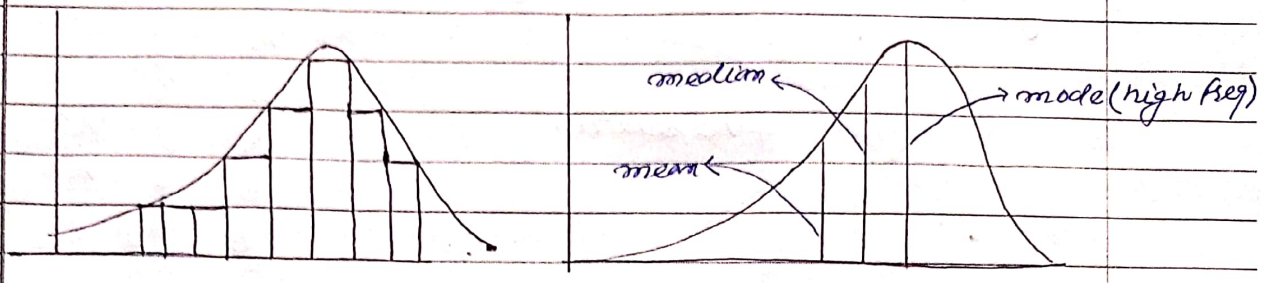
Important *
Example e.g.
Visit ⊕
Points ⇒

Title:
Sub Title:
Highlighted:

Page No.....
Date ...../...../.....

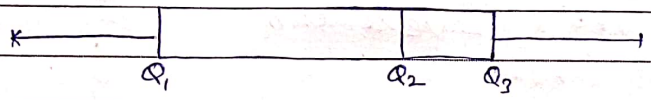
③ Left Skewed Distribution



- Left skewed also called negative skewed
- Relation between mean, median, mode

* mean ≤ media ≤ mode

Boxplot



$Q_2 - Q_1 \geq Q_3 - Q_2$

Covariance and Correlation

Both are using to solve "what is the relationship between x and y?"

Correlation (co-relation between x and y)

x	y
2	3
4	5
6	7
8	9

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

} find relation between $x \in$ and $y \in$

There is two types of relation are possible

(i) $x\uparrow\ y\uparrow$, $x\downarrow\ y\uparrow$ or (ii) $x\downarrow\ y\uparrow$, $x\uparrow\ y\downarrow$

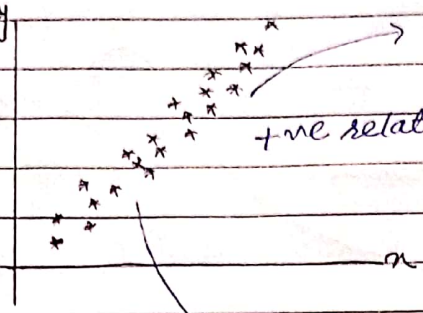
Note ( ):
Note ( ):

Important *
Example e.g
Visit ⊕
Points ⇒

Title:
Sub Title:
Highlighted:

Page No.....
Date ...../...../.....

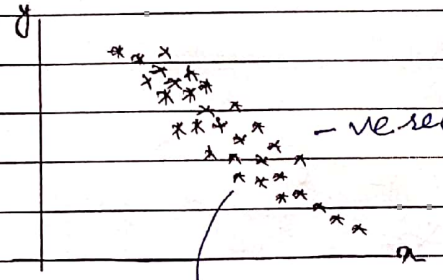
① If [$x\uparrow\ y\uparrow$, $x\downarrow\ y\downarrow$]



x	y	
2	3	$x\uparrow\ y\uparrow$
4	5	$x\downarrow\ y\downarrow$
6	7	
8	9	

$x\uparrow\ y\uparrow$, $x\downarrow\ y\downarrow$ } In this relation scatter plot looks like this shape

② If [$x\uparrow\ y\downarrow$, $x\downarrow\ y\uparrow$]



x	y	
5	4	$x\uparrow\ y\downarrow$
6	5	$x\downarrow\ y\uparrow$
7	9	
9	11	

$x\uparrow\ y\downarrow$, $x\downarrow\ y\uparrow$ } In this correlation scatter plot looks like this

*Scatter plot can be used to find correlation between two random continious variable.

↓Size of House ↑ ↑Price of House↓

This type of relation is a high relation and strong fetcher

*Correlation value range in between −1 to +1.
The value more towards +1 they are more positively corr.
The value more towards −1 the more negatively correlated

$$P_{x,y} = \frac{cov(x,y)}{\sigma x \sigma y}$$

P=.1	−1<P<0	P=+1	P=0	0<P<+1

Note ( ):
Note ( ):

Important * Example e.g. Visit ⊕ Points ⇒

Title: Sub Title: Highlighted

Page No. Date / /

Coveriance

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

x_i ⟶ Datapoints of x
$\bar{x}$ ⟶ Sample mean of x
y_i ⟶ Data points of y
$\bar{y}$ ⟶ Sample mean of y
n ⟶ x or y data size

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

(i) $Cov(x,y)$ $x\uparrow$ $y\uparrow$ / $x\downarrow$ $y\downarrow$ ⟶ +ve correlation

(ii) $Cov(x,y)$ $x\downarrow$ $y\uparrow$ / $x\uparrow$ $y\downarrow$ ⟶ -ve correlation

Eg:

x	y
2	3
4	5
6	7

$\bar{x} = 4$ $\quad$ $5 = \bar{y}$

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

$$= \frac{[(2-4)(3-5) + (4-4)(5-5) + (6-4)(7-5)]}{n-1}$$

$$= \frac{4+0+4}{2} = 8/2 = 4 \quad \text{+ve coveriance}$$

x and y having a +ve coveriance

Advantages	Disadvantages
Relationship between x and y : +ve or -ve value	Coveriance does not have a specific limit value
Covariance can be vary between $[-\infty$ and $+\infty]$	$[-\infty$ and $+\infty]$

You can fix this Problem by using
① Person Coveriance Coefficient
② Spearman's Coveriance Coefficient

Note ( ):
Note ( ):

Important * Example e.g Visit ⊕ Points ⇒ | Title: Sub Title: Highlighted: | Page No. Date ../../..

Person Coversance Coefficient

$$\int_{x,y} = \frac{Cov(x,y)}{\sigma_x \sigma_y} \qquad \text{Range } [-1 \text{ to } 1]$$

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

Spearman Rank Correlation Coefficient :

$$\rho_S = \frac{Cov(R(x), R(y))}{\sigma(R(x)) * \sigma(R(y))} \qquad \text{Range } [-1 \text{ to } +1]$$

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

x more right R(x) more less

* x value more high R(x) more less from 8 to 1

Feature Selection

+ve	+ve	+ve	-ve	
Size of House ↑	No of Rooms ↑	Location ↑	Haunted	Price ↑

① Get only important feature

Correlation and Covariance by using Python

```
df = sns.load_dataset('healthexp')
df.head() ; df.cov()
np.cov(df['columnname'])
```

```
df.corr()
```

Note ( ):
Note ( ):

Important * Example e.g. Visit ⊕ Points ⇒

Title:
Sub Title:
Highlighted:

Page No.....
Date ...../...../.....

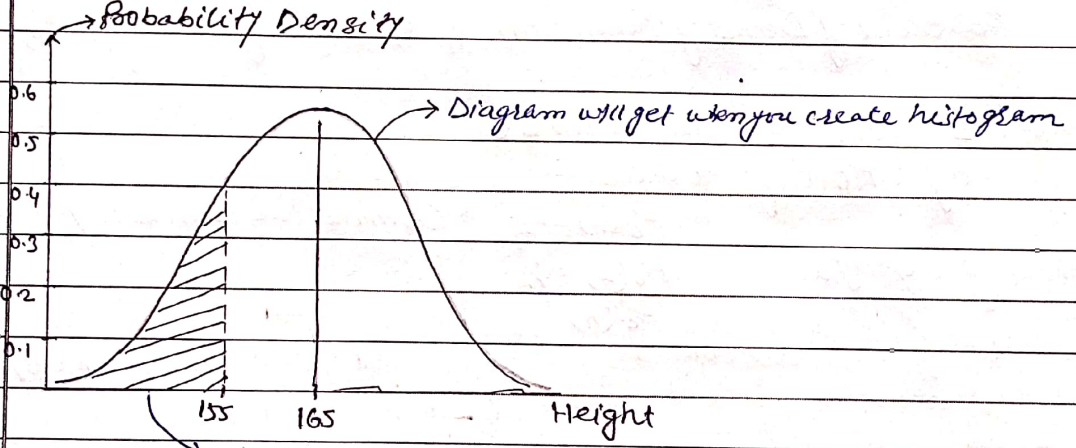
Probability Distribution Function

→ It denotes distribution of a data

Probability Density Function (pdf)	Probability Mass Function (pmf)
Continious Random Variable	Discrete Random Variable
Range [0, 1]	$[-\infty, +\infty]$

