

Probability & Statistics

Vishnu Kumar Chintala

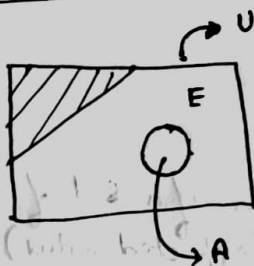
Probability

Probability measures chances of occurrence of a particular event.

$$P(A) = \frac{\text{Number of favourable outcomes}}{\text{Total number of outcomes}}$$

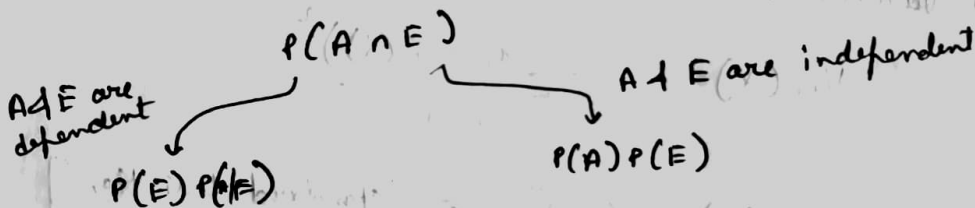
$P(A)$ = Probability of event A happening.

Conditional Probability



Let us assume a sample space U . There is a event E & Event A . Conditional Probability gives the probability occurrence of one event from the space of another event.

$$P(A|E) = \frac{P(A \cap E)}{P(E)}$$



Total Probability

The law of total gives method to count or quantify Probability of an event if we know Probability of that's events occurrence from an other event (conditional Probability) & Probability that other events.

$$P(A) = \sum_{i=1}^n P(A|B_i)P(B_i)$$

ADJ

Bayes theorem

In conditional probability we have event 1 occurring given that event 2 has happened, what if we want to quantify probability event 2 from event 1.

$$P(B/A) = \frac{P(A/B) P(B)}{P(A)} \rightarrow \left[\text{Baye's theorem formula} \right]$$

Expected Value [Mean]

Expected value represents the average outcome you would expect if you were to repeat an experiment many times. (mean)
It is also the measure of the central tendency of probability distribution.

$$E(X) = \sum_{i=1}^n x_i P(X=x_i)$$

Variance

Variance is measure of the spread or dispersion of a set of data points or a random variable from its mean (expected value).
It also indicates how widely range are data points.
Low variance means all values are close to mean.

$$\text{Var}(X) = \frac{\sum (x_i - E(X))^2}{n-1}$$

$$\text{Var}(X) = (\sigma)^2 \quad [\sigma = \text{standard deviation}]$$

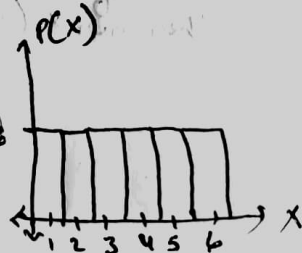
PMF (Probability Mass Function)

Probability Mass Function quantifies Probability of random discrete variables. Example rolling a dice.

$$X = \{1, 2, 3, 4, 5, 6\}$$

$$\text{PMF} = \begin{cases} \rightarrow 1/6, & x \Rightarrow X \\ \rightarrow 0 & \end{cases}$$

PMF $\Rightarrow$



$$f(x_i) \geq 0 \quad \Bigg| \quad \sum_{i=1}^n f(x_i) = 1$$

PMF

Common Distribution Function (CDF)

Cumulative

The cumulative Distribution Function (CDF) of a random variable is a function that gives the probability that the random variable takes a value less than or equal to certain value. It applies to both discrete and continuous random variables and describes the probability distribution of a variable.

$$F_x(x) = P(X \leq x) \rightarrow \text{CDF}$$

$$\Rightarrow F_x(x) = \sum_{i=1}^n P(x_i) \quad [\text{for discrete random variable}]$$

Covariance

Covariance is a measure of the joint variability of two random variables. It indicates whether the two variables tend to vary together or not. Specifically, it measures how much two random variable change together, and whether an increase in one variable is associated with an increase or decrease in the other.

Random variables = X, Y

$$\text{Cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Correlation

Correlation is a measure to quantify strength and direction of the linear relationship between two random variables. It indicates how changes in one random variable associates other.

$$r = \frac{\text{Cov}(X, Y)}{\sigma_x \cdot \sigma_y} \quad (\sigma = \text{standard deviation})$$

$\text{Cov}(X, Y) = \text{Covariance.}$



Binomial Distribution

Binomial Distribution quantifies probability of a specific number of successes in a fixed number of independent trials of binary experiment.

