

Hand Wrtten / Easy explained Notes



Complete Statistics for Data Science



◀ SWIPE ▶



STATISTICS FOR THE DATA SCIENCE

Part - 1

- WHY STATISTICS IN DATA SCIENCE?
- TYPES OF STATISTICS: DESCRIPTIVE VS. INFERENCEAL
- POPULATION VS. SAMPLE DATA
- SCALE OF MEASUREMENT
- MEASURES OF CENTER TENDENCY
- MEASURES OF DISPERSION
- SETS IN STATISTICS

#Value_freeContent



@Krishan kumar

Statistics

Definition Statistics is the science of collecting, organizing and analyzing data

Data facts or pieces of information → It can be major and collect
ex → ① Height in the classroom of the student [175, 180, 160] cm
② IQ [100, 90, 80, 60]

Why statistics ?

To innovate any product, to bring value to any product, the role of data is very important. Statistic give a lots of information about data because statistic provide such tools, we can get many information and conclusion from the providing data.

Types of Statistics :-

1. Descriptive statistic

↳ It consists of organizing and summarizing data.

(A) Measure of centre Tendency,

[Mean, Median, Mode]

(B) Measure of Dispersion

[Variance, Std (Standard deviation)]

(C) Different type of Distribution of data

tools :→ [Histogram, PMF (Probability Function)
• PDF ("Density")]

Sample data

Population data

Problem lets say there are 20 Statistic class at your college and you have collected the height of the student in the class. heights are recorded [175, 180, 175, 180, 176, 160, 135, 180] cm.

* Descriptive Question What are the average height of the entire classroom.

So Mean \rightarrow Average \rightarrow this is the part of descriptive stats.

$$\hookrightarrow \frac{175+180+175+180+176+160+135+180}{8} \rightarrow \text{Average height}$$

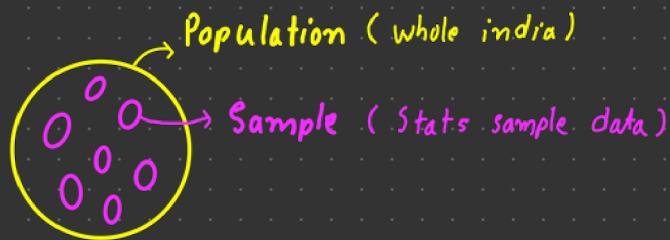
> Inferntion Question Are the height of the student in the classroom similar what you expect in the entire college.
Sample data
 \hookrightarrow population data

★ Population and sample data

* Population Data \Rightarrow The group you are interested in studying

* Sample Data \Rightarrow A subset of population data

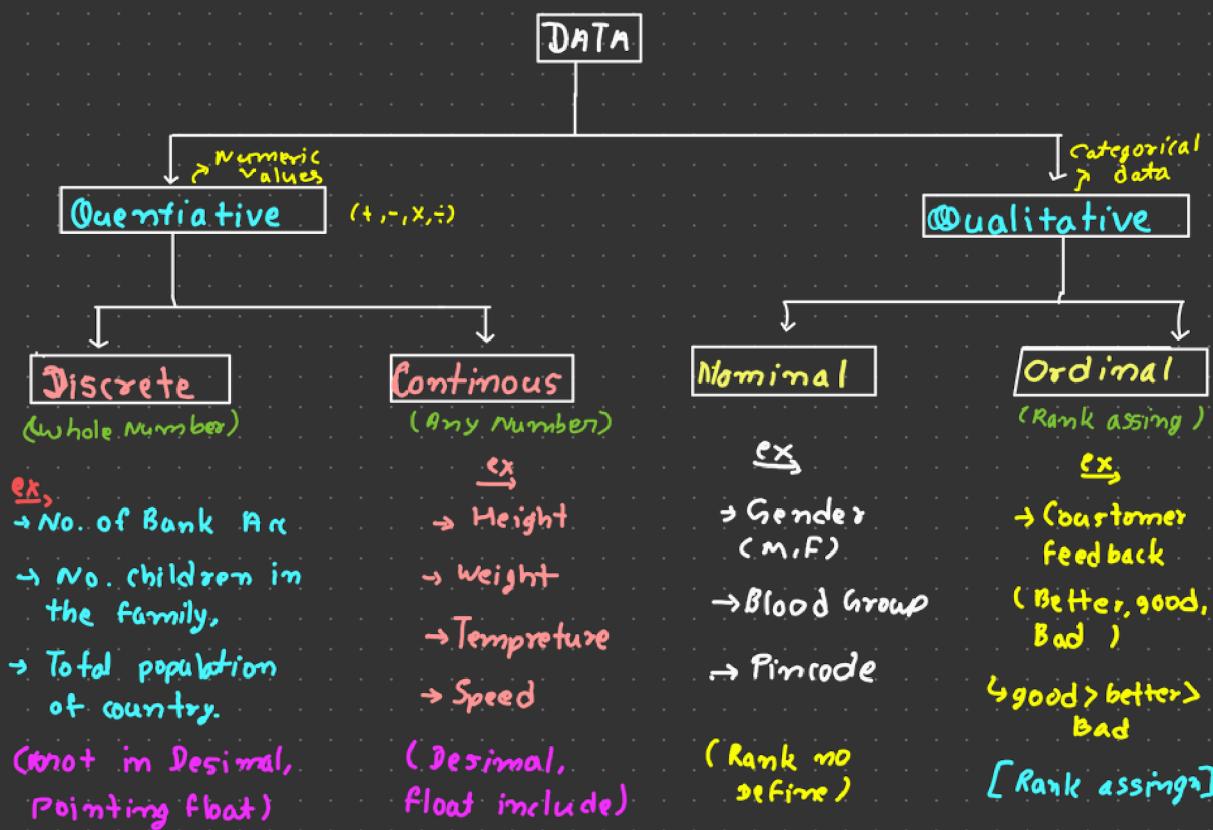
Ex Exit pole of election



Follow:

KRISHAN KUMAR

Types of Data



Follow:

KRISHAN KUMAR

Scale of Measurement

- Nominal scale data.
- Ordinal scale data.
- Interval scale data.
- Ratio scale data.

↳ Nominal scale Data :-

- Qualitative or categorical data
- Order does not matter
- ↳ ex → Gender / color / habit

ex favorite color.
Red → 5 → 50%
Black → 3 → 30%
Blue → 2 → 20%
10 by nominal
data we find
called → cumulative freq.

there are
no any
specific rank

↳ Ordinal scale data

- * Ranking is imp.
- * Order is matter
- * Difference can't be measured

ex 1 → Best
2 → Good
3 → Bad

ex Working Profession
Productivity

Hours	Day
1	Mon.
2	Tue.
3	Wed.
4	Th
5	Fri

Best work on → Friday

ex Race
1st
2nd
3rd

In ordinal Difference
can't be measured.

Follow:

KRISHAN KUMAR



Internal scale data

- * The order matter
 - * Difference can't be measured.
 - * Ratio can't measured
 - * No '0' starting fixed
- It can be -ve OR +ve

Ex, Temp Variable →

$$\begin{array}{l} 60^{\circ}\text{F} \\ 90^{\circ}\text{F} \\ 120^{\circ}\text{F} \\ 180^{\circ}\text{F} \end{array} \left. \begin{array}{l} \text{Ratio} \rightarrow 60:90 \rightarrow 2:3 \\ \text{Diff. can be measured} \\ 90-30 \rightarrow 60^{\circ}\text{F} \\ \frac{120}{180} \rightarrow \frac{60}{90} + \frac{3}{2} \end{array} \right.$$

Ratio does not mean
that Temp. increase
OR ↓ $\frac{3}{2}$ %.



Ratio scale data

- * The order matter
- * Diff. are measurable (including ratio)
- * Contain a '0' Starting point.

Ex, Student marks in a class.
0, 90, 60, 30, 75, 80, 85, 50

Ascending order → 0, 30, 50, 60, 75, 80, 85, 90

$$\begin{aligned} \text{Diff.} \rightarrow 40-30 &= 10 \\ 80-50 &= 30 \end{aligned}$$

Ratio measured in Grading.

question not in Temp. Type question

$\frac{30}{90} = \boxed{3:1}$ This person
gains 3x marks from
that one (30marks)

Measure of centre tendency

► Mean OR Average

► Median

► Mode

Follow:

KRISHAN KUMAR

► Mean or Average :-

Population (N)

$$x = \{1, 1, 2, 2, 3, 3, 4, 4, 5, 5\}$$

$$\text{population mean } (N) = \sum_{j=1}^n \frac{x_j}{N}$$

$$\Rightarrow \frac{1+1+2+2+3+3+4+4+5+5}{10}$$

$$\Rightarrow \frac{32}{10} = 3.2 \Rightarrow \text{Average}$$

$$N = 3.2$$

Sample mean = \bar{x}

$$\bar{x} \Rightarrow \sum_{j=1}^n \frac{x_j}{n}$$

$$\bar{x} = \frac{32}{10} = 3.2$$

$$\boxed{\bar{x} = 3.2}$$

► Median :-

$$x = [4, 5, 2, 3, 2, 1]$$



Steps

→ Sort the random variable $\rightarrow [1, 2, 2, 3, 4, 5]$

→ No. of element , count $\rightarrow 6$

→ if count == even

$$[1, 2, \boxed{2, 3}, 4, 5]$$

$$\therefore \text{Median} \rightarrow \frac{2+3}{2} \cancel{, 2.5}$$

→ if count == odd

$$[1, 2, 2, \boxed{3}, 3, 4, 5]$$

$$\boxed{\text{Median} = 3}$$

Why Median?

→ Median is used to find centre of tendency when outlier is present.

$$x = [1, 2, 3, 4, 5]$$

$$\bar{x} = \frac{1+2+3+4+5}{5}$$

$$\bar{x} = 3$$

Dif. is mean

$$x = [1, 2, 3, 4, 5, 100] \quad \text{→ outlier}$$

$$\bar{x} = \frac{1+2+3+4+5+100}{6} = \frac{115}{6}$$

$$\bar{x} = 19.16$$

In Median

$$x = [1, 2, 3, 4, 5, 100]$$

$$\text{Median} \rightarrow \frac{3+4}{2} \rightarrow \frac{7}{2}$$

$$\text{Median} = 3.5$$

So the result is, Mean with outlier → 3 \curvearrowright 19.16

Median with outlier

3 \curvearrowright 3.5
(every close)

Note

Whenever we find out centre of tendency we should do median.