Probability Density Functions: It is also called Continious Probability Distributions

Eg: Height of students in classroom
lets asume $\bar{x} = 165$



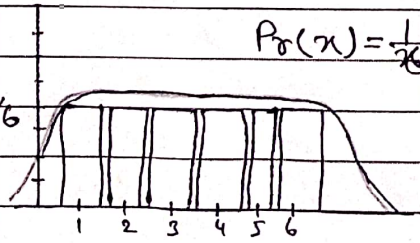
$P_r(x \leq 155)$ = Area under the curve

Kernel Density Estimator smooth histogram to draw a line plot

Probability Mass Function

Show Distribution for Discrete Random Variable

Eg: Rolling a Dice $\{1, 2, 3, 4, 5, 6\}$



$P_r(x) = \frac{1}{6}$ (according to dice)

$P_r(1) = \frac{1}{6}$

$P_r(5) = 1/6$

Problem

$P(x \leq 4) = P_r(x=1) + P_r(x=2) + P_r(x=3) + P_r(x=4)$

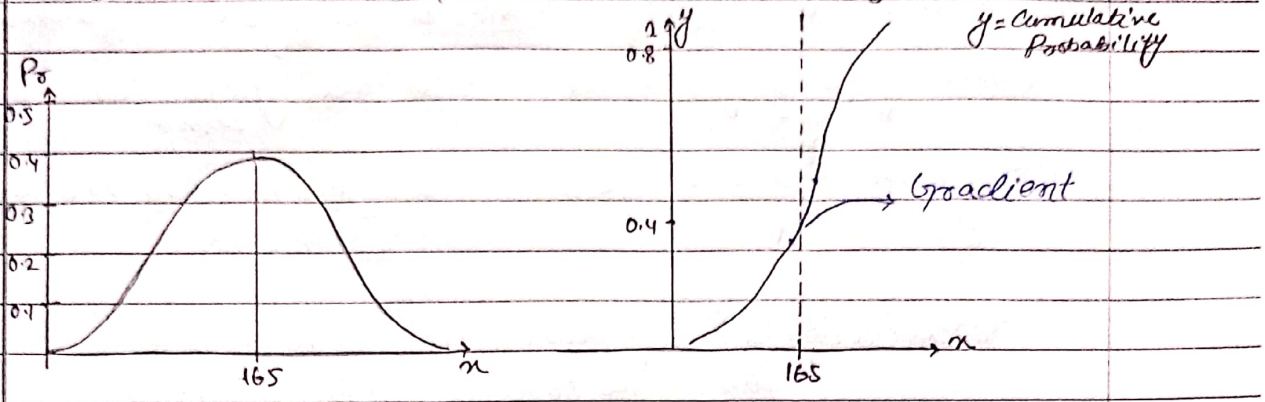
$= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = 2/3$

3 Cumulative Distribution Function (CDF)

Note ( ):
Note ( ):

Important * Example e.g Visit ⊕ Points ⇒
Title:
Sub Title:
Highlighted:
Page No.....
Date ...../...../.....

Cumulative Distribution Function (cdf)



Probabity Density Function & Probability Mass Function

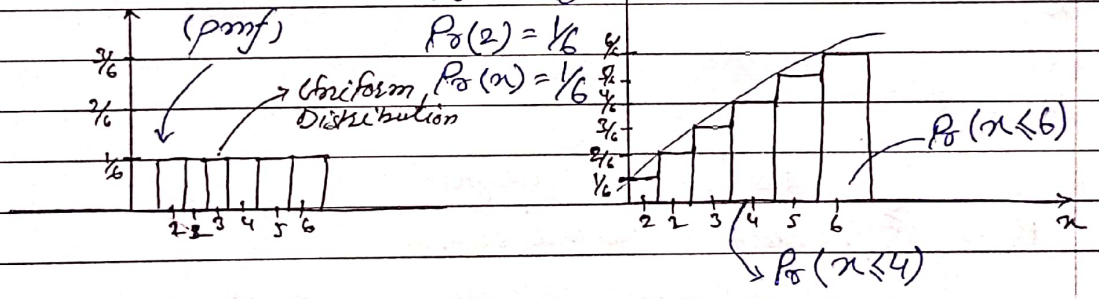
PMF: Discrete Random Variable

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

$P_\sigma(1) = 1/6$

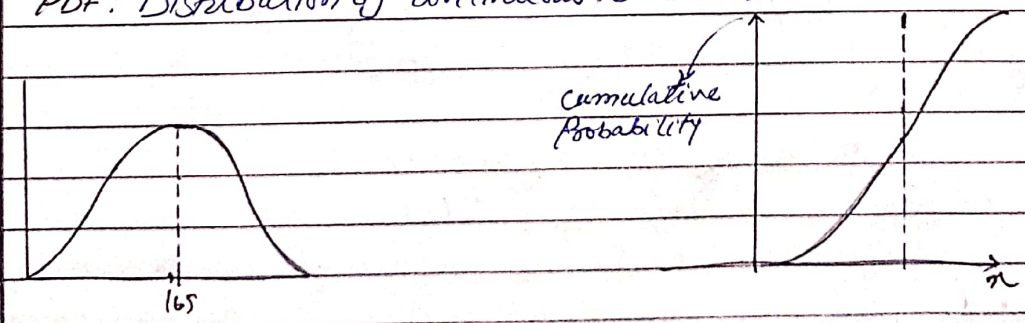
$P_\sigma(2) = 1/6$

$P_\sigma(x) = 1/6$



$$P_\sigma(x \leq 2) = P_\sigma(x=1) + P_\sigma(x=2)$$
$$= 1/6 + 1/6 = 2/6 = 1/3$$
$$P_\sigma(x \leq 6) = 1$$

PDF: Distribution of continuous random variable



Probability density → Gradient of cumulative curve

Note ( ):
Note ( ):

Important * Example e.g. Visit ⊕ Points ⇒

Title:
Sub Title:
Highlighted:

Page No.....
Date ...../...../.....

Types of Probability Distribution

1. Normal / Gaussion Distribution
2. Bernoulli Distribution
3. Uniform Distribution
4. Poission Distribution
5. Log Normal Distribution
6. Binomial Distribution

1. Bernoulli Distribution

In probability theory and statistics, the Bernoulli distribution, named after swiss mathematican Jacob Bernoulli,

(i) is the discrete probability distribution of a random variable.

(ii) which has two outcomes Y and N also known as James or Jacques.

(iii) Or outcomes Binary

Eg: Tossing a coin {H, T}

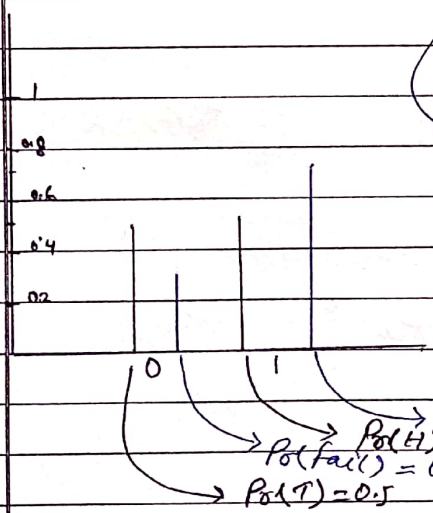
Eg: Whether the person will pass or fail

$P_0(H) = 0.5 = P$

$P_0(T) = 0.5 = 1 - P = q$

$P_0(Pass) = 0.7 = P$

$P_0(fail) = -0.7 + 1 = 0.3 = q$



$P(x=0) = 0.5$ and $P(x=1) = 0.5$

$P(x=0) = 0.3$ and $P(x=1) = 0.7$

Parameters

$0 \leq P \leq 1$

$q = 1 - P$

$k = \{0, 1\}$

0 = fail, 1 = success

PS Pmf

P

Note ( ): Discrite has finite value from 0, 1, 2 ...

Note ( ):

Important * Example e.g Visit @ Points ⇒ | Title: Sub Title: Highlighted: | Page No. Date ../../..