All trials should be independent.

$p \rightarrow$ success

$q \rightarrow$ failure

$$P(X=r) = {}^nC_r p^r q^{(n-r)}$$

$n =$ number of trials

$r =$ number of successes

$p =$ probability of success

$q =$ probability of failure

Expected Value = Mean = np

Variance = $\sigma^2 = npq$

Standard deviation = $\sigma = \sqrt{npq}$

Note:-

$p = q = \frac{1}{2} \Rightarrow$ Symmetrical Distribution.

$p \neq q \Rightarrow$ skewed Distribution.

Poisson Distribution

Poisson Distribution is a probability distribution which describes probability of a given number of events occurring in a fixed interval of time or space. Example if an event occurs k times avg what is probability of getting K times that same event.

Imp Note:-

- Discrete Distribution.
- Fixed Interval.
- Avg Rate of Occurrence.
- Independence of Events.
- Small probability of Event.

Def

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

X = random variable

k = k is the number of occurrences of event

λ = Average number of times event occurs.

e = euler number.

$$\boxed{E(X) = \mu = \sigma^2 = \lambda}$$

→ rate of occurrence or Average

Expected value Mean Variance

The Poisson Distribution is memoryless with respect to its process, meaning that the probability of an event occurring in future interval is independent of past events.

Geometric Distribution

Geometric Distribution quantifies probability of getting first success in k number of attempts of an independent event.

$$P(X=k) = (1-p)^{k-1} p \quad E(X) = \frac{1}{p}$$

p → probability of success in each trial.

k → number of trials until the first success.

$(1-p)$ → probability of failure.

Bernoulli Distribution

Simplest probability to represent probability of a binary event.

$$P(X=x) = \begin{cases} p, & x=1 \\ (1-p), & x=0 \end{cases}$$

1 → denotes success, 0 → denotes failure

Def

Continuous Random Variable (CRV)

Till now values of Random Variable in every example or explanation had Discrete values.

$$\bullet X = \begin{cases} 1, & \text{Heads} \\ 0, & \text{Tails} \end{cases} \quad [\text{Coin}]$$

$$\bullet X = \text{Sum of two dice rolled}$$

$$X \in [2, 3, 4, 5, 6, \dots, 12]$$

In these cases value X can take in unique and Discrete.

Now there can be case where values cannot bound to be discrete, like it can be in decimal and belong from a range.

$$\bullet X = \text{Height of students in class.}$$

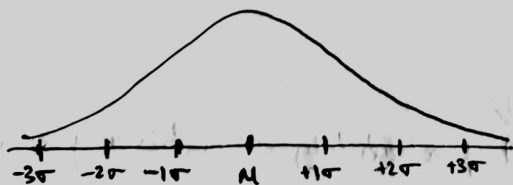
$$\bullet X = \text{Meters athlete ran in a second in world records.}$$

In these cases values can be like 75.6 kg 75.2 kg which are very non-discrete values. These kind of random values or variable comes under CRV (Continuous Random Variable).

Normal Distribution (Z-score)

$$\text{formula} \Rightarrow f(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\left[\frac{(x-\mu)^2}{2\sigma^2}\right]}$$

[PDF]



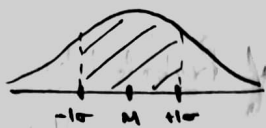
The Bell shaped curve is the representation of Normal Distribution. It is centred around mean and spreads out with decreasing probability of occurrence in either ways.

μ = mean.

σ = standard deviation.

x = value of random variable.

Handwritten signature



68% of the data lies in the range of -1σ to $+1\sigma$.

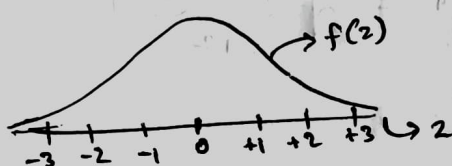


95% of the data lies in the range of -2σ to $+2\sigma$.

Similarly, 99.7% of the data lies in the range of -3σ to $+3\sigma$.

