INFERENTIAL STATISTICS

---------- DAY-1 ---------

DEFINITION:

Inferential statistics is the **practice of using sampled data** to draw conclusions or make predictions **about a larger sample data sample** or population.

EXAMPLE:

Amazon's -- Walmart's Quality Control department.

It wants to know how much company's products in its warehouses defective.

Inferential Statistics is used in the industry in multiple ways like:

1.Healthcare : Clinical Trials

It is used in clinical trials to analyse the effectiveness of new drugs or treatments by drawing conclusions about the entire patient population based on a sample.

2.Finance: Risk Assessment

Financial Institutions use inferential statistics to assess and manage risks. This includes predicting market trends, estimating the likelihood of default on loans, and analysing investment portfolios. (ex. insurance loan...)

3.Marketing: Consumer Surveys

In marketing, inferential statistics are employed to make inferences about the preferences and behaviours of a target market based on survey data, helping businesses make informed decisions about product development and advertising strategies.

4.Manufacturing: Quality Control

It is used in quality control processes to make inferences about the quality of products based on a sample of items, helping manufacturers maintain consistent product quality.

5.Education: Standardized Testing

In education, Inferential statistics are used to draw conclusions about the performance of a larger population of students based on the results of standardized tests taken by a representative sample.

6.Environmental Science: Pollution Monitoring

It helps environmental scientists estimate the level of pollution in a region by analysing sample of air, water or soil, allowing for inferences about the overall environmental health.

7.Human Resources: Employee Satisfaction

HR professionals use inferential statistics to make inference about the overall job satisfaction of employees based on survey data, helping organizations identify areas for improvement.

8.Retail: Demand Forecasting

In retail, inferential statistics are applied to analyse past sales data and make predictions about future demand for products, optimizing inventory management and supply chain logistics.

9.Telecommunications: Network Performance

Telecom companies use inferential statistics to assess network performance by analysing data from a sample of users, helping them make inferences about the quality and reliability of their services for the entire user base.

10.Government: Census Data Analysis

Governments use inferential statistics when analysing census data. By studying a sample of the population, they can make inferences about demographic trends, socioeconomic indicators, and other important factors that inform public policy decisions.

****

PROBABILITY:

Event: is the result of an observation or experiment. Event is a subset of sample space.

(Ex: getting head or tail, if we toss the coin)

Types of Events:

i) Independent Event:

Two or more events are independent if the 'occurrence of one does not affect the occurrence of the other.'

(Ex: If we toss the coins twice, the result of the second throw is not affected by the first result.)

ii) Complementary events:

Not happened events.

(Ex: Take two events, let it be A and B then P(A)+P(B)=1)

iii) Equally likely Events:

Let A and B be the two events and P(A)=P(B), then they are equally likely event.

(Ex: in throwing a dice, all the 6 faces (1,2,3,4,5,6) are equally likely occurring.)

iv) Mutually Exclusive events:

If A and B have no common outcome, then it is said to be mutually exclusive events.

(Ex: S={1,2,3,4,5,6}, A=odd numbers, B=Even numbers then A and B have no common outcome.)

v) Collective Exhaustive events:

It gives Complete sample space.

(Ex: that is AUB=S.)

Sample Space:

A collection of all the possible outcomes of a trial or experiment.

(When tossing a coin: {H, T}, tossing 2 coins: {HH, TT, HT, TH})

Interpreting a Probability:

Tossing a coin twice:

P(HT) = 1/4 or 25% or 0.25

The probability of any event going to be 0<=P(any event)<=1

0 means it is impossible to happen.

1 means, that is definitely going to happen.

****

PROBABILITY DISTRIBUTION: (See Infer-class1.pdf)

It lists all the possible outcomes in the sample space and the probabilities with which they can occur.

Example: throwing a die

we get event: 1, 2, 3, 4, 5, 6

Probability: 1/6, 1/6, 1/6, 1/6, 1/6, 1/6 ---- it is called uniform distribution.