Proof for Bernoulli Distribution

Proof → $P(X=x)=P(x)=\begin{cases} p^x(1-p)^{1-x}; & x=0,1 \ (x \text{ belongs to } 0 \text{ and } 1) \\ 0; \text{ other wise} \end{cases}$

where p = Probability of success

$1-p=q$ = Probability of failure

number of trial = 1 (trial → Experiment)

number of trial = one { Success (p); failure ($1-p$)

If $x=1 \Rightarrow P(x)=p^x(1-p)^{1-x}$

$= P(1)=p^2(1-p)^0$

$= P(1)=p$ #Probability of Success

If $x=0 \Rightarrow P(x)=p^x(1-p)^{1-x}$

$P(0)=p^0(1-p)^{1-0}$

$P(0)=1(1-p)^1$

$P(0)=1-p=q$

* As you know $(x)^0=(x)^1=1$

Verify is this Proof or not, If proof $\sum_{i=0}^{1} P(x)=1$

Step $\sum_{i=0}^{1} p(x)=$

*

Proof { $q=1-p$ if $k=0$: $\sum_{i=0}^{1} p(0)=p^0(1-p)^{1-0}=1-p=q$; p if $k=1$: $\sum_{i=0}^{1}(p(1))=p^1(1-p)^{1-1}=p$

$k=x$

$\therefore \sum_{i=0}^{1} p(x)=1-p+p=1$

Moment Generating function.

$M_x(t)=\sum_{x=0}^{1} e^{tx}p(x \Rightarrow M_x(t)=\sum_{x=0}^{1} e^{tx}p^x(1-p)^{1-x}$

If $x=0$

$M_x(t)=\sum_{x=0}^{1} e^{tx}p^x(1-p)^{1-x}=e^0p^1(1-p)^{1-0}=1-p$

If $x=1$

$M_x(t)=\sum_{x=0}^{1} e^{tx}p^x(1-p)^{1-x}=e^tp(1-p)^0=pe^t$

Now $1-p+pe^t=q+e^tp$

Note ( ):

Note ( ):

Mean of Bernoulli Distribution:

$$E(K) = \sum_{i=0}^{K} K \cdot P(K)$$

$K = x$

Mean → E

lets $P_B(K=1) = 0.6$

$P_B(K=0) = 0.4$

$= [0 * 0.4 + 1 * 0.6]$

$= 0.6 = p$

It means the mean of bernoulli distribution be always p, $q = 0$ does not has any value.

Median of Bernoulli Distribution

$$\text{median} \begin{cases} 0 & \text{if } p < 1/2 \\ [0,1] & \text{if } p = 1/2 \\ 1 & \text{if } p > 1/2 \end{cases}$$

Variance: $Var = p * (1-p)$

Standard division: $std = \sqrt{pq}$

Binomial Distribution

In probability theory and statistics, the binomial distribution with parameters n and p is the discrete probability distribution of the number of success in a sequence of n independent expriments, each asking question < yes → success (with probability p); no → failure $q = 1-p$

(number of success; all expresment numbers)

A single success/failure experiment is also called a Bernoulli trial or Bernoulli Experiment.

A sequence of outcomes is called a Bernoulli process for a single trial; i.e $n=1$, the binomial distribution is a Bernoulli Destribution (in case of single experiment).

① Every experiment outcomes < yes, no

② Then experiment is performed → n trials

Note ( ):
Note ( ):

Important * Example e.g Visit ⊕ Points ⇒

Title:
Sub Title:
Highlighted:

Page No.
Date / /

Notation : $B(n, P)$ — B for binomial distribution

Parameters: $n \in \{0, 1, 2, 3, 4, \ldots\}$ → Numbers of trial / experiment

$P \in [0, 1]$ → Success probability for each trial

$q = 1 - P$

↳ It is a fixed number of experiment

Eg: Tossing a coin 10 times

P = Probability of success/failure

n = numbers of experiment ↳ It is fixed or countable

Support: $n \neq K$

$K \in \{0, 1, 2, 3, \ldots, n\}$ → Number of success

Pmf $P_r(K, n, P) = {}^{n}C_K \, P^K (1-P)^{n-K}$

for $K = 0, 1, 2, 3, 4, \ldots n$ where ${}^{n}C_K = \frac{n!}{K!(n-K)!}$

Binomial Distribution
$X = 1, 2, 3, 4, \ldots n$
no of trials, n = finite
Pr. of success, P = constant

Eg: 35% of all voters support Proposition A. If a random sample of 10 voters is polled what is the probability that exactly three of them support the proposition

find $P(x=3)$ if $n=10$ and $P=0.35$

$$P(x=3) = \frac{n!}{x!(n-x)!} p^x q^{n-x} = \frac{10!}{3!7!} (0.35)^3 (0.65)^7$$
$$= 0.2522$$

Ans: 25.22%.

Mean : Variance : Stdl

mean $= np$ var $= npq$ Stdl $= \sqrt{npq}$

Proof mean of binomial distribution

Note ( ):
Note ( ):

Important *
Example e.g
Visit Φ
Points ⇒

Title:
Sub Title:
Highlighted:

Page No.....
Date ...../...../.....

Note ( ):
Note ( ):

Important *
Example e.g.
Visit ⊕
Points ⇒

Title:
Sub Title:
Highlighted:

Page No.
Date ../../..

Proof mean of binomial Distribution

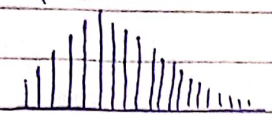
Properties of Binomial Distribution

1. The number of observation n is fixed
2. Each observation is independent
3. Each observation represents one of two outcomes ("success" or "failure").
4. The probability of "success" p is the same for each outcomes.
5. No previous event impact on next event.

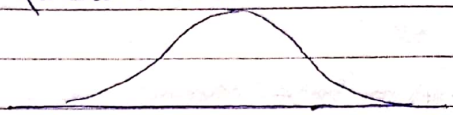
Note ( ):
Note ( ):

Random Variables.

Probability Distribution of a Discrete Random Variable



Probability Distribution of a Continuous Random Variable



Poisson Distribution

In statistics, Poission distribution is a probability distribution used to show how many times an event is likely to occur over a specified period.

In a poison distribution number of expriement is very large (∞) but number is success is very low.

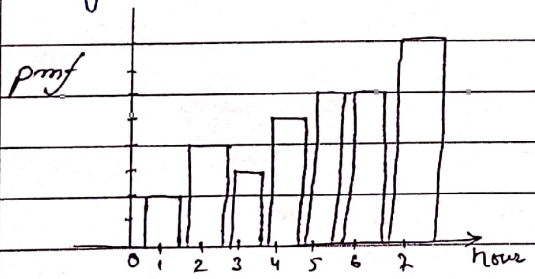
$X = 1, 2, 3, \ldots \infty$
number of trial, $n \to$ (very large)
Prob. of success, $P \to 0$ (very low)

Eg: Lic holder
number of policy holder is very high $= \infty$
but number of policy holder death is very low

Eg2: Numbers of word on book page is very high $= \infty$
but number of miss print word in book very low

*Describe number of event occuring in a fix time inperval

Eg1: Number of people visiting hospital every hour



$\lambda = 3$

lambda is a expected number of events occur at every time inperval.

Question based on graph.

What is the probability of a person to come (in bank) at the fifth hour?

Note ( ):
Note ( ):

Important *
Example e.g.
Visit ⊕
Points ⇒

Title:
Sub Title:
Highlighted:

Page No.
Date / /

Poss

$$P(x=5) = \frac{e^{-\lambda} \lambda^{x}}{x!}$$

x = (this time interval)

lets $\lambda = 3$

$$P(x=5) = \frac{e^{-3} 3^{5}}{5!} = 0.101 \simeq 10\%$$

Question 2: What is the probability for a person to comes less than or equale to fifth hour

Some

$P_o(x \leq 5)$

$P_o(x=1) + P_o(x=2) + P_o(x=3) + P_o(x=4) + P_o(x=5)$

Mean:

Variance:

mean $= E(x) = \mu = \lambda * t$

Expected of x

λ = is a expected number of events

at every time interval

t = time interval

Note ( ):
Note ( ):

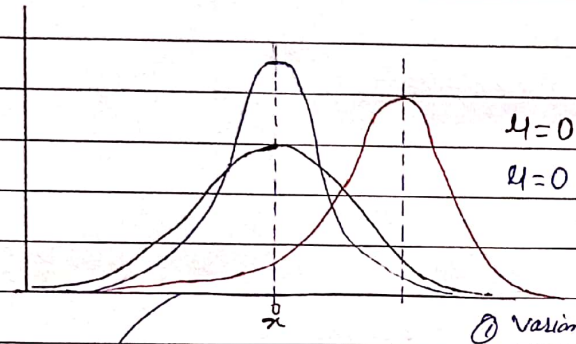
Normal / Gaussion Distribution

In statistics, a normal distribution or Gaussion Distribution is a

→ type of continuous probability distribution
→ for a real-valued random variable
→ and a bell-shaped curve
→ with an equale number of measurements above and below the mean value