► Mode :-

↳ Frequency maximum

↳ Maximum repeating Number

$$X_{\text{show}} = [1, 2, 3, 4, 5, 8, 9, 1, 2, 3, 2, 8, 9, 1, 2, 3, 4]$$

Where and why we use mean, median, mode,

↳ We use these all in EDA and Feature engineering.

for ex →

Age	Numeric fe.		Gender	Mode	Mode
	weight	Salary			Degree
24	70	40K	M	BE	
25	80	50K	F	-	→ Data is missing
27	95	70K	M	-	
24	-	35K	-	BE	
32	-	70K	m	PHD	
-	60	40K	-	Master	
-	65	-	F	BSc	
40	72	-	m	BE	

Note

Missing data filling

- * if data is → Categorical → fill by Mode (best result)
- * if data is → Numerical → fill by Mean (best result)
- * Data with Outlier → fill by Median (best result)

Follow:

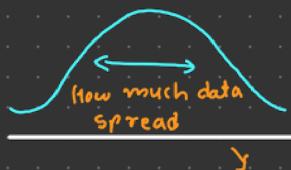
 KRISHAN KUMAR

Measure of Dispersion :- (How much our data spread)

Types

(i) Variance

(ii) Standard deviation



To calculate the this spreaded data we use → Measure of dispersion

(i) Variance :-

(A) Population variance

$$\sigma^2 = \frac{1}{N} \sum_{j=1}^N (x_j - \mu)^2$$

x_j → Data point

μ → Population mean

N → Population size

(B) Sample variance

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

x_i → Data point

\bar{x} → Sample mean

n → Sample size

Q Why we devide Sample variance by $n-1$?

Sq The sample variance devided by $n-1$ So that we can create an unbiased estimator of the population variance

↳ This scenario → called → Bencle correction

Ex [1, 2, 3, 4, 5], $s^2 = ?$

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

$$\Rightarrow s^2 = \frac{10}{4} \Rightarrow 2$$

x	\bar{x}	$(x_i - \bar{x})$
1	3	4
2	3	1
3	3	0
4	3	-1
5	3	2
Mean is 3		10

Sample variance → 2

* What difference in these two sample variance? X, Y

$$\hookrightarrow X = 2.5$$

$$Y = 7.5$$

$$S^2 = 2.5$$

$$S^2 = 7.5$$

Ans S^2 denote \Rightarrow Dispersion of spread

$$S^2 = 2.5$$



\rightarrow When spreadness = ↑↑ (increase)

Height becomes - ↓↓ (decrease)

► Standard deviation :- (std) (σ)

(A) Population Standard deviation

$$\sigma = \sqrt{\text{variance}}$$

(B) Sample std.

$$\text{Sample std} = \sqrt{s^2}$$

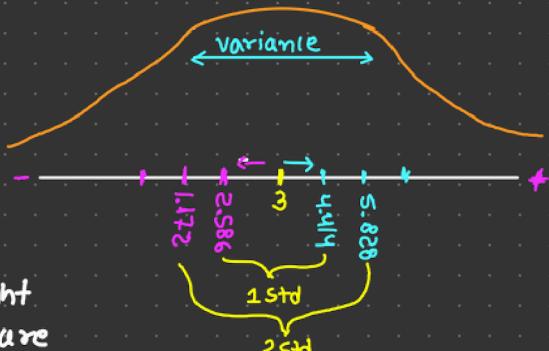
$s^2 \rightarrow$ sample variance

Ex

$$X = [1, 2, 3, 4, 5]$$

$$\text{Sample mean} = \bar{x} = 3$$

$$\sigma = \sqrt{3} = 1.732$$



When one step towards the right and one step towards the left are combined then we called them 'one std'

follow:

Where the 2 element fold in the variance.

Ans \rightarrow One step left to the mean.

$$\begin{array}{r} 3.000 \\ + 1.414 \\ \hline 4.414 \\ - 2.856 \\ \hline 1.558 \\ + 1.414 \\ \hline 3.000 \end{array}$$

$$\begin{array}{r} + 1.414 \\ + 1.414 \\ \hline 5.828 \end{array}$$

Random variable

Random variable is a process of mapping the output of a random process or experiments to a number there are not any fix value.

Ex,

Tossing a coin

$$X = \begin{cases} 0, & \text{if Head} \\ 1, & \text{if Tail} \end{cases}$$

Rolling a dice

$$Y = \{ \text{Sum of the rolling a dice 7 times} \}$$

$$\hookrightarrow P(Y \geq 15) \quad \underline{\text{OR}} \quad P(Y < 10)$$

Probability of

Sets

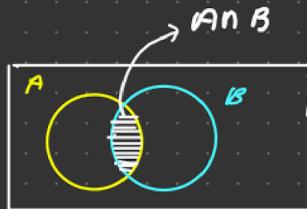
$$A = \{1, 2, 3, 4, 5, 6, 7, 8\}$$

$$B = \{3, 4, 5, 6, 7\}$$

Intersection ($A \cap B$)

(common value)

$$A \cap B = \{3, 4, 5, 6, 7\}$$

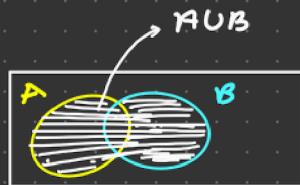


Follow:

KRISHAN KUMAR

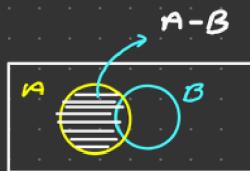
Union

$$A \cup B = \{1, 2, 3, 4, 5, 6, 7, 8\}$$



Difference

$$A - B = \{1, 2, 8\}$$



Subset (son)

$$A \rightarrow B = \text{False} \quad (\because \text{all element of } B \text{ Present in } A)$$

$$B \rightarrow A = \text{True} \quad (\because \text{All element of } A \text{ Present in } A)$$

Superset

$$B \rightarrow A = \text{False}$$

$$A \rightarrow B = \text{True}$$

-

Follow:

 KRISHAN KUMAR



STATISTICS FOR THE DATA SCIENCE

Part - 2

- HISTOGRAM AND SKEWNESS
- COVARIANCE AND CORRELATION
- ADVANTAGE VS DISADVANTAGE OF VARIANCE
- PEARSON CORRELATION COEFFICIENT
- PDF / PMF / CDF
- BERNOULLY DISTRIBUTION
- BINOMIAL DISTRIBUTION

#Value_freeContent



@Krishan kumar

Histograms And Skewness

Age = {10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 45, 51}

0-50
value

$$\hookrightarrow \frac{50}{10} = 5 \text{ bin size} \quad (\text{No. of bins} \rightarrow 10)$$

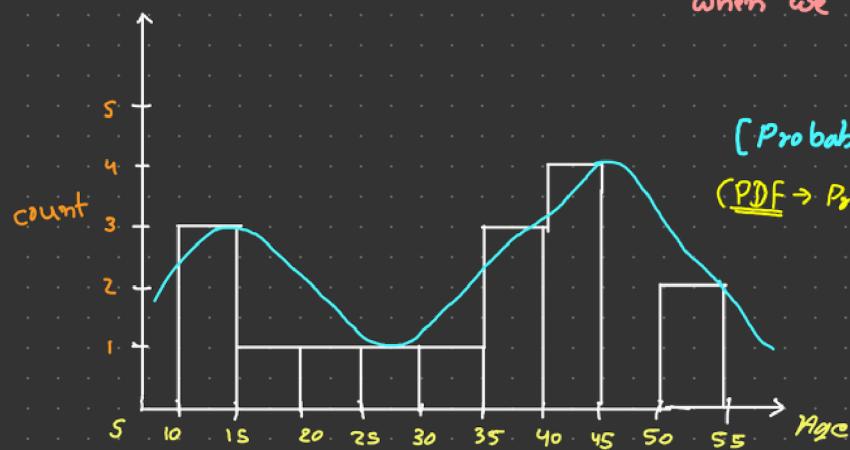
$$\hookrightarrow \frac{50}{20} = 2.5 \text{ bin size} \quad (\text{No. of bins} \rightarrow 20)$$

when we smoothen our histogram
↓

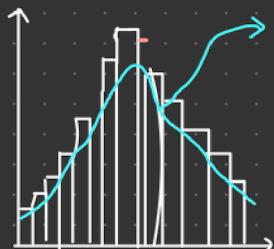
Technique

↓
[Probability distribution fun.]

(PDF → Probability density function)



Skewness



Normal / Gaussian distribution



These types of distribution is
called → Symmetrical distribution.

follow:



KRISHAN KUMAR

* Symmetrical Data



NO Skewness
(No curve)

equal from both
side



$$\Rightarrow Q_3 - Q_2 \approx Q_2 - Q_1$$

The mean, median, mode are all perfectly at the centre.

Mean = Median = Mode

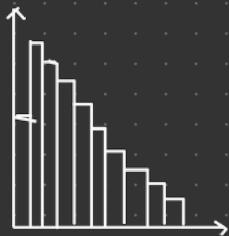


There are no skewness in the symmetrical distribution data and just because there are no any maxi value and minimum value in Right OR Left corner of data.

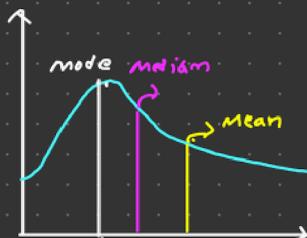
② Right Skewed data

Means

longitude in Right side (Majority data on right side present)



\Rightarrow Positive / Right Skewed \Rightarrow Distribution in PDF



Relation btw

follow:

KRISH
mean > median
mode

Mean > Median > Mode

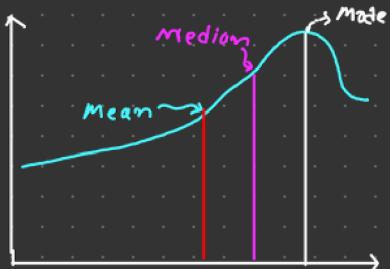


* Box plot for Right Skewed data



$$\Rightarrow Q_3 - Q_2 \geq Q_2 - Q_1$$

* Left Skewed data



* Box plot



$$\Rightarrow Q_2 - Q_1 \geq Q_3 - Q_2$$

Relation

Mean < Median < Mode



Covariance and Correlation

[To reactify Relation btw X and Y]

X	Y
2	3
4	5
6	7
8	9

In this scenario →

X↑	Y↑
X↓	Y↑
X↑	Y↓
X↓	Y↓

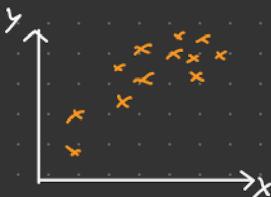
Follow:

KRISHAN KUMAR

For example

$x \uparrow$	$y \uparrow$
$x \downarrow$	$y \downarrow$

Plotted like →



Covariance →

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

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

x_i → Data point of x

\bar{x} → Sample mean of x

y_i → Data point of y

\bar{y} → Sample mean of y

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

$$\Rightarrow \text{Cov}(x, x) = \text{Var}(x)$$

Q What are the difference between covariance and variance?

A Varience of x $\text{Var}(x)$ is nothing but $\text{Cov}(x, x)$. Whenever we talk about variancex it is specifically told about spread of the data.