(uniform distribution- every event is equally likely to happen)

TYPES: 1) Discrete Probability Distribution,

2) Continuous Probability Distribution

NOTE: Random Variables:

Random Variables are variables that represent the outcomes of a random experiment. For example, the collection of outcomes of a series of coin tosses is a random variable. Here the possible set of outcomes is just two-Head & Tails. If we map Heads to the number 1 and Tails to 0. Then the random variable could look something like (1, 1, 0, 1, 0, 0, 1, 0) for eight coin flips.

The values of a random variable can change the next time it is recorded, but they can only contain a specific set of values.

A random variable is denoted with a capital letter (typically X, Y, Z, etc.,) and specific values are denoted with lowercase letters (eg., X=x or X<=x).

Ex: Tossing two coins together

- X=0 if both tosses result in no heads. P(X=0) = 1/4
- X=1 if one of the tosses results in heads. P(X=1) = 2/4 = 1/2
- X=2 if both tosses result in heads. P(X=2) = 1/4

(X= probability of getting heads, ¼, ½, are random variables)

Random Variables are of two types:

1. Discrete RV: They take a fixed set of possible outcomes. Each outcome has an associated probability. Ex: Number of heads in two tosses, The creditworthiness of a loan applicant, Marital status, Gender, etc., (It can predict, it gives finite possible outcomes)
2. Continuous RV: They can take any value within a range. Ex: Age of a person, Income of a person, Subscription of any platform like Netflix, Disney, Hotstar, etc., (not predict, infinite possible outcomes)

DISCRETE PROBABILITY DISTRIBUTION:

These distributions model the probabilities of random variables (A discrete random variable is a variable that can only take on specific, distinct values. These **values are typically finite or countable** and often represent the outcomes of a random event.) that can have discrete values as outcomes.

Example:

The possible values for the random variable X that represents the number of heads that can occur when a coin is tossed twice are the set {0, 1, 2} and not any value from 0 to 2 like 0.1 or 0.6.

Examples: Bernoulli, Binomial, Negative Binomial, Hypergeometric, Geometric distribution, Poisson distribution, multinomial distribution.

TYPES:

1) Bernoulli Distribution:

This distribution is generated when we **perform an experiment once** and it **has only two possible outcomes** – success and failure. The trials of this type are called Bernoulli trials, which form the basis for many distribution discussed below. Let p be the probability of success and 1-p is the probability of failure.

The PMF is given as

$$\text{PMS} = \begin{cases} P, & Success \\ 1-p, & Failure \end{cases}$$

One example of this would be flipping a coin once. p is the probability of getting a head and 1-p is the probability of getting a tail. Please note down that success and failure are subjective and are defined by us depending on the context.

2) Binomial distribution:

(When we **repeat Bernoulli trial for n times**, it is called Binomial distribution)

This is generated for random variables with only two possible outcomes. Let p denote the probability of an event is a success which implies 1-p is the

probability of the event being a failure. Performing the experiment repeatedly and plotting the probability each time gives us the Binomial distribution.

The most common example given for Binomial distribution is that of flipping a coin n number of times and calculating the probabilities of getting a particular number of heads. More real-world examples include the number of successful sales calls for a company or whether a drug works for a disease or not.

The PMF is given as,

$n_{C_x} p^x (1-p)^x$

3) Poisson distribution:

This distribution describes the **events that occur in a fixed interval of time or space.** An example might make this clear. Consider the case of the number of calls received by a customer care centre per hour. We can estimate the average number of calls per hour but we cannot determine the exact number and the exact time at which there is a call. Each occurrence of an event is independent of the other occurrences.

The PMF is given as, (Probability Mass Functions - PMF)

$P(X=x) = \frac{e^{-\lambda}\lambda^x}{x!}$

Where lambda is the average number of times the event has occurred in a certain period of time, x is the desired outcome and e is the Euler's number.

4) Multinomial Distribution:

(**Repeat event multiple times, and it gives more than 2 possible outcomes.**)

In the Binomial distribution, there are only two possible outcomes – success and failure. The multinomial distribution, however, describes the random variables with many possible outcomes. This is also sometimes referred to as categorical distribution as each possible outcome is treated as a separate category. Consider the scenario of playing a game n number of times. Multinomial distribution helps us to determine the combined probability that player 1 will win x_1 times, player 2 will win x_2 times and player k wins x_k times.