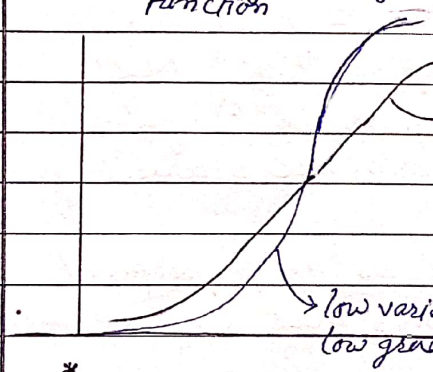
mean = median = mode

Continuous random variable = pdf



Probability density function

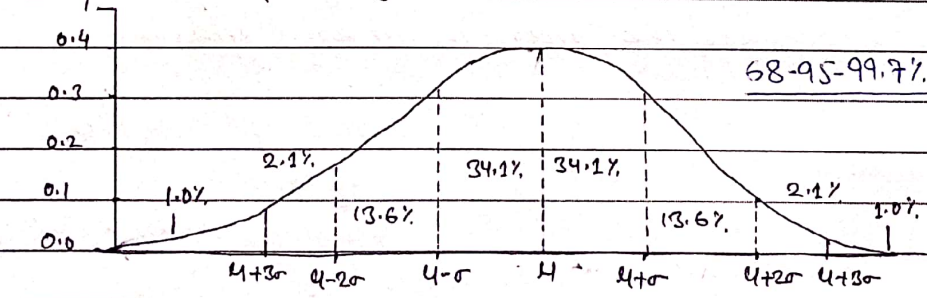
① Variance (spread) will increase if std increase
If varience = x and std = y
then x↑ y↕ } Positive
x↓ y↓ } Relation



② Relation between variane and gradient
If varience = x and gradient = y
then x↑ y↑ } Positive Relation
x↓ y↓

Area under the curve is always = 1

Empirical Rule of Normal Distribution:



$P_r(\mu-\sigma \leq x \leq \mu+\sigma) \approx 68\%$ 1st std
$P_r(\mu-2\sigma \leq x \leq \mu+2\sigma) \approx 95\%$ 2nd std
$P_r(\mu-3\sigma \leq x \leq \mu+3\sigma) \approx 99.7\%$ 3rd std
} Data Distribution.

Note ( ):
Note ( ):

Title:
Sub Title:
Highlighted:

pdf

$$pdf = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}$$

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

Parameters:

$\mu \in R$ = mean

$\sigma^2 \in R > 0$ = Varience

$x \in R$

Eg: Which types of data set follow Normal Distribution

1. Weight of the people
2. Weight of the student in the class
3. Height of student in the class
4. IRIS data set
5. Q-Q plot {Quantile-Quantile Plot}

Probability in Uniform Distribution

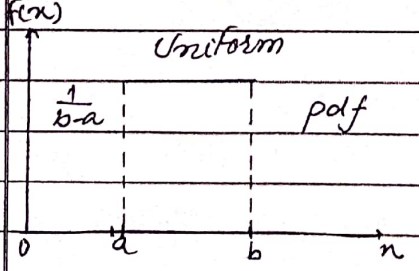
- Continuous
- Discrete

In probability theory and statistics, the continuous uniform distribution or rectangular distribution is a family of symmetric probability distribution.

The distribution describes an experiment where there is an arbitrary outcomes that lies between certain bounds.

↳ Where is minimum and maximum limit

The bounds are defined by the parameters, a and b, which are the minimum and maximum values



Notation: $U(a, b)$

Uniform a, b

Parameters: $-\infty < a < b < \infty$

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

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

Note ( ):
Note ( ):

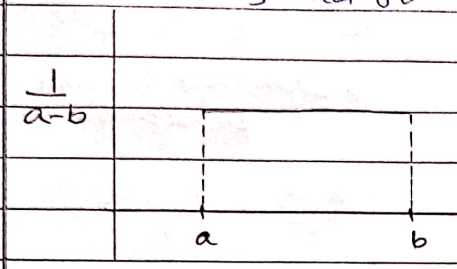
Important * Example e.g Visit ⊕ Points ⇒ Title: Sub Title: Highlighted:

Page No. Date ../../..

Mean	Median	Variance
$\frac{1}{2}(a+b)$	$\frac{1}{2}(a+b)$	$\frac{1}{12}(b-a)^2$

Eg1: The number of candles sold daily at a shop is uniformly Distribution with a maximum of 40 and minum of 10.

1) Probability of daily sales to fall between 15 and 30



(Where $x_1 = 15$ and $x_2 = 30$

$P_o(15 \leq x \leq 30) = (x_2 - x_1)\frac{1}{b-a}$

$= (30-15)\frac{1}{40-10} = \frac{1}{2} = 0.5$

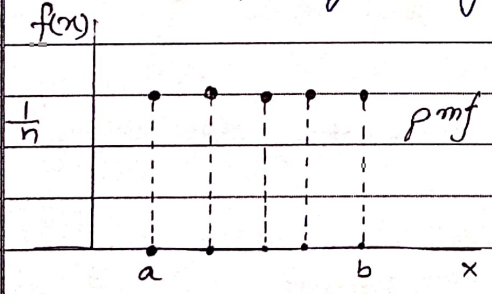
Some 2 $P_o(x \geq 20) = (40-20) * \frac{1}{30}$

$= 0.66 = 60\%$

Discrete Uniform Distribution (pmf)

In probability theory and statistics, the discrete uniform distribution is a symmetric probability distribution wherein a finite number of a value are equally likely to be observed; every one of n values has equale probability 1/n.

An other way of saying "discrete uniform distribution" would be a known. finite number of outcomes equally likely to happen

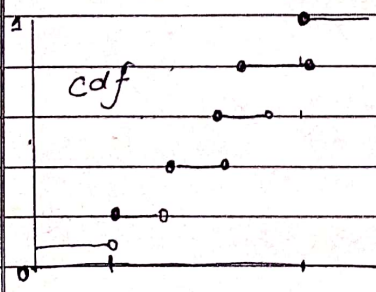


Eg: Rolling a dice

$\{1,2,3,4,5,6\}$ $P_o(1) = 1/6$

$a = 1$ and $b = 6$ $P_o(2) = 1/6$

$\frac{1}{n} \Rightarrow n = b-a+1$ $P(3) = 1/6$



Notation: $U(a,b) \rightarrow$ uniform

Parameters: a, b with $b \geq a$

Pmf $1/n$

Mean, Median $\} \frac{a+b}{2}$

Note ( ):

Note ( ):

Important *
Example e.g
Visit ⊕
Points ⇒

Title
Sub Title
Highlighted

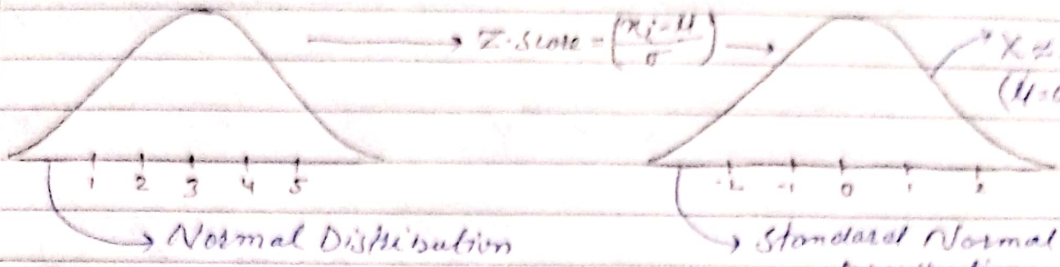
Page No
Date / /

Z-Stat and Z-Test

Standard Normal Distribution And Z-Score

→ where mean is zero

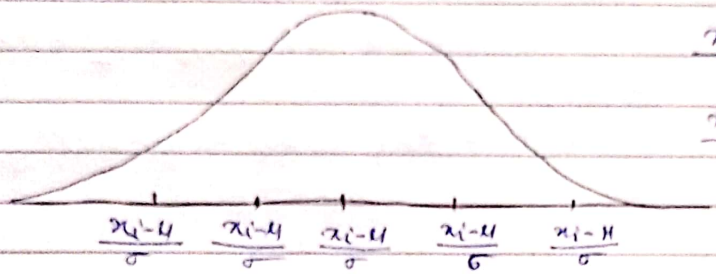
$X = \{1,2,3,4,5\}$ where mean = 3 and $\sigma = 1.414 \approx 1$



→ Normal Distribution

→ Standard Normal Distribution

Solution



$\frac{x_i - 4}{\sigma} = \frac{1-3}{1} = -2$

$\frac{x_i - 4}{\sigma} = \frac{2-3}{1} = -1$

....

