Inferential Statistics

Inferential statistics involves drawing conclusions or making inferences about a population based on data collected from a sample of that population.

Important point of inferences:

a. Population: full dataset
b. sample : part of dataset
c. Sampling : taking random data from dataset by apply some techniques
d. Parameter : It is a measure that could be mean, median, variance, and many more for population data.
e. **Statistic:** It is a measure that could be mean, median, variance, and many more for sample data.

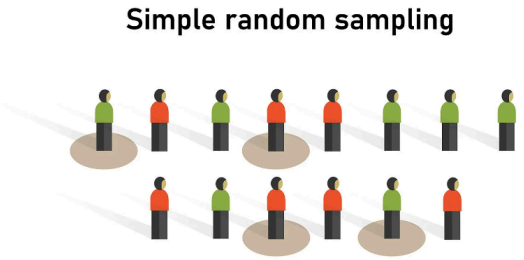
Key Points:

1. Sampling
2. Central limit theorem
3. Estimation
4. Hypothesis and HypothesisTesting

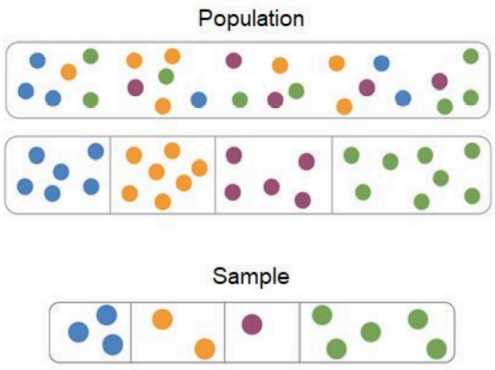
1.Sampling

Sampling is the process of selecting a sub-group of data points from the population based on a certain logic. This logic is provided by the type of technique used.

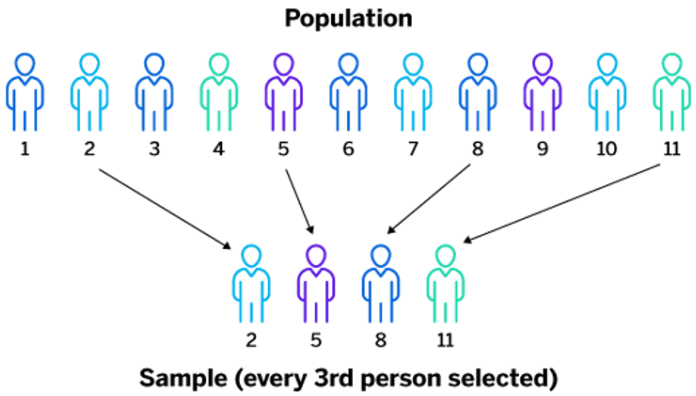
a) Simple Random Sampling: In this type, each of the data points in the population has an equal probability of getting picked up in the sample. It has two methods to do it, namely **sampling with replacement**(the data point taken in as the first is put back in the sample space before choosing the next data point) and **sampling without replacement,** which is the converse of the first.



b) Stratified Sampling: Here, the sample data points are selected based on “strata” or commonality. We use “group by” to partition the common data points.



c) Systematic Sampling: The first data point is selected randomly and the next one is selected at random intervals.



2.Central Limit Theorem

The **central limit theorem,** which is a statistical theory, states that when a large sample size has a finite variance, the samples will be normally distributed, and the mean of samples will be approximately equal to the mean of the whole population.

Central Limit Theorem Statement

The central limit theorem states that whenever a random sample of size n is taken from any distribution with mean and variance, then the sample mean will be approximately a normal distribution with mean and variance. The larger the value of the sample size, the better the approximation of the normal.

$$Z_n = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{X_1 + X_2 + \ldots + X_n - n\mu}{\sqrt{n}\sigma}$$

3.Estimation

Estimation in statistics involves using sample data to make educated guesses about a population's characteristics, such as mean, variance, or proportion. The population refers to the entire interest group, like all people in a country or all products made by a company.

Purpose of Estimation in Statistics

Statistical estimation is essential for making inferences about populations using sample data, helping to determine parameters like mean and variance without individual measurements.

- This evaluation is vital for decision-making in business and healthcare, informing strategies and treatment options.
- It is closely linked to hypothesis testing, contributing to scientific development, political decisions, public health, and economic choices.
- Risk assessment benefits from evaluation in managing probabilities and risk in finance and insurance.
- Quality control also relies on evaluation to ensure products and services meet standards by identifying and correcting deviations.

Types of Estimation
Estimation is of two types that includes:

1. Point Estimation
2. Interval Estimation

Point Estimation
In statistics, the sample mean is used to estimate a population mean, while the sample proportion is used to estimate a population percentage. These measurements help approximate unknown population parameters accurately.

- Identifying a single number to represent a large group is like a point estimate. For instance, measuring the heights of random people can be used to estimate the average height of the entire group

Interval Estimation
Point estimates provide a single value, while interval estimates give a range likely to contain the true parameter. This method recognizes data variability and estimation uncertainty.