⇒

$$\text{Cov}(x, y)$$

$x \uparrow$	$y \uparrow$
$x \downarrow$	$y \downarrow$

→ Then the output will be

+ve covariance

$x \uparrow$	$y \downarrow$
$x \downarrow$	$y \uparrow$

→ Then output will be

-ve covariance

$$\begin{array}{cc}
 X & Y \\
 2 & 3 \\
 4 & 5 \\
 6 & 7 \\
 \bar{x} \rightarrow \frac{6}{4} & \bar{y} \rightarrow 5
 \end{array}
 \quad \text{Cov}(x,y) = \sum_{j=1}^n \frac{(x_i - \bar{x})(y_j - \bar{y})}{n-1}$$

$$\Rightarrow \frac{(2-4)(3-5) + (4-4)(5-5) + (6-4)(7-5)}{3-1}$$

$$\Rightarrow \frac{4+0+4}{2} = 4 \rightarrow +ve \text{ value}$$

positive covariance = +ve value

So, X and Y are having a positive covariance.

Advantage of covariance,

- * Relation between X and Y +ve OR -ve value of covariance
- * Disadvantage of covariance,
- * We don't conclude whether which value is covariance with any other value Because \downarrow

Covariance does not have a specific limit value.



Note:- To fix the disadvantage of covariance we use:-

Pearson correlation coefficient

limitation $\rightarrow +1$ to -1

$$r_{x,y} = \frac{\text{Cov}(x,y)}{\sigma_x \times \sigma_y}$$

- * The more value towards $+1$ the more +ve correlated it is (x,y)
- * The more value towards -1 the more -ve correlated it is (x,y)

Spearman Rank correlation

Rank $\rightarrow -1$ to $+1$

$$\rho_s = \frac{\text{Cov}(R(x), R(y))}{\sigma(R(x)) \times \sigma(R(y))}$$

Rank \rightarrow Highest element in the values
makes \rightarrow 1st Rank

X	Y	R(x)	R(y)
1	2	5	5
3	4	4	4
5	6	3	3
7	8	2	1
0	7	6	2
8	1	1	6

follow:

 KRISHAN KUMAR

* Feature selection

+ve	+ve	+ve	= 0	-ve
Size of House	No. of Rooms	Location	No. of people staying	Houmte \Rightarrow Price $\uparrow\uparrow$

PDF / PMF

* PDF = Probability Density Function

* PMF = Probability mass function.

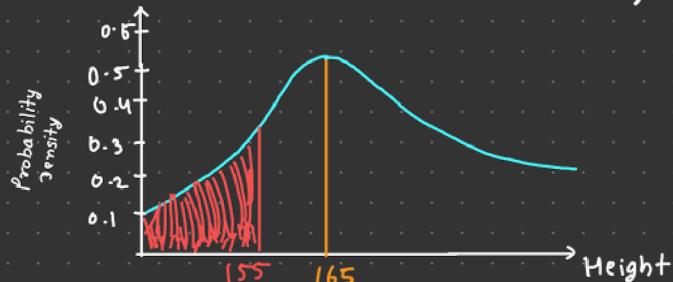
↳ PDF is denote a distribution of data. That helps to understand that how our data will distributed.

► Types of Probability Distribution function

- ① Pdf
- ② Pmf
- ③ Cdf

► Probability density function (Pdf)

↳ Continuous Random variables, ex, Height of students in classroom

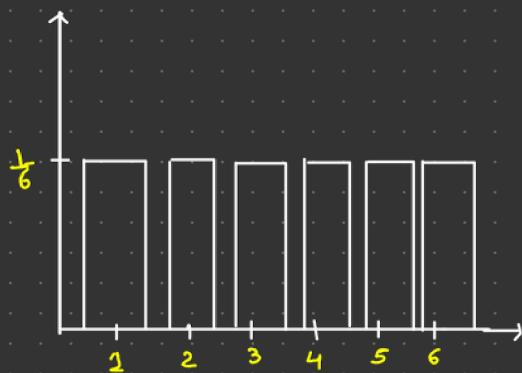


$$\hookrightarrow P(X \leq 155)$$

► Probability mass function (Pmf)

↪ Variable → Discrete Random variable

Ex Rolling a dice $\{1, 2, 3, 4, 5, 6\}$



$$P(X \leq 4)$$

↓

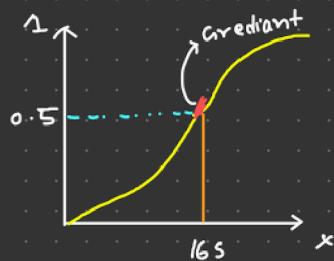
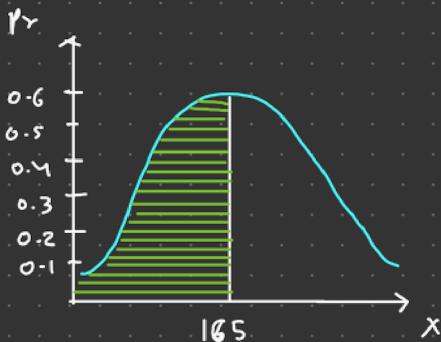
$$P(X=1) + P(X=2) + P(X=3) \\ + P(X=4)$$

$$\Rightarrow \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} \rightarrow \frac{4}{6} = \frac{2}{3}$$

$$P(1) \rightarrow \frac{1}{6}$$

$$P(2) \rightarrow \frac{1}{6}$$

► Cumulative Distribution fun. (cdf),



Follow:

KRISHAN KUMAR

PDF Vs PMF Vs CDF

Relation and difference

① PMF

CDF

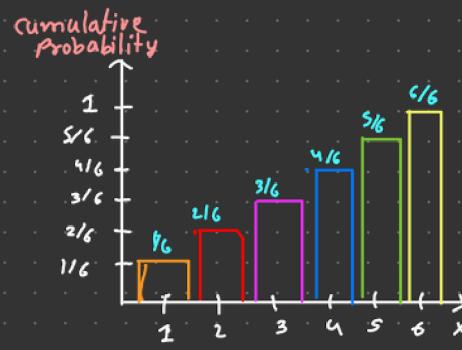
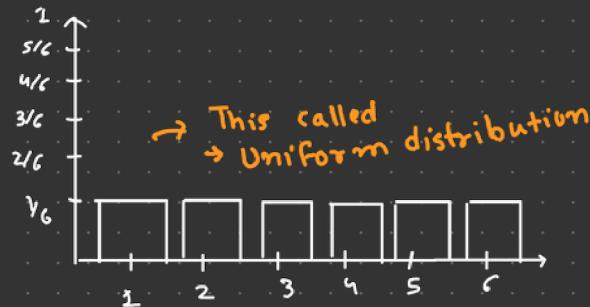
* Discrete Random Variable.

Ex → Rolling a dice $\Rightarrow [1, 2, 3, 4, 5, 6]$

$$P(1) \rightarrow \frac{1}{6}$$

⋮

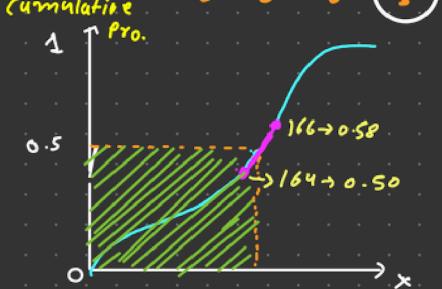
$$P(6) \rightarrow \frac{1}{6}$$



Ex → cdf for $P(X \leq 2)$

$$\begin{aligned} \text{CDF} &= P(X \leq 2) = P(X=1) \\ &\quad + P(X=2) \end{aligned}$$

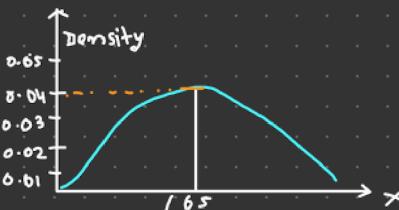
$$= \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$



$$\begin{aligned} \Rightarrow \text{slope} &\Rightarrow \frac{0.58 - 0.5}{1.66 - 1.64} \rightarrow \frac{0.08}{2} \\ &\Rightarrow 0.04 \end{aligned}$$

② PDF

↳ Distribution of continuous Random variable

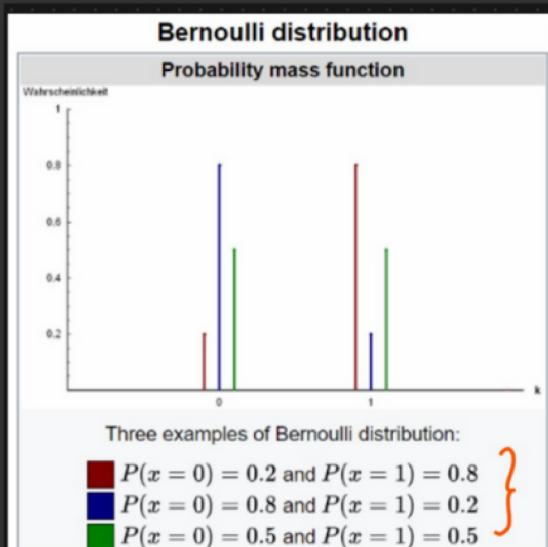


Types of Probability Distribution

- Normal / Gaussian distribution (Pmf)
- Bernoulli distribution (Pmf)
- Uniform distribution (Pmf)
- Poisson distribution (Pmf)
- Log normal distribution (Pmf)
- Binomial distribution (Pmf)

► Bernoulli distribution → In the discrete probability distribution of a random variable which takes the value 1 with probability p and the value 0 with probability $q=1-p$. Less formally it can be thought as a model for the set of possible outcomes of any single experiment that asks a yes-no question.

Pmf → To see the distribution



★ Points

- ① Discrete random variable (Pmf)
- ② Outcomes are binary
 - ex. \cong Tossing a coin (H,T)
 $P(H) \rightarrow 0.5 \rightarrow p$ (Pr. of Success)
 $P(T) \rightarrow 0.5 \rightarrow 1-p \Rightarrow q$
 $(q = \text{Pr. of Failure})$
- ③ When there will fail / pass
 - $P(\text{Pass}) \rightarrow 0.7 \rightarrow p$ Success value
 - $P(\text{Fail}) \rightarrow 1-0.7 \Rightarrow 0.3 = q$ Failure value

$$* \underline{\text{PMF}} = P^k * (1-P)^{1-k}, \quad k \in \{0,1\}$$

if $k=1$

$$\text{PMF} \rightarrow P(k=1) = P^1 * (1-P)^{1-1}$$

$$= P$$

if $k=0$

$$\begin{aligned} \text{PMF} &= P(k=0) = P^0 * (1-P)^{1-0} \\ &= (1-P) = q \\ &\Rightarrow q \end{aligned}$$

Simplified

$$\text{PMF} = \begin{cases} 1-P & , \text{ if } k=0 \\ P & , \text{ if } k=1 \end{cases}$$

\Rightarrow Mean of Bernoulli distribution,

$$\langle \varepsilon(k) \rangle = \sum_{j=1}^k k \cdot P(k)$$

$$, \quad P(k=1) = 0.6 \Rightarrow P$$

$$, \quad P(k=0) = 0.4 \Rightarrow 1-P \Rightarrow q$$

$$\Rightarrow [0 \times 0.4 + 1 \times 0.6] \Rightarrow 0.6 = P$$

Whenever we abstract mean in Bernoulli distribution, then we get P-Value.