* Standardization of data

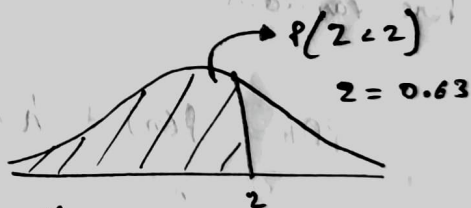
In Normal distribution it can be used to represent Probability of occurrence of many kind of data with different units & properties. So to get rid of that in distribution this technique is used.



$$z = \frac{x - \mu}{\sigma} \Rightarrow \underline{\underline{Z\text{-score}}}$$

To get the probability of a value of random variable Z-score table can be used to get probability by calculating corresponding Z-score of value of random variable.

$$P(Z < z) \rightarrow \text{Z-score}$$



Area of left is the Probability of the Z-score value in table.

Z	0.00	0.01	0.02	...
0.0				
0.1				
0.2				
...				

→ Z-Score Probability

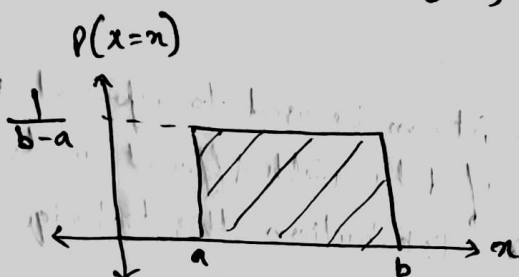
← Example of Z-score table

Def

Uniform Distribution

Uniform Distribution is a type of Probability distribution in which all outcomes are equally probable or likely. Every value in the given range has same probability of occurring.

$$\text{PDF} = f(x) = \begin{cases} 0, & x < a \\ \frac{1}{b-a}, & a \leq x \leq b \\ 0, & x > b \end{cases}$$



$$\text{CDF} = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \leq x \leq b \\ 1, & x > b \end{cases}$$

$$\text{Standard deviation} = \sigma = \sqrt{\frac{(b-a)^2}{12}}$$

$$\mu = \text{mean} = \frac{a+b}{2}$$

Exponential Distribution

Exponential Distribution is same as Poisson's Distribution but here random variable is a range (CRV).

$$\text{PDF} = f(x) = h e^{-hx}, \quad x \geq 0$$

$$\text{CDF} = \int_0^x f(x) dx = \int_0^x h e^{-hx} dx = 1 - e^{-hx}, \quad x \geq 0$$

$$E(x) = \text{mean} = \frac{1}{h}$$

$$\text{Var}(x) = \sigma^2 = \frac{1}{h^2}$$

$$\text{Standard deviation} = \frac{1}{h}$$

ADP

Important $\Rightarrow$ Probability distribution between two random variable. (Joint Probability)

Assume random variables

$\Rightarrow X, Y$

$\{0, 1\}$ $\{0, 1, 2, 3\}$

X/Y	0	1	2	3
0	m	m	m	m
1	m	m	m	m

(Probability distribution)

• Marginal Probability $\Rightarrow$

$$P(X=0) = P(X=0, Y=0) + P(X=0, Y=1) + P(X=0, Y=2) + P(X=0, Y=3)$$

• include every combination includes $X=0$.

• Conditional Probability $\Rightarrow$

$$P(X=0/Y=1) \Rightarrow \frac{P(X=0, Y=1)}{P(Y=1)}$$

• Independence of the events $\Rightarrow$

$$P(X=x, Y=y) = P(X=x) \cdot P(Y=y)$$

for all x, y .

this condition should be true for Independence of the events.

Properties of Covariance

$$\bullet \text{Cov}(aX + b, cY + d) = ac \text{Cov}(X, Y)$$

$$\bullet \text{Cov}(X_1 + X_2, Y) = \text{Cov}(X_1, Y) + \text{Cov}(X_2, Y)$$

$$\bullet \text{Cov}(X, X) = \text{Var}(X)$$

$$\bullet \text{Cov}(X, Y) = E(XY) - \mu_X \mu_Y$$

• X and Y are independent $\text{Cov}(X, Y) = 0$ (Converse is not true)

ADG

Sampling Distribution

A Sampling Distribution is the distribution of population data of size (n) with samples of a fixed size (n) .

These samples statistics form their own distribution.

That distribution will have mean μ_s & standard deviation σ_s . CLT makes the derived sample distribution centred along μ_s .