The PMF is given as,

$$P(X=x_1, X=x_2,\ldots X=x_k) = \frac{n!}{x_1!x_2!\ldots.x_k!} P_1^{x_1} P_2^{x_2} \ldots P_k^{x_k}$$

Where n is the number of trials, p1, … ,pk denote the probability of the outcomes x1,…xk respectively.

CONTINUOUS PROBABILITY DISTRIBUTION:

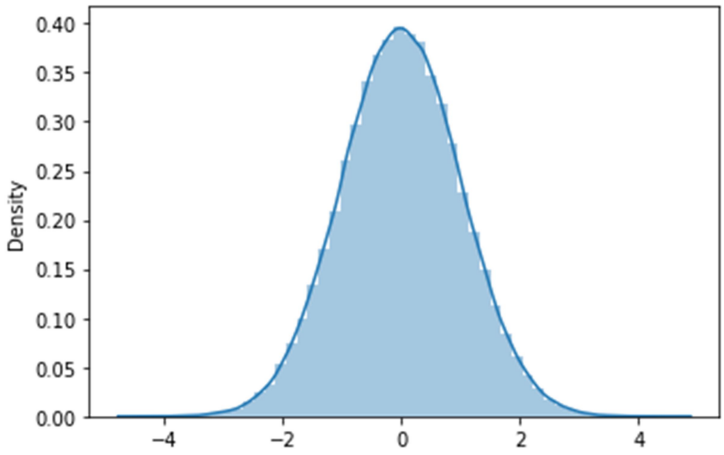
These distributions model the probabilities of random variables that can have any possible outcome. For example, the possible values for the random variable X that represents weight of citizens in a town which can have any value like 34.5, 47.7, etc..

Examples: Normal, Student's T, Chi-square, Exponential, salary, temperature, age,...

1) Normal Distribution:

This is the most commonly discussed distribution and most often found in the real world. Many continuous distributions often reach normal distribution given a large enough sample. This has two parameters namely mean and standard deviation.

This distribution has many interesting properties. The **mean has the highest probability and all other values are distributed equally on either side of the mean in a symmetric fashion**. The standard normal distribution is a special case where the mean is 0 and the standard deviation is 1.



It also follows the empirical formula that 68% of the values are 1 standard deviation away, 95% percent of them are 2 standard deviations away, and 99.7% are 3 standard deviations away from the mean. This property is greatly useful when designing hypothesis tests. (https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/)

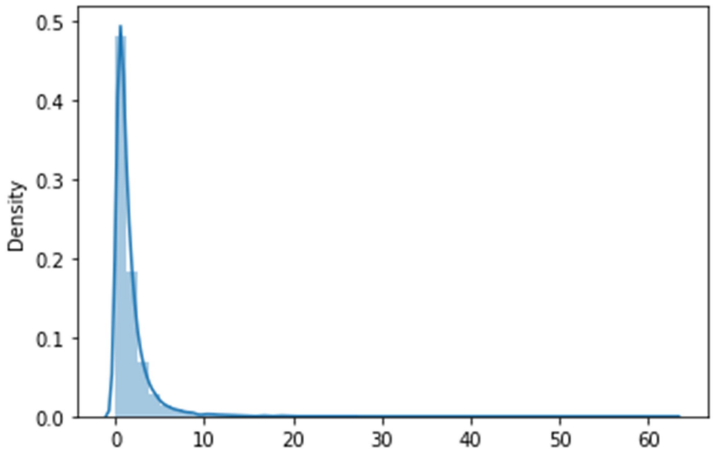
The PDF is given by,

$$f(\mathrm{x}) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

where μ is the mean of the random variable X and σ is the standard deviation.

2) Log-normal Distribution:

This distribution is used to plot the random variables whose logarithm values follow a normal distribution. Consider the random variables X and Y. Y = ln(X) is the variable that is represented in this distribution, where ln denotes the natural logarithm of values of X.



The PDF is given by,

$$f(\mathrm{x}) = \frac{1}{x\sqrt{2\pi}\sigma} exp\left(-\frac{1}{2}\left(\frac{lnx-\mu}{\sigma}\right)^2\right)$$

where μ is the mean of Y and σ is the standard deviation of Y.