\Rightarrow Median of Bernoulli dis,

$$\text{Median} = \begin{cases} 0 & , \text{ if } P < \frac{1}{2} \\ [0,1] & , \text{ if } P = \frac{1}{2} \\ 1 & , \text{ if } P > \frac{1}{2} \end{cases}$$

\Rightarrow Variance of Bernoulli dis.,

$$\text{Var} = P(1-P)$$

$$\boxed{\text{Var} = pq}$$

\Rightarrow Std of Bernoulli dis.,

$$\boxed{\text{Std} = \sqrt{pq}}$$

► **Binomial Distribution** In Statistic the binomial dis. with parameter n and P is the discrete P.d. Dis of the number of success in a sequence of n independent experiments. Each asking Yes or No. question. and each its own boolean-value outcome. success (with Pro. P) or failure ($q=1-p$).

A single success / failure experiment is also called a bernoulli trial OR Bernoulli experiment and a sequence of outcome is called a bernoulli process.

For a single trial ex, $n=1$ the binomial dis. is a bernoulli dis. the binomial dis. is the basis for the popular binomial test of statistical significance.

$$B(n,p)$$

Notation: $B(n,p)$

- * Every experiment outcome is binary
- * This experiment is Performed for n trials
- * Group of bernoulli dis \rightarrow Binomial Distribution

Parameters

- * $n \in \{0, 1, 2, \dots\} \rightarrow$ No. of trials
- * $p \in [0, 1] \rightarrow$ Success Pro. of each trial
- * $q = 1-p$
- * Ex Tossing a coin 10 times

Support: $k \in \{0, 1, 2, 3, \dots, n\}$
 \rightarrow Number of success

Follow:

KRISHAN KUMAR

$$* \text{Pmf} = P(k, n, p) = {}^n C_k p^k (1-p)^{n-k}$$

for $k = 0, 1, 2, 3, \dots, n$ where

$${}^n C_k = \frac{n!}{k! (n-k)!}$$

\Rightarrow Mean :-

$$\text{Mean} = np$$

\Rightarrow Variance :-

$$\text{Var} = npq$$

\Rightarrow Std

$$\text{Std} = \sqrt{npq}$$

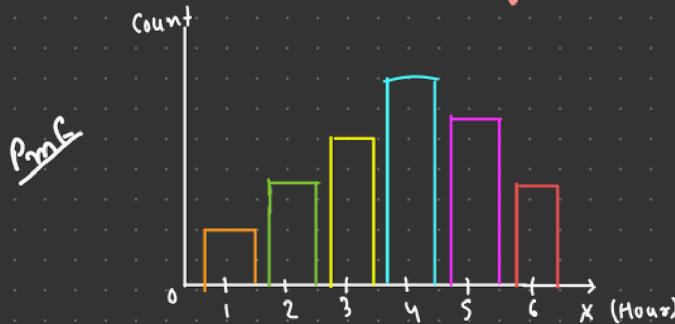
► Poisson Distribution

* Discrete random variable. (Pmf)

* Describe the number of event occurring in fixed time interval.

Ex) ① No. of people visiting hospital every hour:

② No. of people visiting bank every hour.



$\lambda=3$, Expected no. of event occur at every time interval.

Means How many people come at every hour.



STATISTICS FOR THE DATA SCIENCE

Part - 3

- POISSON DISTRIBUTION
- NORMAL / GAUSSIAN DISTRIBUTION
- UNIFORM DISTRIBUTION
- Z - SCORE
- CENTRAL LIMIT THEOREM
- ESTIMATOR
- HYPOTHESIS AND TESTING MECHANISM
- P - VALE
- Z - TEST

#Value_freeContent



@Krishan kumar

Q What is the probability of a person to come at 5th hour.

Ans

$$\text{Pmf} = P(X=5) = \frac{e^{\lambda} - \lambda^x}{\lambda^x}, \text{ if } \lambda = 3 \quad (\text{Ans})$$
$$= \frac{e^3 - 3^5}{5^5} \Rightarrow 0.101 \rightarrow 10.1\%$$

there is 10% possibilities that at 5th hour 3 person come.

Q What is the probability to visit at 5th hour OR 4th hour.

Ans
Pmf

$$P(X=5) + P(X=4)$$

* Mean of poission distribution

$$\text{Mean} = \{ \text{m} = \mu \}$$

$$\boxed{\mu = \lambda \times t}$$

λ = expected no. of event
occur at every time
interval

t = time interval

$$\boxed{\text{Variance of poission} = \lambda \times t}$$

Follow:

 KRISHAN KUMAR

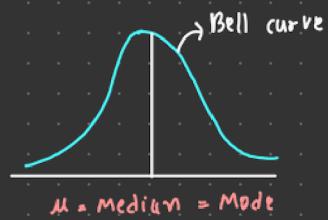
► Normal / Gaussian Distribution \rightarrow (P.d.F)

Notation :- $N(\mu, \sigma^2)$

Parameter :- $\mu \in \mathbb{R}$ ^{Real number} (Mean)

$\sigma^2 \in \mathbb{R} > 0$ = variance

$x \in \mathbb{R}$ = Data points



$$\text{PDF} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i - \mu)^2}{\sigma^2}}$$

⇒ Mean of normal distribution

Mean = μ = Average

⇒ Variance

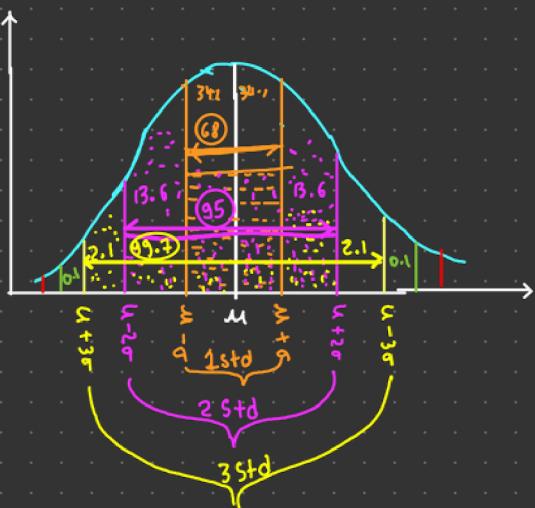
$$\text{Var} = \sigma^2$$

⇒ Std

$$\sigma = \sqrt{\text{Var}}$$

This rule follow

68-95-99.7% Rule



Follow:

KRISHAN KUMAR

$$P(\mu - \sigma \leq x \leq \mu + \sigma) \approx 68\%$$

$$P(\mu - 2\sigma \leq x \leq \mu + 2\sigma) \approx 95\%$$

$$P(\mu - 3\sigma \leq x \leq \mu + 3\sigma) \approx 99.7\%$$

- Ex,
- ① Weight of the students in the class.
 - ② Height of " " " "
 - ③ IRIS Dataset [Sepal width]

► Uniform Distribution →

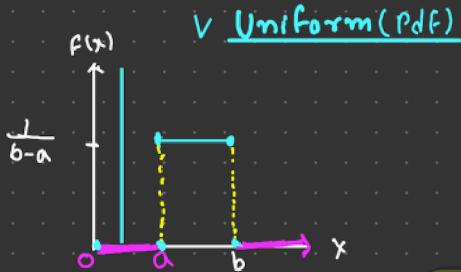
(A) Continuous uniform distribution (Pdf)

(B) Discrete Uniform distribution (Pmf)

(A) Continuous Uniform Distribution,

In statistic, the continuous Uniform dis. OR Rectangular dis. is a family of symmetric probability dis. The dis. describe an experiment where there is an arbitrary outcome that lie between certain bounds.

the bounds are defined by the parameter a and b, which are the minimum and maximum va



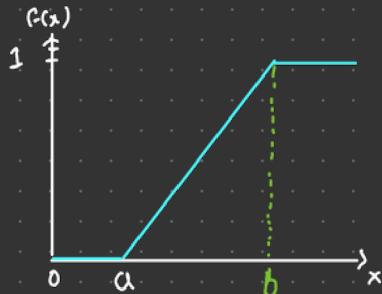
V Uniform(Pdf)

Notation → $V(a,b)$

Parameter,

$$-\infty < a < b < \infty$$

► Cdf



$$pdf \rightarrow \begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & , \text{ otherwise} \end{cases}$$

$$cdf \rightarrow \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a, b] \\ 1 & \text{for } x > b \end{cases}$$

Mean $\rightarrow \frac{1}{2}(a+b)$

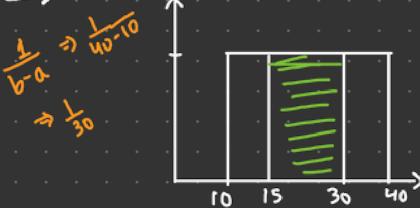
Variance $\rightarrow \frac{1}{12}(b-a)^2$

Median $\rightarrow \frac{1}{2}(a+b)$

Eg The no. of candies sold daily at a shop is uniformly dis. with a maxi at 40 and mini of 10.

① Probability of daily sales falls between 15 and 30.

Sol:



$$\lambda_1 = 15$$

$$\lambda_2 = 30$$

$$P(15 \leq X \leq 30) = (\lambda_2 - \lambda_1) \times \frac{1}{b-a}$$

$$\Rightarrow 15 \times \frac{1}{30} \Rightarrow \frac{1}{2} \Rightarrow 0.5$$

② $P(X \geq 20)$

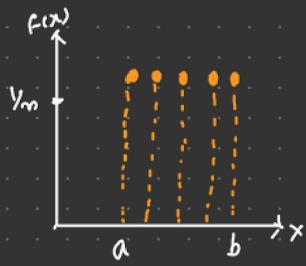
Sol. $(40-20) \times \frac{1}{30}$

$\Rightarrow 0.66 \rightarrow 66\%$

follow:

(B) Discrete uniform dis. (Pmf)

In statistic the discrete uniform dis. is symmetric probability dis. wherin a finite number of value are equally likely to be observed every one of n value has equally probability $\frac{1}{n}$. Another way of saying that "discrete uniform Dis" would be a known finite number of outcome equally likely to happen.



