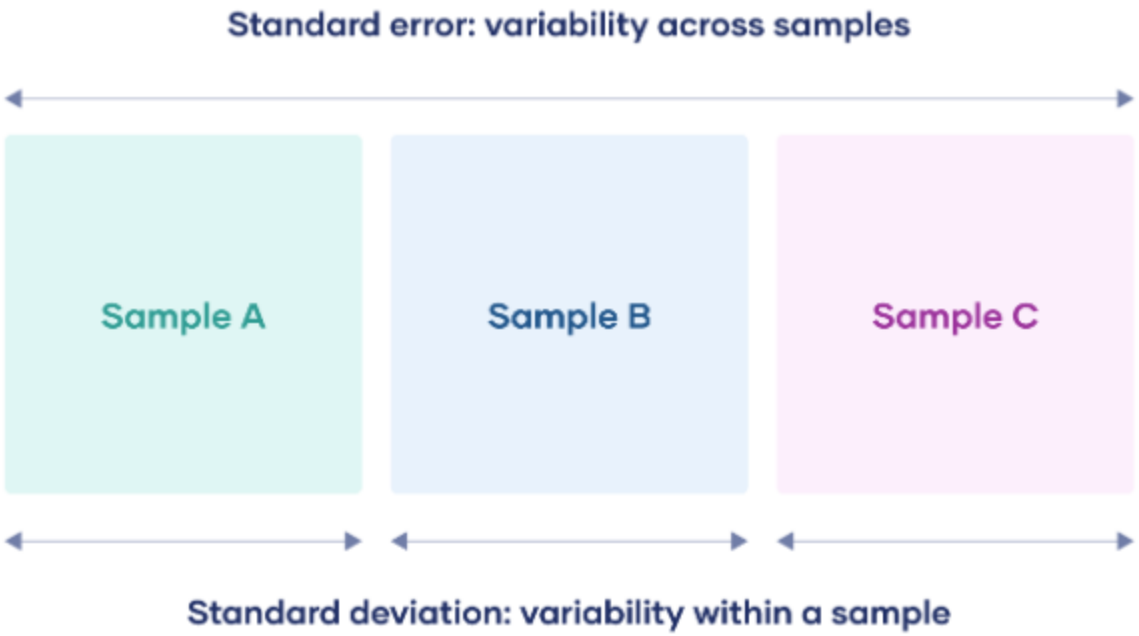
The Law of Large Numbers states that as the **sample size of a random experiment increases, the average value of the outcomes will converge to the expected value or true probability.**

In simple terms:

- It states that **as the sample size increases, the sample mean gets closer and closer to the population mean**.

Difference between standard deviation and standard error.

- Standard deviation measures the variability of individual data points within a single sample or a population,
- Standard error measures the variability of sample means when multiple samples are drawn from a population.



Uniform Distribution

A Uniform Distribution is a probability distribution where all possible outcomes are equally likely to occur.

In this distribution, each value within a specified range has the same probability of occurring.

Example

- A deck of cards has within it uniform distributions because the likelihood of drawing a heart, a club, a diamond, or a spade is equally likely.
- A coin also has a uniform distribution because the probability of getting either heads or tails in a coin toss is the same.

Types of uniform distribution

1. Discrete Uniform Distributions:
    - The possible results of rolling a die provide an example of a discrete uniform distribution
    - In discrete uniform distribution: **$P(x) = 1/n$$P(x) = 1/n$**.

      Where, P(x) = Probability of a discrete variable, n = Number of values in the range
        - It is possible to roll a 1, 2, 3, 4, 5, or 6, but it is not possible to roll a 2.3, 4.7, or 5.5.
        - Therefore, the roll of a die generates a discrete distribution with p = 1/6 for each outcome.
2. Continuous Uniform Distributions:
    - A random number generator would be considered a continuous uniform distribution.
        - Suppose we generate random numbers between 0 .0and 1.0,
        - With this type of distribution, every point in the continuous range has an equal opportunity of appearing, yet there is an infinite number of points between 0.0 and 1.0

1) The Distribution function of discrete uniform distribution(PMF)

In a discrete uniform distribution, the probability is calculated using the probability mass function (PMF)

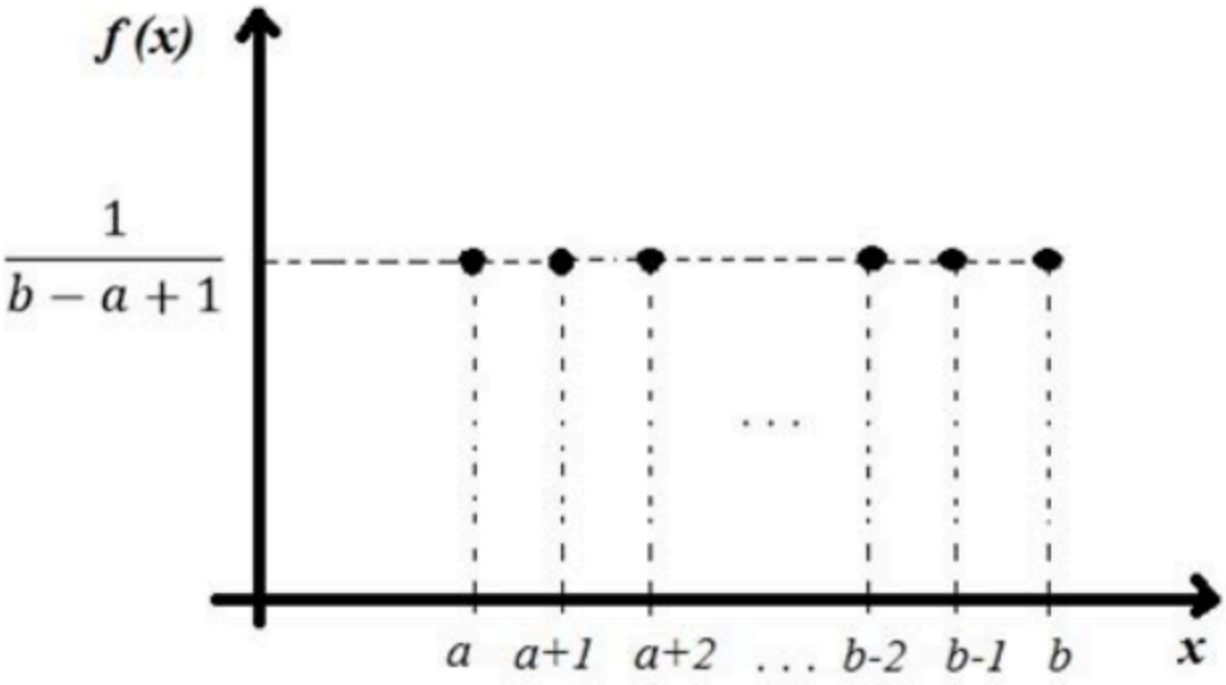
To calculate the probability of a specific value in a discrete uniform distribution:

- PMF(x) = **P(X = x) = 1 / (b - a + 1)** for a ≤ x ≤ b

  PMF(x) = **P(X = x) = 0** otherwise

Where:

- PMF(x) is the probability mass function, representing the probability of a random variable having a specific value within the range [a, b].
- "a" and "b" are the minimum and maximum values within the range.



2) The Distribution function of continuous uniform distribution(PDF)

In a continuous uniform distribution, the probability is calculated using probability density function (PDF) which is a constant value within a given range, and it's defined as:

- PDF(x) = 1 / (b - a) for a ≤ x ≤ b

  PDF(x) = 0 otherwise

Where:

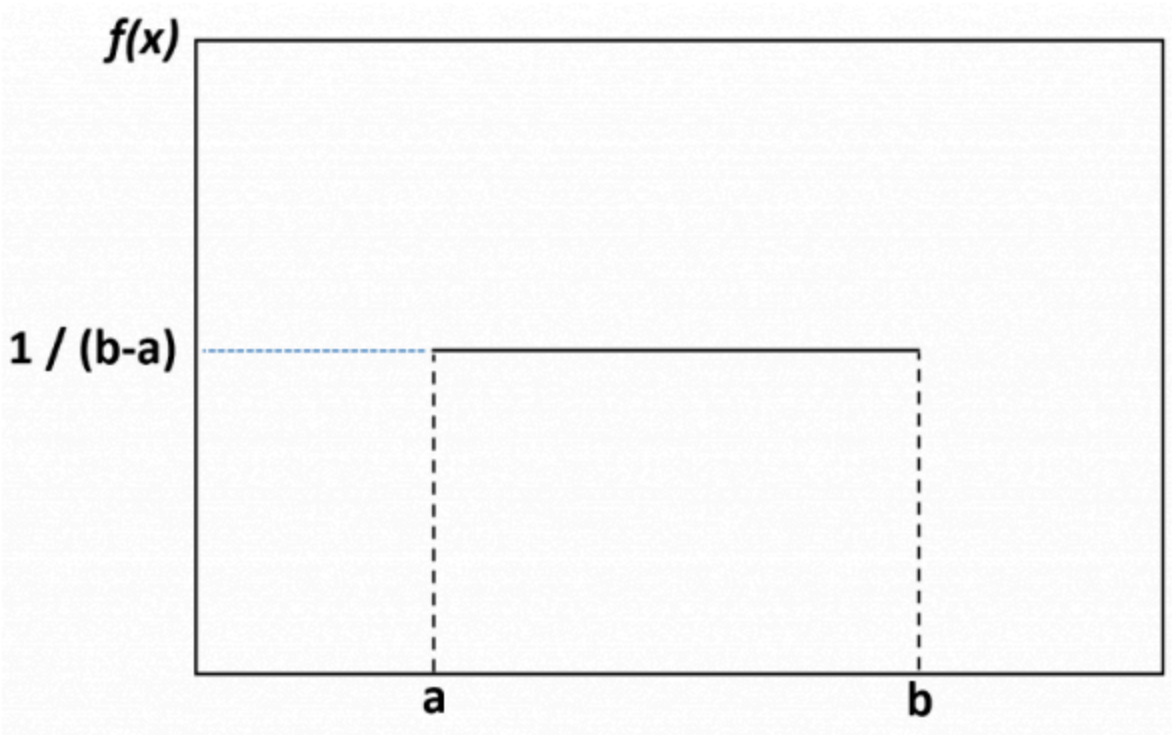
- PDF(x) is the probability density function, representing the probability of a random variable falling within the range [a, b].
- "a" and "b" are the minimum and maximum values within the range.

- Every value between "a" and "b" is equally likely to occur and any value outside of those bounds has a probability of zero.

It is also known as height of the graph

If a random variable X follows a uniform distribution, then the probability that X takes on a value between x1 and x2 can be found by the following formula:

$P(x_1 < X < x_2) = \Large \frac{(x_2 – x_1)}{(b – a)}$$$P(x_1 < X < x_2) = \Large \frac{(x_2 – x_1)}{(b – a)}$



Properties of the Uniform Distribution

The uniform distribution has the following properties:

- Mean: $\Large \frac{(a + b)}{2}$$\Large \frac{(a + b)}{2}$
- Variance: $\Large \frac{(b – a)^2}{12}$$\Large \frac{(b – a)^2}{12}$, for Continuous uniform distributions
- Variance: $\Large \frac{(b – a+1)^2 -1}{12}$$\Large \frac{(b – a+1)^2 -1}{12}$, for Discrete