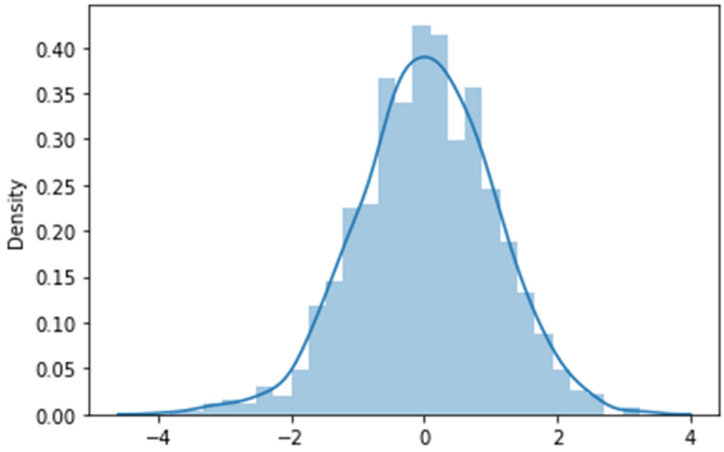
('ln' basically refers to a logarithm to the base e.)

3) Student's T Distribution:

The student's t distribution is similar to the normal distribution. The difference is that the tails of the distribution are thicker. This is used when the sample size is small and the population variance (or standard deviation) is not known. This distribution is defined by the degrees of freedom (p) which is calculated as the sample size minus 1(n – 1).

As the sample size increases, degrees of freedom increases the t-distribution approaches the normal distribution and the tails become narrower and the

curve gets closer to the mean. **This distribution is used to test estimates of the population mean when the sample size is less than 30** and population variance is unknown. The sample variance/standard deviation is used to calculate the t-value.



The PDF is given by,

$$f(t) = \frac{\Gamma\left(\frac{\mathrm{p}+1}{2}\right)}{\sqrt{\mathrm{p}\pi}\,\Gamma\left(\frac{\mathrm{p}}{2}\right)}\left(1+\frac{t^2}{\mathrm{p}}\right)^{-\left(\frac{p+1}{2}\right)}$$

where p is the degrees of freedom and Γ is the gamma function. Check this link for a brief description of the gamma function.

The t-statistic used in hypothesis testing is calculated as follows,

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

where $\bar{x}$ is the sample mean, μ the population mean and s is the sample variance.

NOTE:

Population and sample, not exactly same. We have some + or - margin of error involved in.

Ex: if margin of error + or – 2k,

income of family 25k ---- this exact value is point estimate.

Then range between 23k to 27k ----- it is confidence interval

Highlight: Sample mean + or – margin of error = confidence interval.

Types of Sampling techniques:

1) Random Sampling: Is the most common sampling technique. We take samples randomly. Most of the time the chances of error is very high because of we select the samples or data randomly.

2) Stratified Sampling: Most advance and structured technique.

In this technique, we first split samples into groups or sections (split Disproportionate Sampling or Proportionate Sampling manner) and then we randomly select the samples.

Note: each samples in group have similar characteristics.

Note:

Disproportionate Sampling: we take equal amount of samples.

Ex: Split population according to Classes (rich, middle, poor).

Only take 1000 family, then take rich in 333 families and middle in 333 families and poor in 334 families. Totally we take 1000 families. Ie., we take equal number of families for sampling.

Proportionate Sampling: Taking samples proportionally.

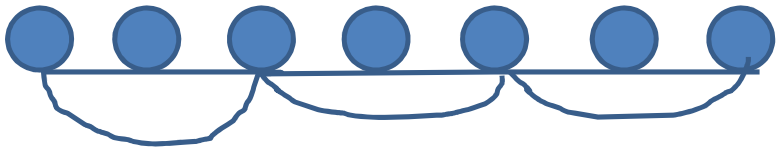
If rich:middle:poor = 10%:60%:30%

Take 1000 families, rich is 100 families, middle is 600 families and poor is 300 families for sampling.

3) Systematic Sampling:

This one is specially used to take samples from dynamic / continuously changing data. Data regularly updating.

Ex: k = 2, k is the any sample interval.