Ex Rolling a dice

$$[1, 2, 3, 4, 5, 6]$$

$$\begin{aligned} \hookrightarrow P(1) &\rightarrow \frac{1}{6} & a &\rightarrow 1 \\ &\vdots && \\ P(6) &\rightarrow \frac{1}{6} & b &\rightarrow 6 \end{aligned}$$

Notation $\rightarrow m(a, b)$

$$\frac{1}{n} = n = b - a + 1$$

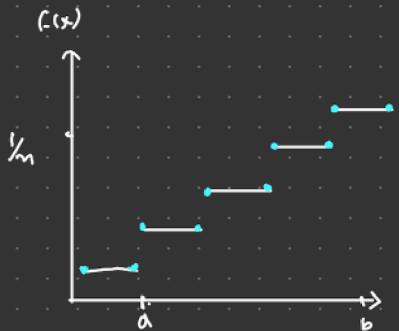
Parameter,

a, b with $b \geq a$

$$\text{Pmf} = \frac{1}{n}$$

$$\text{Mean} = \frac{a+b}{2}$$

Median



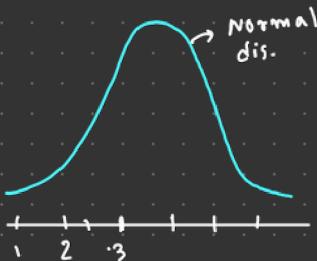
Follow:

KRISHAN KUMAR

Standard Normal Distribution and Z-score

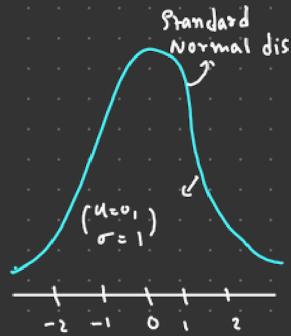
$$X \in [1, 2, 3, 4, 5], \mu = 3$$

$$\sigma = 1.414 \approx 1$$



Transition Technique

When
 $\mu = 0$
 $\sigma = 1$



\Rightarrow Z-Score

Z-score is used for the help for Normal dis. to transform in standard normal distribution.

$$\text{Z-Score} = \frac{x_i - \mu}{\sigma}$$

$$x_i = 1, \frac{1-3}{1} \Rightarrow -2$$

$$x_i = 4 \Rightarrow \frac{4-3}{1} \rightarrow 1$$

$$x_i = 2, \frac{2-3}{1} \Rightarrow -1$$

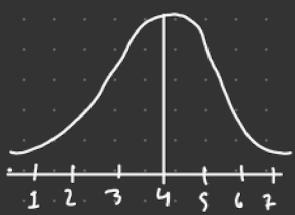
$$x_i = 3 \Rightarrow \frac{3-3}{1} \rightarrow 0$$

$$x_i = 3, \frac{3-3}{1} \Rightarrow 0$$

o Away from the mean for a specific mean how much Std fall for the x_i .

by Z-score,

$$x_i = 4 \Rightarrow \frac{4-3}{1} \rightarrow 1$$



$$\left\{ \begin{array}{l} \mu = 4 \\ \sigma = 1 \end{array} \right\} \quad (\text{Normal dis.fun.})$$

Q How many standard deviation 4.5 is away from the mean?

Ans $x_i = 4.5$

$$\text{Z-score} \rightarrow \frac{4.5 - 4}{1} \rightarrow 0.5$$

0.5 Std from the right.

Q What percentage of data is falling above 4.5



$$\mu \rightarrow 4$$

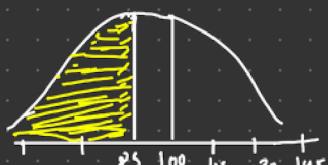
$$\sigma \rightarrow 1$$

$$\text{Z-score} \rightarrow \frac{4.5 - 4}{1} \Rightarrow 2.5$$

Area under the curve (≤ 2.5) $\Rightarrow 1 - 0.6681 \rightarrow 6.6\%$

Problem In India average IQ is 100, with a std of 15. What is the percentage of the population whose you expect to have an IQ lower than 85.

$$\mu \rightarrow 100, \sigma = 15$$

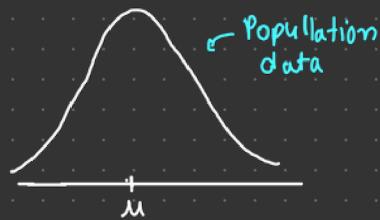


$$\text{Z-score} \rightarrow \frac{85 - 100}{15} \rightarrow -\frac{15}{15} \rightarrow -1$$

Area under the curve (≤ 85) $= 0.15866 \rightarrow 15.8\%$

Central limit theorem

type, ① $X \sim N(\mu, \sigma)$



$n = 20$
sample distribution

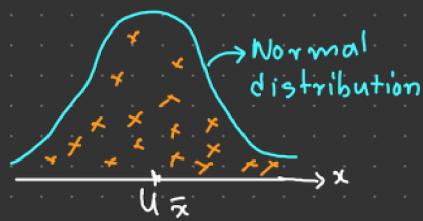
$$S_1 \rightarrow [x_1, x_2, \dots, x_{10}] = \bar{x}_1$$

$$S_2 \rightarrow [x_2, x_3, \dots, x_{20}] = \bar{x}_2$$

⋮

$$S_n \rightarrow \bar{x}_n \Rightarrow \text{Sample mean.}$$

$$\bar{x} = \{\bar{x}_1 + \bar{x}_2 + \dots + \bar{x}_n\}$$

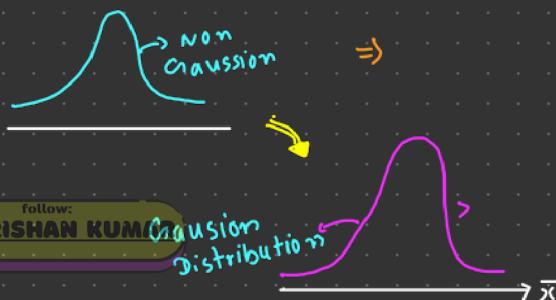


If we have population data who is normally distributed. On this we applied Sampling distri. and after that calculating sample mean.

↳ when we plotted all those mean \rightarrow We got normal distribution

② $X \not\sim N(\mu, \sigma)$

$n \geq 30$



$$S_1 \rightarrow \{x_1, x_2, \dots, x_{30}\} \rightarrow \bar{x}_1$$

$$S_2 \rightarrow \{x_2, x_3, \dots, x_{30}\} \rightarrow \bar{x}_2$$

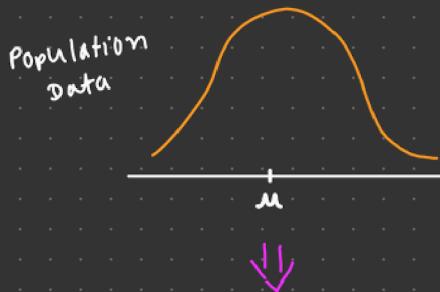
⋮

$$S_n \rightarrow \bar{x}_n$$

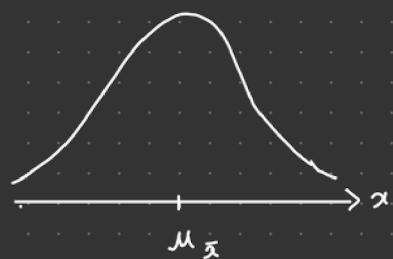
⇒ CLT Properties

the central limit theorem says that the sampling dis. of the mean will be always be normally distributed, as long as the sample size is large enough. Regardless of whether the population has a normal, poisson, binomial or any other dist. the sampling distribution of the mean will be normal.

Important for the interview



Sampling distribution
of mean ((LT))



$$x \sim N(\mu, \sigma)$$

σ → population std

μ → Population mean

n → Sample size.

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

n can be any value

Q What is the Std in sample distribution?

S

$$\text{Std} \Rightarrow \frac{\sigma}{\sqrt{n}}$$

Inferential Statistics

Estimate :

It is an observe numerical value used to estimate an unknown population parameter.

① Point estimate

Single numerical value used to estimate the unknown population parameter.

* Sample mean is a point estimate of a population mean.



There is a huge gape, to counter those gape we use Interval estimate.

② Interval estimate

Range of value used to estimate the unknown population parameter.

* Interval estimate of population parameter are called confidence interval.



55 - 65



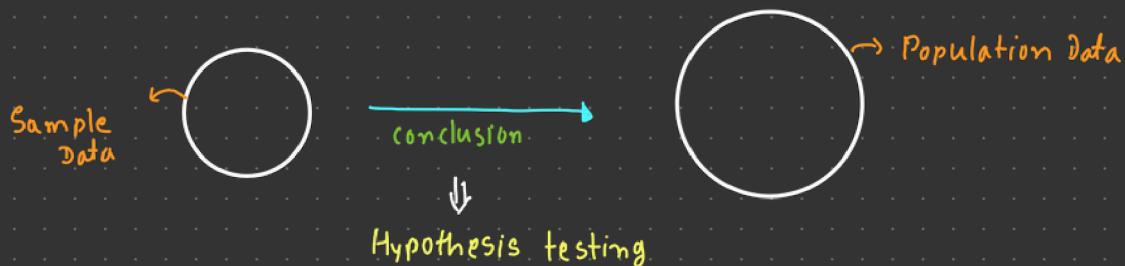
Follow:

KRISHAN KUMAR



A Hypothesis and Hypothesis testing mechanism:-

↳ In inferential Stats. used for Conclusion or Inferences.



★ Hypothesis testing Mechanism

↔ Person ~~not~~ done crime

(1) Null Hypothesis (H_0) → The person is not Guilty
↳ The assumption that you are beginning with.

(2) Alternate Hypothesis (H_1), The person is Guilty

↳ Opposite of Null Hypothesis

(3) Experiments → prove collect (by help of \rightarrow Statistical Analysis)
↳ { DNA, Tests, Finge prints } \rightarrow (p-value)

(4) Accept the null hypothesis OR Reject the null hypothesis

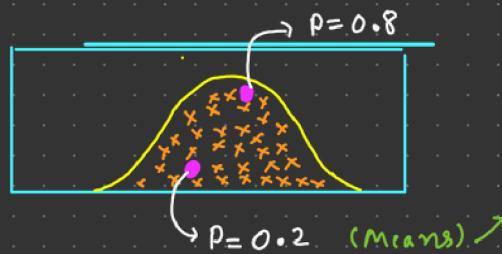
for example, College at district a stats says its average passed percentage of statistic are 85%. A New college opened in the district and it was found that a sample of 100 Students have a pass % of 90 with a Standard deviation of 40%. Does this school have a different Passed%.

Ans → Null Hypothesis $\rightarrow (H_0) \rightarrow \mu = 85\%$
Alternate hypothesis $\rightarrow (H_1) \rightarrow \mu \neq 85\%$

P-Value

The P-value is a number, calculated from a statistical test, that describes how likely you are to have found a particular set of observations if the null hypothesis were true. P-values are used in hypothesis testing to help decide whether to reject the null hypothesis.