$\frac{x_1-4}{\sigma}$ $\frac{x_2-4}{\sigma}$ $\frac{x_3-4}{\sigma}$ $\frac{x_4-4}{\sigma}$ $\frac{x_5-4}{\sigma}$

Eg1: In $X = \{1,2,3,4,5\}$, 4 how much away from standard divition

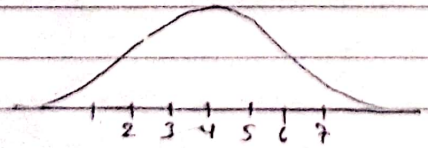
Some $Z\text{-score} = \frac{x_i - 4}{\sigma}$, where $x_i = 4$ and $4 = 3$

$= \frac{4-3}{\sigma} = 1$ (+ve to the right)

So 4 is 1 Standard divition away from mean

Z-score define how much away from the mean a single random variable

Eg2: How many standard divition 4.5 is away from mean?
where $4 = 4$ and $\sigma = 1$



$Z\text{-score} = \frac{x_i - 4}{\sigma}$ $x_i = 4.5$, $4 = 4$, $\sigma = 1$

$= \frac{4.5 - 4}{1} = 0.5$

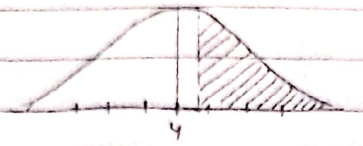
Note ( ):
Note ( ):

Important *
Example e.g.
Visit ⊕
Points ⇒

Title:
Sub Title:
Highlighted.

Page No.
Date / /

Eg: How much % of data is greater than 4.5



given $x_i = 4.5$, $\mu = 4$, $\sigma = 1$

Solve $z\text{-score} = \frac{4.5-4}{1} = 0.5$ (+ve)

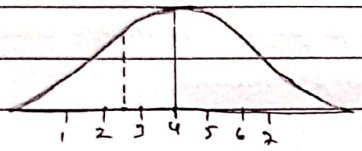
Area under the curve (≥ 4.5) = $1 - 0.69146$

(1 → whole area under the curve; 0.69146 → +Z table value)

$= 0.30854$

In Percentage = 30.85%

Eg: What % of data is falling below 2.5?



given $x_i = 2.5$, $\mu = 4$, $\sigma = 1$

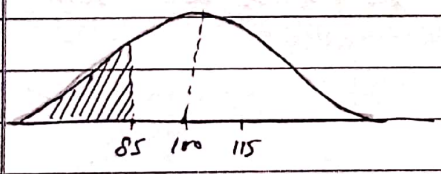
$\frac{x_i - \mu}{\sigma} = \frac{2.5-4}{1} = -1.5$

Area under the curve (≤ 2.5) = 0.06681

= 6.6%

Eg: In India the average IQ is 100, with a standard division of 15. With a percentage of the population would you except to have an IQ lower than 85?

Solve



given

$x_i = 85$, $\sigma = 15$, $\mu = 100$

$\frac{x_i - \mu}{\sigma} = \frac{85-100}{15} = -1$

See negative Z-table −1

Area under the curve = 0.15866.

= 15.866

Area under the curve (≥ 85) = 1 − 0.15866

≈ 84%

Note ( ):
Note ( ):

Important * Example e.g Visit ⊕ Points ⇒

Title:
Sub Title:
Highlighted:

Page No.....
Date ...../...../.....

Central Tendency Vs Central limit Theorem

Central Tendency

The central tendency refers to a measure that describe the typical or central values of a data set. The most common measures are the mean, median, and mode.

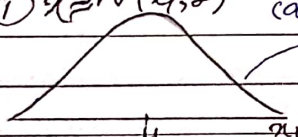
Central limit Theorem

The CLT states that if you have a large sample size from a population with a finite mean & variance, then the distribution of the sample mean will approximate a normal distribution, regardless of the distribution of the original population. This theorem is important because it is allows us to make inference about the population mean based on the sample mean.

Central Limit Theorem (CLT)

The central limit theorem says that (a) the sampling distribution of the mean will always be normally distributed, as long as the sample size is large enough.

① $X \sim N(\mu, \sigma)$



Let $n = 20$ and S = Sample

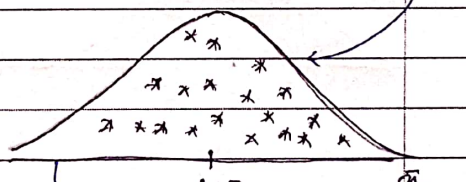
$S_1 = \{x_1, x_2, x_3 \ldots x_{20}\} = \bar{x}_1$
$S_2 = \{x_1, x_3, x_4, \ldots x_{20}\} = \bar{x}_2$
$S_3 = \{x_5, x_6, x_2, \ldots x_{20}\} = \bar{x}_3$
$S_4 = \{x_{10}, x_9, x_3, \ldots x_{20}\} = \bar{x}_m$
$S_m = \{ \ldots \ldots \}$

→ Sample data can be repeated

Gaussian Distribution

n = any size from population in GD

$\bar{x} = \{\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots \bar{x}_m\}$



→ Normal Distribution

If you have population that is normally distributed. And if you get different sample from population $S_1, S_2, S_3, \ldots S_m$ (Some data can be repeated) and number of sample is n, then find all samples mean where $S = \mu\bar{x}$ and $n = \bar{x}$

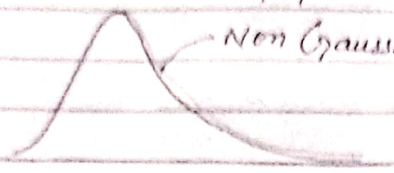
Now you get Normal Distribution.

② (b) Regardless of whether the population has a normal, Poission, binomial, or any other distribution.

Note ( ): Relate to (a) and (b)
Note ( ):

$\{ X \approx N(\mu, \sigma) \rightarrow \boxed{n \geq 30}$

μ = population mean, σ = population std, n = sample size



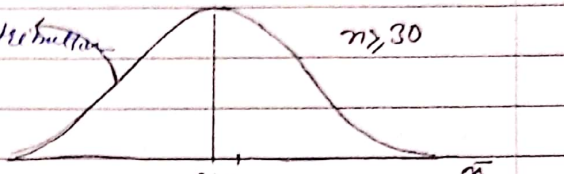
$S_1 = \{x_1, x_2, x_3, \ldots x_{30}\} = \bar{x}_1$

$S_2 = \{x_2, x_5, x_{10}, \ldots x_{30}\} = \bar{x}_2$

$S_3 = \{x_9, x_8, x_7, \ldots x_{30}\} = \bar{x}_3$

$S_m = \{x_9, x_{11}, x_{13} \ldots x_{30}\} = \bar{x}_m$

⇓



CLT is important for Interview

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

Estimate

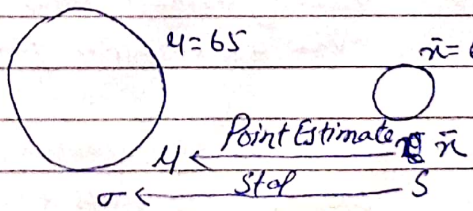
It is an observed numerical value used to estimate an unknown population parameter

① Point Estimate
② Interval Estimate

Point Estimate : Single numerical value used to estimate the unknown population parameter

* Sample mean is a point estimate of a population mean

→ Some time we are not able to find population mean due to data loss (null value, NaN) etc then, we use estimate to solve



Problem: It may be large different between population mean and sample mean or σ and S. Then we go with Interval Estimate.

Note ( ):
Note ( ):

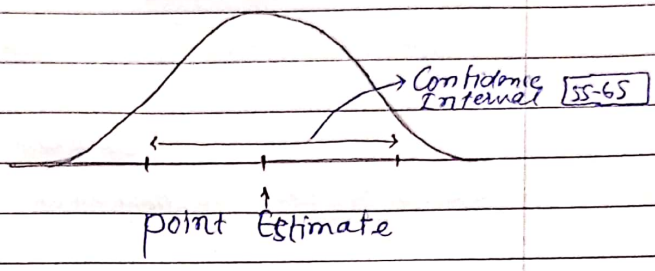
Important * Example e.g Visit ⊕ Points ⇒

Title:
Sub Title:
Highlighted:

Page No.....
Date ...../...../.....

Interval Estimate: Range of value used to estimate the unknown population parameters

*Interval estimate of population parameters are called of confidence intervals.