4) Cluster Sampling:

Ex: We take 20 samples. Divide it in 5 groups and each group have 4 samples (not necessary to each sample have similar characteristic). Finally we choose one or some group of samples for sampling, and other groups of samples are ignored.

****

---------DAY-2---------

CENTRAL LIMIT THEOREM:

Hints:

- Take the dataset (ex: population)
- Make multiple samples. Each sample > 30
- Take the mean of all samples.
- Plot and check the distribution of the data.
- It will always be a 'Normal Distribution'.

We have to get population mean using sample mean:

Population mean = Sample mean $\pm$ Margin of error

= Confidence interval

(If interval range is increasing, confidence percentage might also increase. Margin of error decides the range increasing or decreasing)

Population mean Formula:

Population Mean = $\bar{x} \pm z^* \times \frac{s}{\sqrt{n}}$

$\bar{x}$: Sample mean

z^* : z – score for a certain confidence level (z^* -- Only used for normal distribution)

s : Standard deviation of the sample

n : No. of data points in the sample

z^*	Confidence Level
1.65	90%
1.96	95%
2.58	99%

Example: Estimate whether mean lead content in Maggi is within the allowed range or not. (Note: ppm - Parts Per Million)

Allowed = 2.5ppm, n = 100, $\bar{x}$ = 2.3ppm, s = 0.3ppm,

Find confidence interval when confidence level is 99% (because it is dangerous to life)

Confidence interval = $\bar{x} \pm z^* \times \frac{s}{\sqrt{n}}$

$= 2.3 \pm 2.58 \times \frac{0.3}{\sqrt{100}}$

$= 2.3 \pm 0.07$

= 2.23 to 2.37

CLT used in Real World:

1) **Quality Control in Manufacturing:**

   Scenario: A manufacturing plant produces a large number of products each day, and the quality control team is interested in the average weight of the products.

   Use of CLT: By collecting random samples of product weights and calculating the sample means, the quality control team can apply the CLT to assume that the distribution of sample means is approximately normal. This allows them to make statistical inferences about the average weight of all products.

2) **Financial Modelling and Risk Assessment:**

   Scenario: A financial analyst wants to assess the average return on investment (ROI) of a portfolio of stocks.

   Use of CLT: BY taking multiple random samples of historical ROI data and calculating sample means, the analyst can apply the CLT. This enables them to make more reliable predictions about the average ROI of the entire portfolio.

3) **Marketing and A/B Testing:**

   Scenario: A marketing team is running an A/B test to compare the effectiveness of two different versions of an advertisement.

Use of CLT: By collecting random samples of user responses to each version and calculating sample means, the marketing team can use the CLT to make statistical inferences about the average effectiveness of each advertisement version for the entire target audience.

4) **Healthcare and Clinical Trials:**

Scenario: In a clinical trial for a new drug, researchers want to estimate the average reduction in symptoms.

Use of CLT: By repeatedly collecting random samples of patient data and calculating sample means, researchers can apply the CLT. This allows them to make inferences about the average impact of the drug on the entire population of interest.

5) **E-Commerce and Customer Behavior:**

Scenario: An e-commerce platform wants to understand the average time spent by customers on their website.

Use of CLT: By taking random samples of user engagement data and calculating sample means, the data science team can leverage the CLT to make statistically valid predictions about the average time spent on the website for all users.

****

HYPOTHESIS TESTING

Difference between Inferential Statistics and Hypothesis Testing:

Let's understand the basic difference between inferential statistics and hypothesis testing.

Inferential statistics is used to find some population parameter (mostly population mean) when you have no initial number to start with. So, you start with the sampling activity and find out the sample mean. Then, you estimate the population mean from the sample mean using the confidence interval.

Hypothesis testing is used to confirm your conclusion (or hypothesis) about the population parameter (which you know from EDA or your intuition). Through hypothesis testing, you can determine whether there is enough evidence to conclude if the hypothesis about the population parameter is true or not.

Hypothesis Testing starts with the formulation of these two hypotheses:

Null hypothesis (H_0): The status quo

Alternate hypothesis (H_1): The challenge to the status quo

(See Hypothesis_final.pdf)