Ex) Using keyboards spare top.



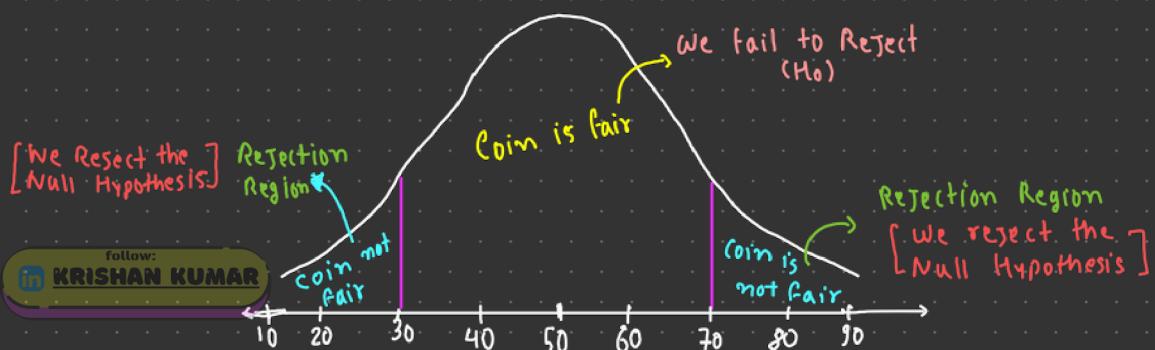
out of 100 touches in this key the probability of touching in this region ≥ 0.8

↳ Hypothesis testing

Ex) Whether coin is fair or Not. f 100 Tosses?

Ans)

- ① Null Hypothesis (H_0): coin is fair
- ② Alternate hypothesis (H_1): Coin is not fair
- ③ Experiment;



★ Significance value (α)

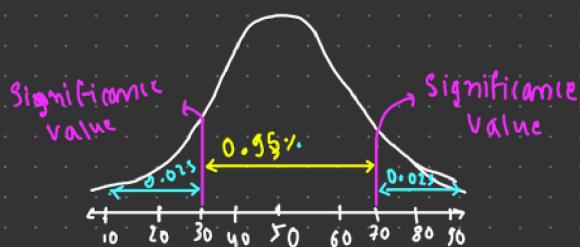
Let,

$$\alpha = 0.05$$

confidence interval (C.I.)

$$C.I. \Rightarrow 1 - 0.05$$

$$\Rightarrow \underline{\underline{0.95}}$$



* Conclusion

Let, $P = 0.01\%$

$\Rightarrow P < \text{Significance}$

\therefore We reject the null hypothesis

else:

We fail to reject null hypothesis.

Follow:

KRISHAN KUMAR

Hypothesis testing and statistical Analysis:-

- (A) Z - Test } Average
- (B) T - Test }
- (C) CHI SQUARE } categorical data
- (D) ANNOVA } variance

(A) Z-test :-

Condition : Z-score only apply on where,

- (i) Population std
- (ii) $n \geq 30$

Problem

The average height of all the residents in a city is 168cm. with a $\sigma = 3.9$. A Doctor believe that mean to be different. He measure the height of 36 individuals and found the average height to be 169.5 cm.

(a) State Null and alternate hypothesis.

(b) At 95% confidence level, is there enough evidence to reject the null hypothesis.

Method 1 \rightarrow Z-test

$$\mu = 168\text{cm}, \sigma = 3.9, n = 36, \bar{x} = 169.5\text{cm}$$

(a) Null hypothesis (H_0) = $\mu = 168\text{cm}$

Alternate hypothesis (H_1) = $\mu \neq 168$ {2 Tail test}

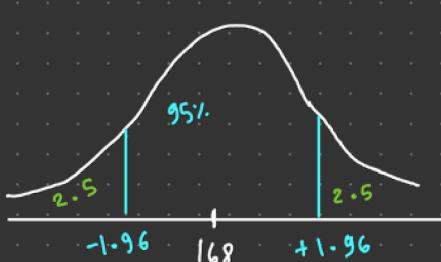
It can be greater OR
lesser than also.

Follow:

KRISHAN KUMAR

$$(b) C.I. = 0.95, \alpha = 1 - 0.95 \\ = \underline{\underline{0.05}}$$

Decision boundary



* Area under the curve

$$1 - 0.95 \Rightarrow 0.05 \\ 0.25 \leftarrow \overbrace{\quad}^{0.25}$$

$$\Rightarrow 1 - 0.25 \Rightarrow \boxed{0.9750}$$

Area under the curve

$$Z\text{-Score} \rightarrow 0.9750 \rightarrow +1.96$$

Z-Score for population data

$$Z\text{-Score} = \frac{x_i - \mu}{\sigma}$$

Z-Score for sample data

$$Z\text{-Score} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

X Statistical Analysis

$$Z\text{-Score} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \Rightarrow \frac{169.5 - 168}{3.9 / \sqrt{36}} \Rightarrow 2.31$$

If Z-test value is less than -1.96 OR greater than +1.96 we

Reject the Null hypothesis

else

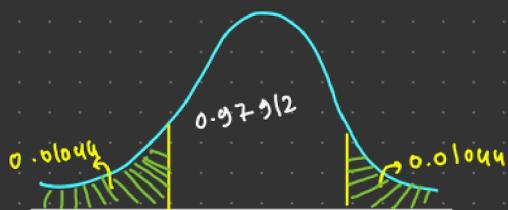
we Accept the Null hypothesis

$2.31 > +1.96$ { So, we Reject the null hypothesis }

Method 2 → p-value

* Statistical Analysis →

$$Z\text{-Score} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{169.5 - 168}{3.9/\sqrt{36}} \Rightarrow 2.31$$



In Z-table 2.31 ⇒ 0.98956

$$\begin{aligned} & 1 - \text{Area under the curve} \\ & \Rightarrow 1 - 0.98956 \\ & \Rightarrow \boxed{0.01044} \end{aligned}$$

* How to calculate p-value →

$$\Rightarrow 0.01044 + 0.01044$$

$$\Rightarrow \underline{\underline{0.02088}}$$

If P value < Significance

$$0.02088 < 0.05$$

{We Rejecting the null hypothesis}

else-

we accept the null hypothesis.

↳ Here we accepting the null hypothesis.

Problem: 2

A factory manufacture bulbs with a average warranty of 5 yrs with standard deviation of 0.50%. A worker believe that the bulb will manufacture in less than 5 yrs. He test a sample of 40 bulbs and find the average time to be 4.8 years.

(a) State the null hypothesis and alternate

(b) At a 2% significance level, is there enough evidence to support the idea that the warranty should be revised.

follow:

MSHARJUMIR

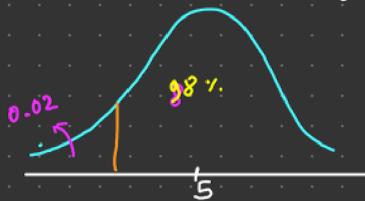
$$\rightarrow \mu = 5, \sigma = 0.50, \bar{x} = 4.8, n = 40$$

\Rightarrow (A) Null hypothesis (H_0) $\Rightarrow \boxed{\mu = 5}$

\hookrightarrow Alternate hypothesis (H_1) $\Rightarrow \underline{\mu < 5}$ { 2 Tail test }
 (only checking one side)

* Method (P-Value)

Decision Boundary

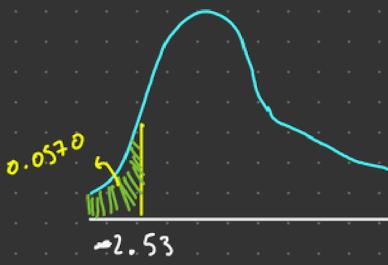


C.I. $\rightarrow 0.98$

$$\alpha = 1 - 0.98 \\ \Rightarrow \boxed{0.02}$$

\rightarrow P-value \rightarrow

$$z\text{-test} \rightarrow \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \rightarrow \frac{4.8 - 5}{\frac{0.05}{\sqrt{40}}} \rightarrow \frac{-0.2}{0.079} \rightarrow -2.53$$



\hookrightarrow Area under the curve of
 -2.53 , z value is $= \boxed{0.0570}$

$$P\text{-value} = 0.0570$$

if

Pvalue $<$ Significance

$$0.0570 < 0.02 \rightarrow \text{False}$$

* Conclusion

The warranty needs to be revised. So, we accept the Null hypothesis.



STATISTICS FOR THE DATA SCIENCE

Part - 4

- STUDENT T-DISTRIBUTION
- BAYES THEOREM
- CHI - SQUARE TEST
- F - DISTRIBUTION
- ANOVA AND ITS TYPES

#Value_freeContent



@Krishan kumar

 **(B) Student t-distribution** → Statistical analysis using Z-score
we need population standard deviation (σ).

* How do we perform an analysis when we don't know the population standard deviation.

Ans. Student t-distribution

in Z-score,

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

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



s = Sample std.

* Degree of freedom →

$$d.f. = n - 1$$

n ⇒ Sample size

we 3 people

X □ □

► We can find area under the curve by using t-value OR dof.

for t-distribution where population std (not given)

only
Sample std (given)

Follow:

 KRISHAN KUMAR

Problem

In the population the average IQ is 100. A team of researchers want to test a new medication to see if it has either a +ve or -ve effects on intelligence, or no effect at all. A sample of 30 participants who have taken the medicine has a mean of 140 with a std of 20. did the medication affect intelligence? C.I. \Rightarrow 95%.

Ans $\mu = 100$, $\sigma = 20$, $n = 30$, $\bar{x} = 140$, C.I. \Rightarrow 95%, $\alpha = 0.05$

① Null hypothesis (H_0) $\rightarrow \mu = 100$

② Alternate " (H_1) $\rightarrow \mu \neq 100$ { 2 Tail test }

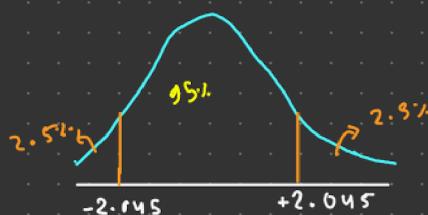
only apply when the value of a mean is either $\uparrow\downarrow$ OR $\uparrow\downarrow$

③ $\alpha = 0.05$

④ Degree of freedom :-

$$d.f \rightarrow 30 - 1 \rightarrow \underline{\underline{29}}$$

⑤ Decision boundary :-



by using t-table,

$$d.f = 29$$

$$\alpha = 0.05$$

we got \Rightarrow +2.045

if t-test is less than -2.045 and greater than 2.045 , we reject the null hypothesis.