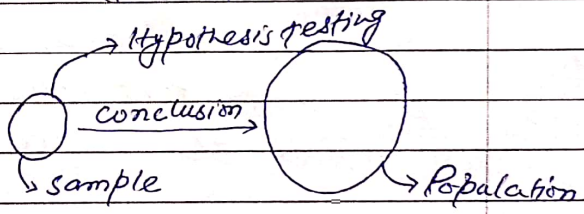


Hypothesis.

Hypothesis and hypothesis testing.

A Hypothesis is a assumption about a population parameter that is being tested using data from a sample.

It's a part of inferential stats.



→ by using hypothesis testing we get unknown population parameter

→ Hypothesis testing happing for specific population data

Hypothesis Testing Mechanism

Eg: Person Crime → Court

Step1: Null Hypothesis (H_0) = The person is not guilty
- The assumption, you are begining with.

Step2: Alternate Hypothesis (H_1) = The persion is guilty
- opposite of null hypothesis

Step3: Expromonts → Proof collect (DNA Test, Finger Point)
Stalistics Analysis using

Step4: Accept the null hypothesis or reject hypothesis

Note ( ):
Note ( ):

Eg: Colleges at District A stats it is average passed % of student are 85%, A new collage opened in the district and it was found that a sample of student 100 homa a pass % of 70% with a standard division of 4%.

Does this home a different passed percentage.

Solution:

Null Hypothesis H_0 = $\mu = 85\%$.

Alternate Hypothesis H_1 = $\mu \neq 85\%$.

P Value (Hypothesis 3rd step)

The P value is a number, calculated from a statistical test, that describes how likely you are to have found a particular set of observation if the null hypothesis were true.

Pvalue are used in hypothesis testing to help decide whether to reject the null hypothesis

Eg: in case of your keyboard space key

Out of 100 touches in this key → P=0.8
the probability of touching in this region 80%.

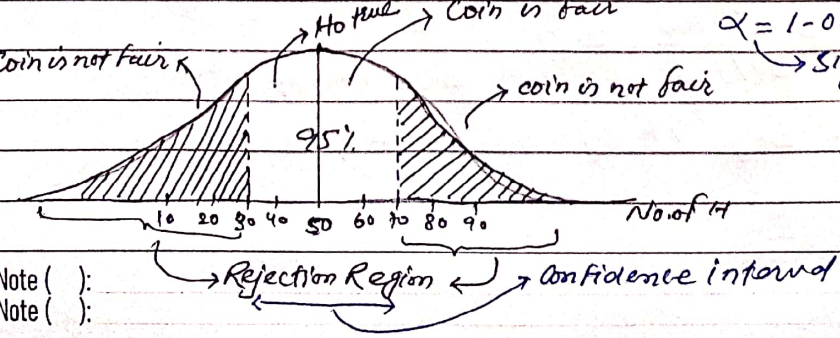
Eg: Hypothesis Testing Exp = Coin in fair or Not with 100 tossing

Confidence Interval ← $X = \{H, T\}$

① H_0 = Coin is fair $30 \leq Pr(H) \leq 70$

② H_1 = Coin is not fair C.I = 0.95

$\alpha = 1 - 0.95 = 0.5$ → Significance Value



Note ( ):
Note ( ):

Important *
Example e.g.
Visit ⊕
Points ⇒

Title:
Sub Title:
Highlighted:

Page No.....
Date ...../...../.....

Hypothesis Testing And Statistical Analysis

Z-table for Z-Test
t-table for t-Test

Average { Z-Test, t-Test }
Categorial Data ← Chi-Test
Variance ← ANOVA

Use z-test { given population std, given sample size $n \geq 30$ }

Eg: The average of all height residents in a city is 168cm with $\sigma = 3.9$. A doctor belive the mean to be different. He measured the height of 36 individual and found the average height to be 169.5cm
(a) State null and Alternate Hypothesis
(b) At a 95% confidence level,

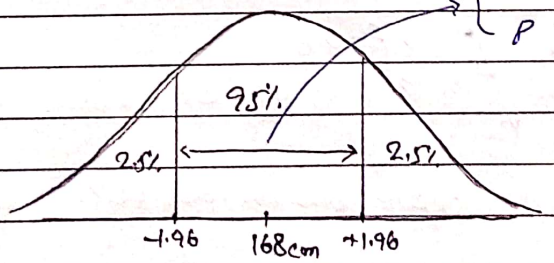
Solution:

① given $\mu = 168cm$, $\sigma = 3.9$, $n = 36$, $\bar{x} = 169.5$ cm

a) Null Hypothesis $H_0 = \mu = 168$ cm

b) Alternate Hypothesis $H_1 = \mu \neq 168$cm {2 tail Test}
if less than or grater than true H_0 rejected

c) C.I = 95 so $\alpha = 1 - 0.95 = 0.05$
Right and left $\alpha = 0.05/2$



{ C.I by using z-test, p value C.I }

$1 - 0.25 = 0.9750$

⇒ Statistical Analysis for sample data

$$z\text{-test} = \frac{\bar{x}_0 - \mu}{\sigma/\sqrt{n}}$$

$$= \frac{169.5 - 168}{3.9/\sqrt{36}} = 2.31$$

$2.31 > +1.96$ {we reject the H_0}

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

for population data

② P-value: Says if significance value > H_0 reject H_0

Note ( ):
Note ( ):

Important * Example e.g Visit ⊕ Points ⇒

Title:
Sub Title:
Highlighted:

Page No.
Date ../../..

Eg2: A factory manufactures bulbs with a average warranty of 5 years with standard division of 0.50. A worker belives that the bulb will manufacture in less than 5 years. He test a sample of 40 bulbs and find the average time to be 4.8 years.

a) State null and Alternate Hypothesis

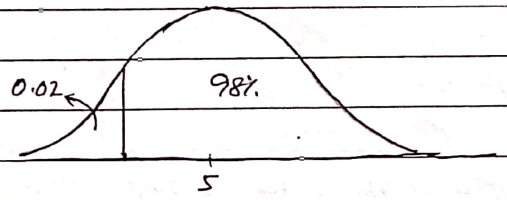
b) At a 2% significant level, is through enough evidence to support idea that the warranty should be revised?

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

① $H_0 = \mu = 5$

② $H_1 = \mu < 5$ {1 tail test}

③ Decision Boundary



Pvalue:

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

$$= \frac{4.8 - 5}{0.50/\sqrt{40}} = -2.53$$

If Pvalue < significent = False

0.0570 < 0.02 => false

{we accept H_0}

Area under the curve = -2.53

Ztable value is = 0.0570

Pvalue = 0.0570

Conclusion: The warrenty needs to be revised.

Student t distribution

In statistical Analysis using z-score we need (σ):

but "How do we performe an analysis when we dont know the population standard deviation?

Solution: By using "Student t distribution"

$$z\text{-test} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \Rightarrow t = \frac{\bar{x} - \mu}{s/\sqrt{n}}, \quad s = \text{Sample Std}$$

Degree of freedome

dof = n - 1, n = Sample size

Note ( ):

Note ( ):

Important *
Example e.g
Visit @
Points =>

Title:
Sub Title:
Highlighted:

Page No
Date / /

T-Stats t-test (one sample t test)

① 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 positive or negative effect on intelligence, or no effect at all. A simple of 30 participants who have taken the medication has a mean of 140 with with a standard deviation of 20. Did the medition effect intelligence? C.I = 95%.