μ_p = Mean of Population data

σ_p = Standard Deviation of Population data

μ_s = Mean of Sample data

σ_s = Stand Deviation of Sample data

n = Sample Size

$$\sigma_s < \sigma_p$$

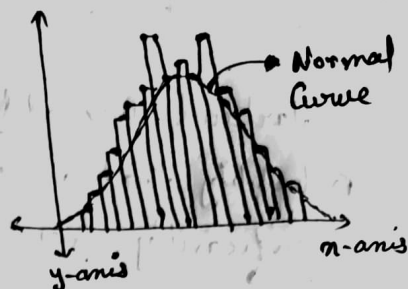
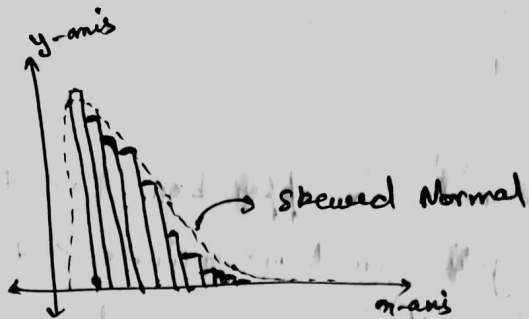
$$\mu_s = \mu_p$$

$$\sigma_s = \frac{\sigma_p}{\sqrt{n}}$$

Central Limit Theorem

The Central Limit Theorem says that if we take many random samples of a large enough size (n) from any population, the averages (sample means) of those samples will form a distribution that looks like a normal bell-shaped curve, even if the population itself isn't normal. The larger the sample more it gets normal.

Handwritten signature/initials.



Confidence Interval Estimation

• Z-Method

Interval for confidence level of $(1-\alpha) \times 100\%$

$$\text{Confidence Interval} = \left(\bar{n} \pm Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} \right)$$

$\bar{n}$ = Sample Mean

$Z_{\alpha/2}$ = Z Critical Value

σ = Standard deviation of Population

n = Sample Size

α = Significance level

• T-Method

Interval for confidence level of $(1-\alpha) \times 100\%$

$$\text{Confidence Interval} = \left(\bar{n} \pm T_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \right)$$

$T_{\alpha/2}$ = T-critical value for the given confidence level of degree of freedom $(n-1)$.

s = Sample standard deviation

Z-Method for $n \geq 30$, T-Method for $n < 30$.

Handwritten signature

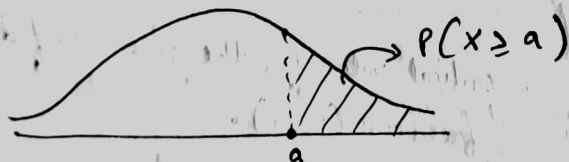
Markov's Inequality

Markov's Inequality provides an upper bound on the probability that a non-negative random variable, X , deviates significantly from mean. It is a very general inequality that applies to any random variable with a finite expected value, regardless of its distribution. For example, if you know the average outcome of a process, Markov's Inequality can estimate the probability of outcome being much larger than the average, even if you lack detailed info about process.

$$P(X \geq a) \leq \frac{M}{a}, \text{ where } a > 0.$$

- $P(X \geq a)$ Probability of X atleast a .
- $M = E[X]$ Expected value of X
- $a = \text{threshold value.}$

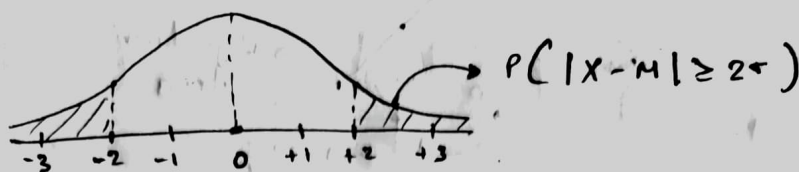
Used to find upper bounds for probability when only the mean of X is known.



Chebyshev's Inequality

Chebyshev's Inequality bounds the probability of a random variable deviating from its mean by more than k standard deviations, applicable to any distribution with finite variances.

Aditya



$$P(|X - \mu| \geq K\sigma) \leq \frac{1}{K^2} \quad K > 0$$

- $P(|X - \mu| \geq K\sigma)$ Probability that X deviates from μ by at least K standard deviations.
- K number of standard deviation ($K > 0$)

Hypothesis Testing

Hypothesis testing is a statistical tool used to evaluate a claim about a population using sample data. It determines whether observed results are due to random chance or if they are statistically significant. The process begins with setting two hypotheses.

Null Hypothesis (H_0) which assumes no effect.

Second Hypothesis

Alternate Hypothesis (H_1) which assumes reflects what we claim

We compare our calculated test values with chosen significance level (α , often 0.05). This comparison governs to reject H_0 or not.