⑥ Calculate t-test statistic,

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \Rightarrow \frac{140 - 100}{\frac{20}{\sqrt{30}}} \Rightarrow \frac{40}{3.65} \Rightarrow \underline{\underline{10.96}}$$

⑦ Conclusion

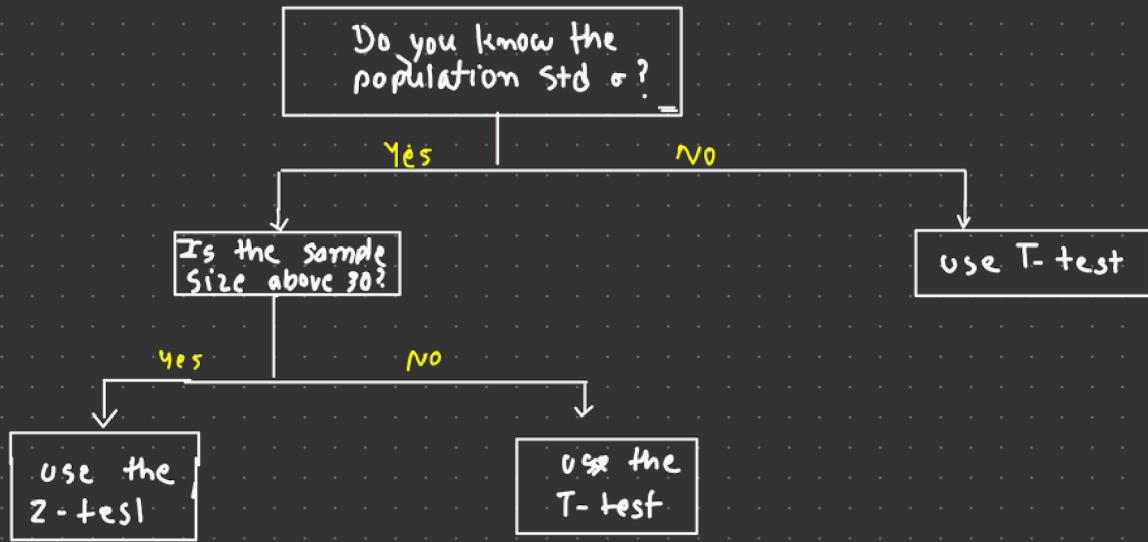
Decision Rule :- if t is less than -2.045 and greater than 2.045 we reject the null hypothesis.

$$t = 10.96 > 2.045 \quad \{ \text{we Reject the null hypothesis} \}$$

Final → Medication has increased the intelligence.

★ Interview Questions

When to use t-test Vs Z-test



* Types 1 and Type 2 Errors

Reality : Null hypothesis is true OR H_0 is false

Decision : H_0 is True OR H_0 is False.

Conclusion

↓

* outcome : 1

we reject the H_0 (Null hypothesis) when in reality it is false \rightarrow Good

* outcome : 2

we reject the null hypothesis when in reality it is True \rightarrow True \rightarrow Type 1 Error

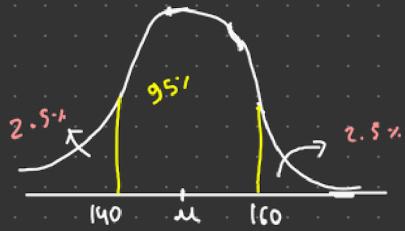
* outcome : 3

we retain (accept) the H_0 , when in reality it is false \rightarrow Type 2 Error

* outcome 4:

we retain the H_0 , when in reality it is True
 \downarrow
Good

Confidence Interval and Margin of Error:



$$\mu \rightarrow 160$$

$$C.I. \rightarrow 95\%$$

$$140 \leftarrow 160$$

Point estimate: A value of any statistics that estimate the value of an unknown population parameter is called "point estimate".

$$\bar{x} \rightarrow \mu$$

Confidence interval: We construct a C.I. to help the estimate what the actual value of the unknown population mean is:-

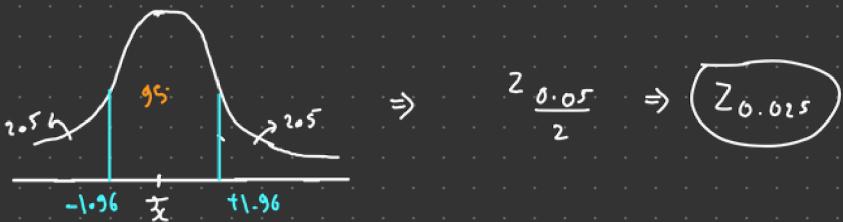
Point estimate \pm Margin Error

Z-test

$$\bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

$\{\alpha = \text{Significant value who decided}$
 How much C.I. is

$$\alpha = 0.05$$



Problem: On a verbal section of CAT Exam, the standard deviation is known to be 100. A sample of 25 test taken has a mean of 520. Construct a 95% C.I. about the mean.

Ans $\bar{x} = 520, \sigma = 100, n = 25, \text{C.I. } \rightarrow 0.95, \alpha = 0.05$



$$\bar{x} \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Follow:

KRISHAN KUMAR

$$\rightarrow \text{Lower C.I.} = 520 - (1.96) \times \frac{100}{\sqrt{25}} = 480.8$$

$$\rightarrow \text{Higher C.I.} = 520 + (1.96) \times \frac{100}{\sqrt{75}} = 559.2$$



\rightarrow Conclusion: I am 95% sure (confident) that the mean CAT exam score lies between 480.8 and 559.2

★ Bayes theorem



probability $\left\{ \begin{array}{l} \text{Independent Event} \\ \text{Dependent Event.} \end{array} \right.$

* Independent Event

ex. Rolling a dice

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

$P(1) \rightarrow \frac{1}{6}$ ex, Tossing a coin

$$\vdots$$

$$P(6) \rightarrow \frac{1}{6}$$

$$P(T) \rightarrow \frac{1}{2}$$

* Dependent Event

ex. Blue balls

0	0	0	$\rightarrow P(R) \rightarrow \frac{2}{5}$
0	0	$\rightarrow P(B) \rightarrow \frac{3}{4}$	

Here one event effect another events also.

ex,

$$P(R \text{ and } B) = P(R) \times P(B/R)$$

$$\frac{2}{5} \times \frac{3}{4} \rightarrow \frac{6}{20}$$

Follow:

$$P(A \text{ and } B) = P(B \text{ and } A)$$

$$P(A) \times P(B/A) = P(B) \times P(A/B)$$

$$P(B/A) = \frac{P(B) \times P(A/B)}{P(A)}$$

Bayes theorem

$P(A/B) \rightarrow$ Conditional Probability

OR

$$P(A/B) = \frac{P(A) \times P(B/A)}{P(B)}$$

A, B → Events

$P(A/B)$ → Probability of A given B is already happened

$P(B/A)$ → Pro. of B given A is true.

$P(A), P(B)$ → The independent Pro. of A and B

* Uses of Bayes theorem in ML : (ex)

Dataset / Predict the price $\rightarrow (y=?)$

Size of the House	No. of Room	Location	Price
x_1	x_2	x_3	y

By Bayes theorem →

$$P\left(\frac{y}{(x_1, x_2, x_3)}\right) = \frac{P(y) \times P\left(\frac{x_1, x_2, x_3}{y}\right)}{P(x_1, x_2, x_3)}$$

Analysis by using bayes theorem → Bayesian Statistics

(() CHI SQUARE-TEST :-

It is non parametric test that

performed categorical {ordinal, nominal} data.

→ CHI SQUARE TEST for Goodness of fit test claims about population proportions [categorical variance].

ex ①

	Theory	Sample	Applied Goodness Fit
yellow bike	1/3	22	
orange bike	1/3	17	
Red bike	1/3	59	
	→ Observed categorical distribution		
	Theory categorical distribution		

ex ② Goodness of fit test :

In a student's class of 100 Student where 30 are right hand. Does this class fit the theory 12% of people are right handed.

→

	O	E	O → Observed value E → Expected value
Right handed	30	12	
Left handed	70	88	$\frac{12}{100} \times 100$
Sample info	$\frac{70}{100}$	$\frac{88}{100}$	Theory categorical distribution

Follow:

KRISHAN KUMAR

Problem: In 2010 Census of city, the weight of the individuals in a small city were found to be the following.

$\leq 50\text{kg}$	50 - 75	> 75
20%	30%	50%

In 2020, age of $n = 500$ individuals were sampled. Below are the results.

<50	50-75	>75
140.0	160	200

using $\alpha = 0.05$ would you conclude the population difference of weights has changed in the last 10 years.

Ans:

Expected Value	<50	50-75	>75	(2010)
	20%	30%	50%	

$$n = 500$$

Observed Value	<50	50-75	>75	(2020)
	140	160	200	

<50	50-75	>75
0.2×500	0.3×500	0.5×500
$\Rightarrow 100$	$\Rightarrow 150$	$\Rightarrow 250$

$$20\% \text{ of } 500 \rightarrow \frac{20 \times 500}{100}$$

$$30\% \text{ of } 500 \rightarrow 0.3 \times 500$$

$$50\% \text{ of } 500 \rightarrow 0.5 \times 500$$

① Null hypothesis: H_0 : The data meet the expectation
 KRISHAN KUMAR
 Alternate Hypothesis: H_1 : The data does not meet the expectation.

$$② \lambda = 0.05, C.I. \rightarrow 95\%$$

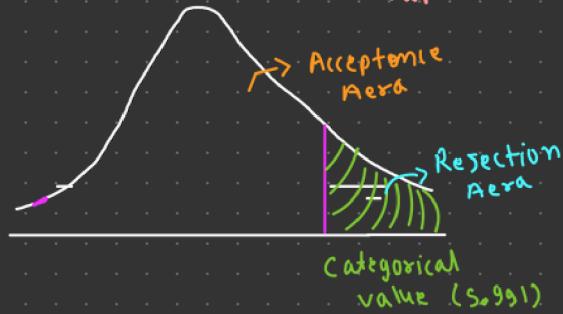
③ Degree of freedom

$$d.f = k - 1$$

$k \rightarrow$ No. of category

$$3 - 1 \rightarrow 2$$

④ Decision Boundary : We used CHI-Square table for the categories value.



denote:

$\chi^2 \rightarrow$ CHI Square (Sumbal)

If χ^2 is greater than 5.991, Reject H_0

else:

We fail to Reject the H_0

⑤ Calculate CHI-SQUARE Test-Statistic :

$$\chi^2 = \frac{\sum (O - E)^2}{E}$$

$O \rightarrow$ observed, $E \rightarrow$ expected value

E value

<50	50-75	>75
140	160	200

O value

<50	50-75	>75
100	150	250

$$\Rightarrow \frac{140}{100} + \frac{10}{150} + \frac{250}{250}$$

$$\Rightarrow \frac{160}{100} + \frac{100}{150} + \frac{2500}{250}$$

$$\Rightarrow 1.6 + 0.66 + 10 \Rightarrow 26.66$$

Final: If χ^2 greater than 5.99 Reject the H_0 , else Fail to reject the H_0 .