Solution

Ans → given $\mu = 100$, $n = 30$, $\bar{x} = 140$, $\sigma = 20$

C.I = 95%, $\alpha = 5\%$

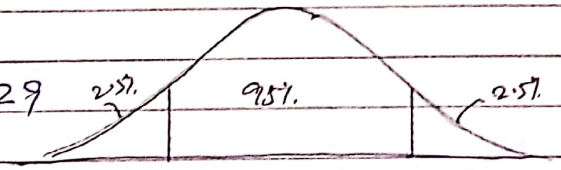
① Null Hypothesis $H_0 : \mu = 100$

Alternate " $H_1 : \mu \neq 100$ {2 tail test}

② $\alpha = 5\% = 0.05$

③ degree of freedom

dof $= n - 1 = 30 - 1 = 29$



Desision Rule

→ If it test less than −2.045 and grater than 2.045, Reject the Ho

⑤ Calculate t-test statistics

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

⑥ Decision Rules: If t is less than −2.045 and grater than 2.045, Reject the null hypothesis

$t = 10.96 > 2.0452 \rightarrow$ Reject the null hypothesis

Answer:

Conclusion: Mediantion has increased the intelligence

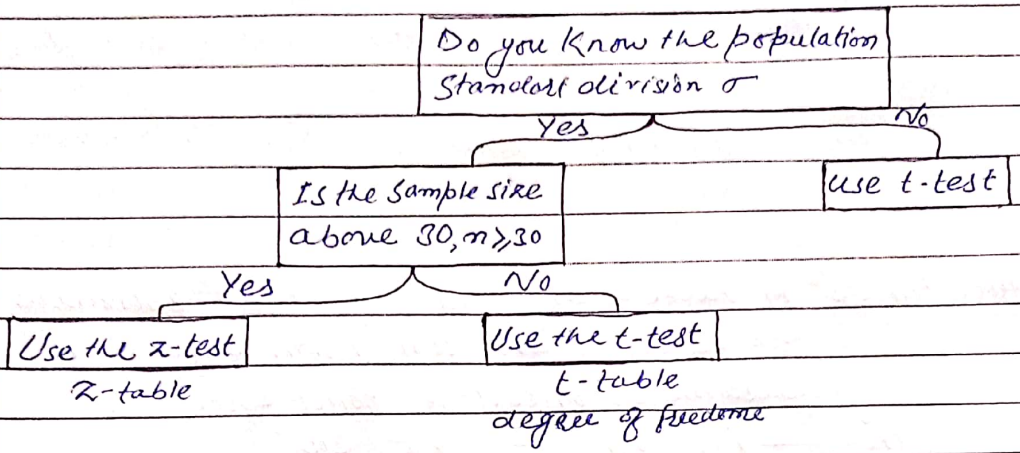
Note ( ):
Note ( ):

Important * Example e.g Visit ⊕ Points ⇒

Title:
Sub Title:
Highlighted:

Page No.....
Date ...../...../.....

When to use t-test Vs z-test



Type 1 & Type 2 errors

Decision	Reality: H_0 True	Reality: H_0 False
Fail to Reject H_0	✓	Type II (β) Error
Reject H_0	Type I Error (α)	✓

Reality: Null Hypothesis is true, or null hypothesis is fail

Decision: Null hypothesis is true, or null hypothesis is fail
↳ Conclusion

Outcomes 1: We reject the null hypothesis when in reality it is false → Good

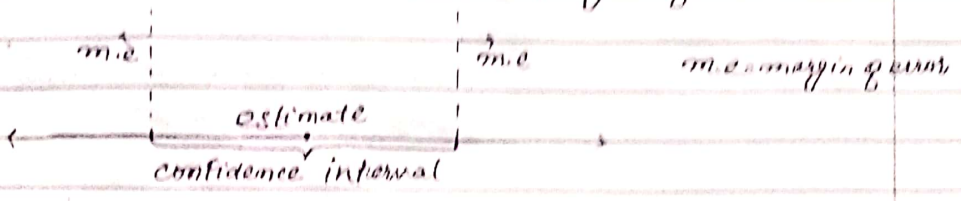
Outcomes 2: We reject the null hypothesis when in reality it is True → Type I error

Outcomes 3: We retain the null hypothesis, when in reality it is false → Type II error

Outcomes 4: We retain the null hypothesis, when in reality it is true → Good

Note ( ):
Note ( ):

Confidence Interval and Margin of Error



Point Estimate: A value of any statistics that estimate the value of an unknown population parameter is calle of Point estimate

$\bar{x}$ ——→ μ or s ——→ σ etc.

Confidence Interval: We construct a confidence interval to help estimate what the actule value of the unknown population mean is.

Point Estimate ± Margin of error

2-test C.I = $\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{2}}$

Margin of Error
$E = Z_{\alpha/2} * \frac{\sigma}{\sqrt{2}}$
(Formula)

α = is a significant value
$\bar{x}$ = Sample mean
n = Sample size
$Z_{\alpha/2}$ = Two tailtest
↳ Get value from z-table

Margin of Error
Margin of error has based on z-test and t-test table has different values

Eg: On the verbal section of cat exam, the σ is to be 100. A sample of 25 test takes has a mean of 520. constsuct 95% C.I about the mean?

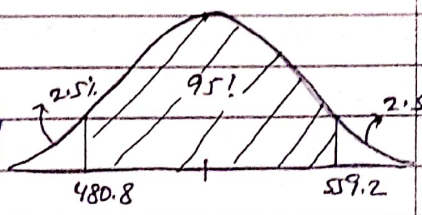
Solve,

given $\sigma = 100$, $n = 25$, $\bar{x} = 520$, C.I = 95%. $\alpha = 5\%$.

z test C.I = $\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{2}}$

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

Higher C.I = $520 + (1.96) * 100/\sqrt{25} = 559.2$



I am 95% confident that the mean CAT score lies b/w 480.8 and 559.2 confidence interval

Note ( ):
Note ( ):

Important * Example e.g Visit ⊕ Points ⇒ Title: Sub Title: Highlighted:

Page No. Date ../../..

Bayes theorem / Bayes Statistics

Bayes statistics is an aproach to data analysis and parameter estimation based on Bayes' theorem.

Bayes theorem
- Independent events
- Dependent events

Independent Events

The probability of each event remains the same value whether or not the other event occuss.

Eg 1: Rolling a dice {1,2,3,4,5,6}

$P(1) = 1/6$

$P(2) = 1/6$ ....

Eg 2: Tossing a coin

$P(H) = 0.5$

$P(T) = 0.5$

Dependent Events

The probability of the second event is influenced by the outcomes of the first event

$P(R) = \frac{2}{5}$ (1)

→ after occuring this event

$P(B) = \frac{3}{4}$

So first event effect next event

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

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

$P(A) * P(B/A) = P(B) * P(A/B)$

→ considering B event happed after A event

Bayes theorem ←

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

→ Conditional Probability.

⇓ it is also written

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

→ A event happened then B event

Parameters:

A, B = events

$P(A/B)$ = Probability of A given B is True (happend)

$P(B/A)$ = Probability of B is given A is true (happened)

$P(A)$, $P(B)$ = They are independent probability of A and B

In machine learning:

Dataset

Size of House	No. of rooms	Location	Price
x_1	x_2	x_3	y

x_1, x_2, x_3 → independent

y → dependent

Bayes theorem ←

$$P(y/x_1, x_2, x_3) = \frac{P(y) * P(x_1, x_2, x_3 | y)}{P(x_1, x_2, x_3)}$$

Note ( ):

Note ( ):

Important * Example e.g Visit @ Points ⇒

Title:
Sub Title:
Highlighted:

Page No.....
Date ...../...../.....

CHI SQUARE TEST

The chi square test for goodness of fit test Claims about population proportions. (categorical variance).

It is a non parameteric test that is performed on categorical (ordinal, nominal) data

Eg: There is a population of mail who likes different colors of bike

	Theory	Sample
Yellow bike	1/3	22
Orange bike	1/3	17
Red bike	1/3	59

Theory about population → Theory categorical distribution

→ Observed categorical distribution

Observer word used for sample data set

Goodness of fit test

(i) In a student class of 100 students, 30 are Right handed. Does this class fit the theory 12% of people are right handed.

	observation	Expected
Right handed	30	12
left handed	70	88
	100	100

Sample ← → Theory categorical Distribution

CHI SQUARE For Goodness of fit

In 2010 census of the city, the weight of the individuals is a small city were found to be the following

Theory about population	<50kg	50-75	>75
	20%	30%	50%

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

Observed / sample data	<50	50-75	>75
	140	160	200

Note ( ):
Note ( ):

Important *
Example e.g
Visit ⊕
Points =>

Title
Sub Title
Highlighted

Page No.....
Date ...../...../.....

Using $\alpha = 0.05$, would you conclude the population difference of weight has changed in the last 10 years?