When estimating the number of jelly beans in a jar, it is better to provide a range, known as a confidence interval, rather than a single guess. This range, such as 80 to 120 jelly beans, allows for uncertainty in the estimate and acknowledges the margin of error.

Confidence intervals give us a sense of freedom in our estimations, while point estimates only provide a single number without considering this uncertainty.
It helps us understand that there is some level of uncertainty in the estimation process.

4.Hypothesis and HypothesisTesting

What is a Hypothesis?
Hypothesis in statistics is any testable claim or assumption about the parameter of the population. It should be capable of being tested, either by experiment or observation. Example- The new engine developed by R & D gives more mileage than the existing engine.

Type of Hypothesis:
a) Null Hypothesis(H0): In the type, we say that there is no variation in the outcome. That means, there is no real effect.
Examples :

- Special training on students does not affect.
- Different teaching method does not affect students' performance
- The drug used for headaches does not affect the application.

b) Alternate Hypothesis(Ha): It is the contrasting statement to H0 where it says there is a real effect in the outcome. This is the statement we are trying to prove.
Examples:

- Special training on students has a significant effect.
- Different teaching methods have a significant effect on students' performance.
- The drug used for headaches has a significant effect after application.

Hypothesis Testing Process:
As we already know, a hypothesis is a testable claim and only either H0 or Ha can be proved. The process of proving either of them is called the "**Hypothesis Testing Process". Note that if we accept H0, automatically Ha is rejected and vice-versa.**

Claim or Assumption	Hypothesis Statements
Bank Claims : Average waiting time at its ATM is 15 minutes	H0: Avg. waiting time at a bank ATM = 15 minutes Ha: Avg. waiting time at a bank ATM ! = 15 minutes
Tata research claims: Nano model gives mileage better than 24 kmpl.	H0: Avg. mileage for nano <= 24 kmpl Ha: Avg. mileage for nano > 24 kmpl
Company claims LED bulbs have a life of less than 18000 hours.	H0: Avg. life of LED bulbs >= 18000 hours Ha: Avg. life of LED bulbs < 18000 hours

Hypothesis Testing:
After framing the hypothesis statements H0 and Ha for a given claim, it is now time to prove either of them wrong. This is done by 3 well-defined methods namely,
a) Critical value approach
b) p-value approach
c) Confidence interval approach

a) Critical value approach:
Steps involved:

Critical Value Approach	
Step 1	State H0 and Ha
Step 2	Define "alpha". Compute the test statistic(test_stat=(xbar-mu_population)/Standard Error))
Step 3	Compute the critical value
Step 4	Compare test and critical value. If test_stat > critical for positive values of test_stat and critical value, "Reject H0". If test_stat > critical for negative values of test_stat and critical value, "Accept H0"

To compute critical values, the kind of test that we observe from the problem statement is very important.
For a left tailed test, the "test_stat" and the "critical" values will lie on the left of the mean of the normal curve. Hence, their values will be **negative**. Then,

critical= scipy.stats.norm.ppf(α***)*** *using Z-distribution for "σ "(known)*
critical= scipy.stats.t.ppf(α***,n-1)*** *using T-distribution for "σ "(unknown)*

For a right-tailed test, the “test_stat” and the “critical” values will lie on the right of the mean of the normal curve. Hence, their values will be positive. Then,

critical= scipy.stats.norm.isf(α) *using Z-distribution for “σ “(known)*
critical= scipy.stats.t.isf(α,n-1) *using T-distribution for “σ “(unknown)*

b) p-value approach:
Steps involved:

p-value approach	
Step 1	State H0 and Ha
Step 2	Define "alpha". Compute the test statistic(test_stat=(xbar-mu_population)/Standard Error))
Step 3	Compute the p-value
Step 4	Compare p-value and the alpha value If p-value < alpha."Reject H0".

To compute a p-value, the kind of test that we observe from the problem statement is very important.
For a left-tailed test, the “test_stat” and the “critical” values will lie on the left of the mean of the normal curve. Hence, their values will be **negative**. Then,

p_value= scipy.stats.norm.cdf(test_stat***)*** *using Z-distribution for “σ “(known)*
p_value= scipy.stats.t.cdf(test_stat***,n-1)*** *using T-distribution for “σ “(unknown)*

For a right-tailed test, the “test_stat” and the “critical” values will lie on the right of the mean of the normal curve. Hence, their values will be **positive**. Then,

p_value= scipy.stats.norm.sf(test_stat***)*** *using Z-distribution for “σ “(known)*
p_value= scipy.stats.t.sf(test_stat***,n-1)*** *using T-distribution for “σ “(unknown)*

c) Confidence Interval approach:
Steps involved:

Confidence Interval approach	
Step 1	State H0 and Ha
Step 2	Define "alpha". Compute the test statistic(test_stat=(xbar-mu_population)/Standard Error))
Step 3	Compute the population parameter confidence interval
Step 4	If the the mu_population lies in the interval, then accept H0. Else, reject H0.