Z-test $\rightarrow$ for ($n \geq 30$) [n = samples]

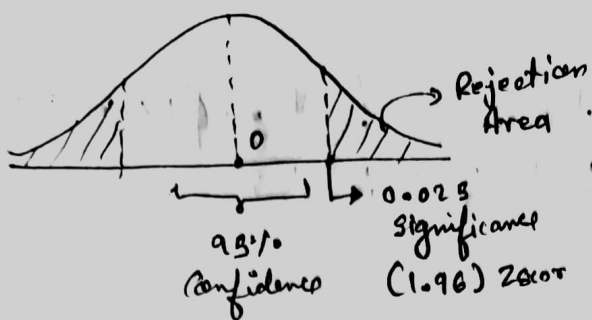
t-test $\rightarrow$ for ($n < 30$)

for Z-test $P \Rightarrow Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

for t-test $P \Rightarrow t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$

$\bar{x}$ = Mean of Sample, μ = Population mean, σ = Population SD
 n = sample size, s = sample SD

Q. Q. Q.



P value will be our test outcome. P value should be either less than 2-score of α significance level & more than (+) significance level.

Example :- A factory has an mean of 80 ml distribution in a single bottle. An employee says it is not true by conducting sample test. with sample size $(n) = 40$ & sample mean was 78 ml. Standard deviation = 2.5

(a) Form Null & Alternative Hypotheses.

Confidence level of 95%. ($\alpha = 0.05$)

(b) Is that Employee right or wrong with 95% Confidence.

$\Rightarrow H_0 = \mu = 80$ (population mean is 80)

$H_1 = \mu \neq 80$ (population mean is not 80 i.e 80 more or less)

$\because n \geq 30$ [$n = 40$] we will apply 2-test

$$p\text{value} = z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

$$\bar{X} = 78, \mu = 80, \sigma = 2.5, n = 40$$

$$\Rightarrow z = \frac{78 - 80}{2.5 / \sqrt{40}} = -5.06$$

$\because$ We can reject H_0 Hypotheses.

Ques

Linear Regression

Linear regression is a method used to find the relationship between two or more variables. It predicts one variable (called dependent variable) based on the values of other variables (called independent or feature variable). In short, it finds a simple linear line in between of the data points with least standard error.

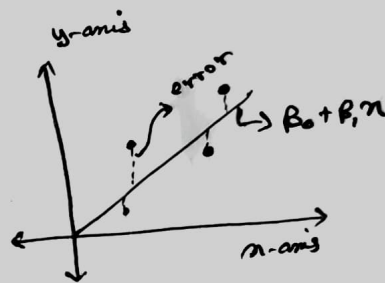
Simple Linear Regression

One dependent & One Independent variable.

$$\left[\begin{array}{c} \text{line to find} \\ \text{or} \\ \text{prediction line} \end{array} \right] \rightarrow y = \beta_0 + \beta_1 x$$

$$\text{Standard Error} \Rightarrow \sum (\hat{y} - y)^2$$

Predicted y value Actual y value



For least standard error we will derive the sum of all errors with our coefficient values.

$$\frac{\partial \sum (\hat{y} - y)^2}{\partial \beta_0} \Rightarrow \frac{\partial \sum (\beta_0 + \beta_1 x - y)^2}{\partial \beta_0}$$

$$\frac{\partial (\sum (\hat{y} - y)^2)}{\partial \beta_1} \Rightarrow \frac{\partial (\sum (\beta_0 + \beta_1 x - y)^2)}{\partial \beta_1}$$

that gives us $\Rightarrow$

$$\beta_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}, \quad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

Q. Q. Q.

Multiple Linear Regression

One dependent & N number of Independent variable

Normal
Eqn we
have to find

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} \dots$$

let us assume all in vectors now

→ This is in
multi dimension
plane

⇒

$Y \rightarrow$ dependent variable actual
values

(vector)
Terms
Used
in Prediction
Line

$X \rightarrow$ feature variable

$\beta \rightarrow$ Coefficient vector (which we have to find)

$\epsilon \rightarrow$ Errors

$$Y = X\beta + \epsilon$$

$$\Rightarrow \epsilon = Y - X\beta$$

$$\Rightarrow \frac{\partial(\epsilon)}{\partial \beta} = 0 \quad \text{for minimum error} \quad \text{--- (1)}$$

from Eqn(1) we got ⇒

$$\beta = (X^T X)^{-1} (X^T Y)$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n3} & \dots & x_{np} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}$$

Ques