Solution:

2010	<50kg	50-75	>75
Expected	20%	30%	50%

2020	<50	50-75	>75
n=500 Observed	140	160	200

	<50	50-75	>75
Expected	0.2×500 = 100	0.3×500 = 150	0.5×500 = 250

① Null Hypothesis H_0: The data meets the expections
Alternate Hypothesis H_1: The data does not meet the expection

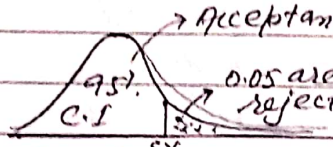
② $\alpha = 0.05$ and $C.I = 0.95$

③ degree of freedome

$df = K - 1 = 3 - 1 = 2$

* K = numbers of categories

④ Decision Boundary



CV = Critical value
χ^2 = Chi square notation

If χ^2 is grater than 5.99, Reject H_0
else:
We fail to reject the Null hypothesis

⑤ calculate chi square test statistics

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

$$= \frac{(140-100)^2}{100} + \frac{(160-150)^2}{150} + \frac{(200-250)^2}{250}$$

$$= \frac{1600}{100} + \frac{100}{150} + \frac{2500}{250} = 16 + 0.66 + 10$$

$$= 26.66$$

$\chi^2 = 26.66$ (Chi square)

Now $\chi^2 = 26.66 > 5.99$, Reject H_0

Answer: The weight of 2020 population are difference than those expected in the 2010 population

Check chi square table to find value by using degree of freedome ($df = 2$) and $\alpha = 0.05$ value

Note ( ):
Note ( ):

F-Test (Variance Ratio Test)

① Following data shows the number of bulbs produced daily for same days by 2 workers A and B

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

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

$\alpha = 0.05$

Solution

① Null Hypothesis : $H_0 \Rightarrow \sigma_1^2 = \sigma_2^2$

Alternate Hypothesis : $H_1 \Rightarrow \sigma_1^2 \neq \sigma_2^2$

② calculate the variance

A			B		
x_1	$\bar{x}$	$(x_1-\bar{x})^2$	x_2	$\bar{x}_2$	$(x_2-\bar{x}_2)$
40	37	9	39	37	4
30	37	49	38	37	1
38	37	1	41	37	36
41	37	16	33	37	16
38	37	1	32	37	25
35	37	4	39	37	4
			40	37	9
			34	37	9

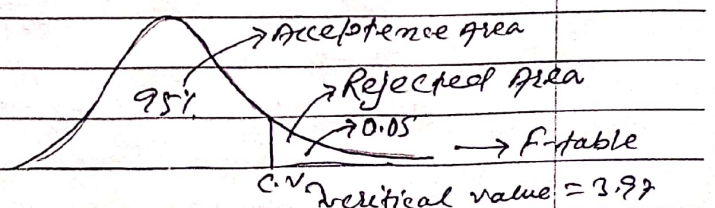
$$S_1^2 = \sum_{i=1}^{n} \frac{x_1 - \bar{x}}{n-1}$$

$$S_1^2 = \frac{80}{5} = 1$$

$$S_2^2 = \frac{\sum_{i=2} (x_2 - \bar{x}_2)^2}{n-1} = \frac{84}{7} = 12$$

→ Calculating of variance Ratio F-test

$$F = \frac{S_1^2}{S_2^2} =$$



③ Decision Rule

$df1 = 6-1 = 5$

$df2 = 8-1 = 7$

$\alpha = 0.05$

Note ( ):
Note ( ):

Important * Example e.g Visit @ Points =>

Title:
Sub Title:
Highlighted:

Page No.....
Date ...../...../.....

If F-test is grater than 3.77, Rejected the null Hypothesis H_0

1.33 < 3.77 we fail to reject the null Hypothesis.

Conclusion:

Worker B is not efficent when worked to worker A.

ANOVA (Analysis of Variance)

Defination: ANOVA is a statistical method used to compare the means of 2 or more group.

ANOVA two important thing
- (i) Factors (variable)
- (ii) levels

Eg1: Factor = Medicines
Level = 5mg 15mg 20mg (Dosage)

Eg2: Mode of Payment

Level → GPay PhonePE IMPS NEFT

ASSUMPTIONS IN ANOVA

(i) Normality of Sampling Distribution of means.
The distribution of sample mean is normaly distributed.

(ii) Absence / Remove of outliers
Outliering score need to be removed from dataset.

(iii) Homogenity
Each and every one of the population that (Sample/column) used in level has same varience
$(\sigma_1^2 = \sigma_2^2 = \sigma_3^2)$

Note ( ):
Note ( ):

Important * Example e.g Visit ⊕ Points ⇒

Title:
Sub Title:
Highlighted:

Page No.....
Date ...../...../.....

population variance in different levels of each independent variable are equale

(iv) Samples must be independent and random

Types of ANOVA

(i) One Way ANOVA: One factor with atleast 2 levels, These levels are independent

Eg: Doctors want to test a new medication to decrease headache.
They split the participant in 3 conditions. (5mg, 10mg, 15mg).
Doctors ask the participent to rate the headache (1-10).

Medication → Factor (single)

levels → 5mg	10mg	15mg	
5	7	2	all levels are independent
9	8	7	
-	-	-	
-	-	-	

(2) Repeated Measures ANOVA: One factor with at least 2 levels, levels are dependent

Running → factor

levels →	Day 1	Day 2	Day 3
Day 2 dependon day 1 etc	8	5	6
	7	4	3
	9	8	7

(3) Factorial ANOVA: Two or more factors, each of which atleast 2 levels, levels can be either independent and dependent.

Running → factor

Independent level	Day 1	Day 2	Day 3 → levels dependent
Gender: Man	8	5	6
	7	4	3
Woman	6	5	4

Note ( ):
Note ( ):

Hypothesis Testing for ANOVA

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

Alternate Hypothesis H_1 : Atleast one of the mean is not equale

Test Statistics

$$F = \frac{\text{Variance between Sample}}{\text{Variance within sample}}$$

Variance between Sample

Variance within Sample

	X_1	X_2	X_3
	1	6	5
	2	7	6
	4	3	3
	5	2	2
	3	1	4
	$\Sigma x_1 = 15$	$\Sigma x_2 = 19$	$\Sigma x_3 = 20$
	$\bar{x}_1 = 3$	$\bar{x}_2 = 3.8$	$\bar{x}_3 = 4$

$H_0 : \bar{x}_1 = \bar{x}_2 = \bar{x}_3$

H_1 : Atleast one sample meant is not equale

One way ANOVA

1) Doctors want to test a new medication with reduce headache. They split the participant into 3 condition (15mg, 30mg, 45mg). Later on the doctor ask the patient to refer the headache between (1-10). Are there any difference between 3 conditions using $\alpha = 0.05$?

Solution:

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

(1) Define H_0 and H_1

$H_0 : \mu_{15} = \mu_{30} = \mu_{45}$ and H_1 : not all μ_j are equale

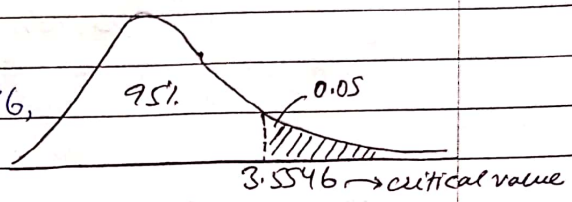
Important * Example e.g. Visit ⊕ Points ⇒
Title:
Sub Title:
Highlighted:
Page No.
Date ../../..

② State significant value
$\alpha = 0.05$ and $C.I = 0.95$

③ Calculate Degree of freedom
$N = 21 \quad a = 3 \quad n = 7$

$df_{between} = a - 1 = 3 - 1 = 2$
$df_{within} = N - a = 21 - 3 = 18$ (2, 18)
$df_{total} = N - 1 = 21 - 1 = 20$ F table

④ State Decision Rule
If F is grater than 3.5546, reject the H_0.



⑤ Calculate Test statistics

$SS_{between} \quad SS_{within} \quad SS_{total}$

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

15mg = 9+8+7+8+9+8+8 = 57
30mg = 7+6+6+7+8+7+6 = 47
15mg = 4+3+2+3+4+3+2 = 21

$= \frac{57^2 + 47^2 + 21^2}{7} - \frac{(57+47+21)}{21}$

$= 98.67$

② $SS_{within} = \Sigma y^2 - \frac{\Sigma (\Sigma a_i)^2}{n}$

$= \Sigma y^2 - \left(\frac{57^2 + 47^2 + 21^2}{7}\right)$

$\Sigma y^2 = 9^2 + 8^2 + 7^2 + 8^2 + 8^2 + 9^2 + 8^2 + 7^2 + 6^2 + \ldots$

$= 853$

$= 853 - \left(\frac{57^2 + 47^2 + 21^2}{7}\right) = 10.29$

③ $SS_{total} = \Sigma y^2 - \frac{T^2}{N}$

$= 853 - \frac{125^2}{21} = 108.95$

Note ( ):
Note ( ):

Important *
Example e.g
Visit ⊕
Points ⇒

Title:
Sub Title:
Highlighted:

Page No.....
Date ...../...../.....

	SS	df	MS	F
Between	98.67	2	49.34	
Within	10.29	18	0.54	
Total	108.95	20		

$$F = \frac{MS\ between}{MS\ within}$$

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

If F is grater than 3.5546, reject the H_0

86.56 > 3.5546, Reject the null hypothesis

Note ( ):
Note ( ):