So, $\chi^2 = 26.66 > 5.99$ { we Reject the Null Hypothesis }

Cons: The weights of 2020 population are different than those expected on the 2010 population.

F-distribution:-

In probability theory and statistic the F-distribution or F-ratio also known as snedecor's F-dis. or the Fisher - snedecor distribution is continuous probability distribution that arises frequently as the null dis. of a test test statistic, most notably in the analysis of variance (ANOVA) and other F-test.

* F-test used to compare variance of mean between two group. F-test also called 'variance ratio test'

parameters:

$d_1, d_2 > 0$ (degree of freedom)

Support: $x \in (0, +\infty)$

F-distribution with d_1 and d_2 degree of freedom is the dis. of

$$\chi^2 = \frac{s_1^2/d_1}{s_2^2/d_2}$$

$\begin{cases} s_1 \rightarrow \text{Independent random variable} \\ s_2 \rightarrow \dots \end{cases} \quad \begin{cases} \{\text{CHI Square distribution}\} \\ \dots \end{cases}$

$d_1 \rightarrow \text{Degree of freedom}$
 $d_2 \rightarrow \text{Degree of freedom}$

F-test:

Problem: The following data show the no. of bulbs produced daily for same day by 2 workers A and B.

A	B
40	39
30	38
38	41
41	33
38	32
35	39
40	
34	

Can be consider based on the data
worker B is more stable and
efficient

$$d = 0.05$$

Ans → ① Null Hypothesis: $H_0: \sigma_1^2 = \sigma_2^2$
 $H_1: \sigma_1^2 \neq \sigma_2^2$

② Calculation of variance:

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

A		
x_i	\bar{x}	$(x_i - \bar{x})^2$
40	37	9
30	37	49
38	37	1
41	37	16
38	37	1
35	37	4
$\bar{x} = 37$		
$\sum (x_i - \bar{x})^2$		
$\Rightarrow 80$		

$$S_1^2 \Rightarrow \frac{(x_1 - \bar{x})^2}{n-1}$$

$$S_1^2 \Rightarrow \frac{80}{6-1} \Rightarrow \frac{80}{5} \Rightarrow 16$$

B		
x_i	\bar{x}	$(x_i - \bar{x})^2$
39	37	4
38	37	1
41	37	36
33	37	16
32	37	25
39	37	4
40	37	9
34	37	9
$\sum (x_i - \bar{x})^2$		
$\Rightarrow 84$		

$$S_2^2 = \frac{(x_2 - \bar{x})^2}{n-1}$$

$$S_2^2 = \frac{84}{8-1} \Rightarrow \frac{84}{7} \Rightarrow 12$$

→ Calculation of variance Ratio (F-test):

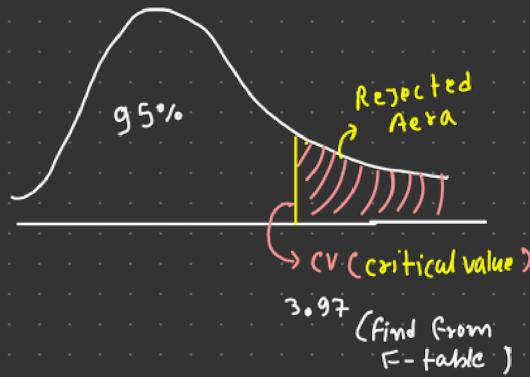
$$F = \frac{s_1^2}{s_2^2} \Rightarrow \frac{13}{12} \Rightarrow 1.33$$

③ Decision Rule:

$$df_1 \rightarrow 6-1 \rightarrow 5$$

$$df_2 \rightarrow 8-1 \rightarrow 7$$

$$\alpha \rightarrow 0.05$$



If F-test greater than 3.97,
Reject the Null Hypothesis

$1.33 < 3.97$, {so, we Reject the H_0 }

Final conclusion :-

Worker B is not efficient when compare to worker A.



(ANOVA) Analysis of variance:

Anova is a statistical method to compared the mean of 2 or more group.

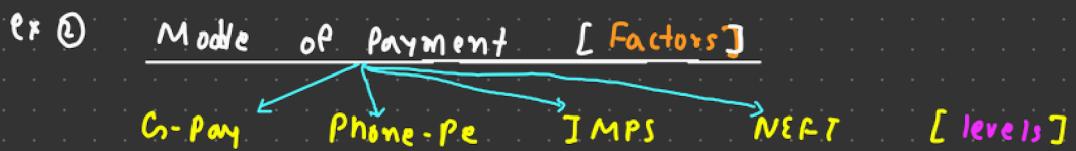
ANOVA:

① factors (variable)

② levels

Ex ① Factors = Medicine

level = 5mg 10 mg 15mg [Dosage]



⇒ Assumptions in ANOVA:

- ① Normality of Sampling Distribution of mean The dis. of Sample mean is normally distributed.
- ② Absence of outliers. Outlier score need to be Remove from dataset.
- ③ Homogeneity of variance Each one of the population has same variance $[\sigma_1^2 = \sigma_2^2 = \sigma_3^2]$ Population variance in different levels of each independent variable are equal.
- ④ Sample are independent and Random.

⇒ Types of ANOVA:

- ① One way ANOVA
- ② Repeated Measured ANOVA
- ③ Factorial ANOVA

- ① One way ANOVA: One factor with at least 2 levels, these levels are independent.

② Repeated Measured anova:

One factor with at least 2 levels, levels are dependent.

③ Factorial Anova: Two or more factor (each of which with after 2 level). levels can be either dependent or independent.

⇒ Hypothesis testing in ANOVA:

Null Hypothesis: $H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_n$

$H_1:$ At least one of the mean is not equal

$$X \boxed{\mu_1 = \mu_2 = \mu_3 = \dots = \mu_n}$$

Test statistic:

$$F = \frac{\text{Variation between sample}}{\text{Variation within sample}}$$

$X_1 \quad X_2 \quad X_3$

1 6 5

2 7 6

4 3 3

5 2 2

3 1 4

$$\sum x_1 = 15$$

$$\bar{x} = \frac{15}{3} \Rightarrow 3$$

$$\sum x_2 = 14$$

$$\bar{x}_2 = \frac{14}{5}$$

$$\sum x_3 = 20$$

$$\bar{x}_3 = \frac{20}{5} \Rightarrow 4$$

$$H_0: x_1 = x_2 = x_3$$

$$H_1: \text{at least one sample mean is not equal}$$

* One way anova: One factor at least 2 levels, levels are independent

Problem: Doctor want to test a new medication which reduce headache. They split the participant into 3 condition [15mg, 30mg, 45mg]. later on the doctor ask the patient to rate the headache between [1 → 10] are there any difference between 3 condition using Alpha (α) $\rightarrow 0.05$?

Ans:

① Define null Hypothesis:

$$H_0: \mu_{15} = \mu_{30} = \mu_{45}$$

H_1 : not all μ 's are equal

15mg	30mg	45mg
9	7	4
8	6	3
7	6	2
8	7	3
8	8	4
9	7	3
8	6	2

② State significant value:

$$\alpha = 0.05$$

$$C.I. \rightarrow 0.95$$

③ Calculate degree of freedom:

$$N = 21, a = 3, n = 7$$

(a → Number of Sample)

$$df_{\text{between}}: a-1 \rightarrow 3-1 \Rightarrow 2$$

(2, 18)

$$df_{\text{within}}: N-a \rightarrow 21-3 \Rightarrow 18$$

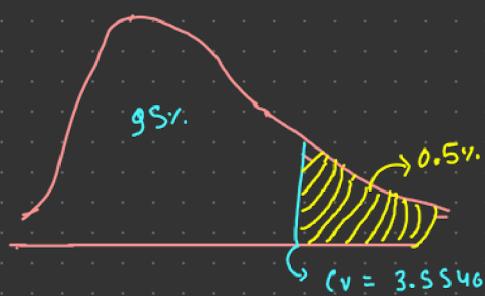
super useful to finding critical value in table

$$df_{\text{total}}: N-1 \rightarrow 21-1 \Rightarrow 20$$

F - Table

Follow:

KRISHAN KUMAR



Critical value $\rightarrow 3.5546$

④ State decision rule:

If F is greater than 3.5546 , we reject the H_0 .

⑤ Calculate test Statistic:

SS = Sum of Square

SS_{between} SS_{within} SS_{total}

15 mg	30 mg	45 mg
9	7	4
8	6	3
7	6	2
8	7	3
8	8	4
9	7	3
8	6	2

$$① SS_{\text{between}} = \frac{\sum (\sum a_i)^2}{n} - \frac{-2}{N}$$

$$15 \text{ mg} \rightarrow 9+8+7+8+8+9+8$$

$$30 \text{ mg} \rightarrow 7+6+6+7+8+7+6$$

$$45 \text{ mg} \rightarrow 4+3+2+3+4+3+2$$

$$SS_{\text{between}} = \frac{57^2 + 47^2 + 21^2}{7} - \frac{[57 + 47 + 21]}{21}$$

$$\Rightarrow 98.17$$

Follow:

KRISHAN KUMAR

$$\textcircled{2} \quad SS_{\text{within}} = \sum y^2 - \frac{\sum (a_i)^2}{n}$$

$\left\{ \begin{array}{l} y^2 \rightarrow \text{Given table value } \vec{A} \\ \vec{A} \text{ element का square करना} \end{array} \right.$

$$\Rightarrow \sum y^2 - \left[\frac{s_7^2 + u_7^2 + z_7^2}{7} \right]$$

$$\sum y^2 \Rightarrow 9^2 + 8^2 + 7^2 + 8^2 + \dots + 3^2 + 2^2 \Rightarrow 853$$

$$\Rightarrow 853 - \left[\frac{s_7^2 + u_7^2 + z_7^2}{7} \right]$$

$$\Rightarrow 10.29$$

$$\textcircled{3} \quad SS_{\text{Total}} : \quad \sum y^2 - \frac{T^2}{N}$$

$$\Rightarrow 853 - \frac{125^2}{21}$$

$$\Rightarrow 108.95$$

	SS	df	MS	F
Between	98.67	2	49.34	
Within	10.29	18	0.54	
Total	108.95	20	49.88	$\{ MS = \frac{SS}{df} \}$

$\Rightarrow \frac{MS_{\text{between}}}{MS_{\text{within}}} \text{ is equal to } \Rightarrow F = \frac{\text{variation between sample}}{\text{variation within sample}}$

$$\Rightarrow \frac{49.34}{0.54} \Rightarrow 86.56 \Rightarrow F = 86.56$$

Final:

If F is greater than 3.5546, we Reject the H₀

86.56 > 3.5546, so we Reject H₀.

