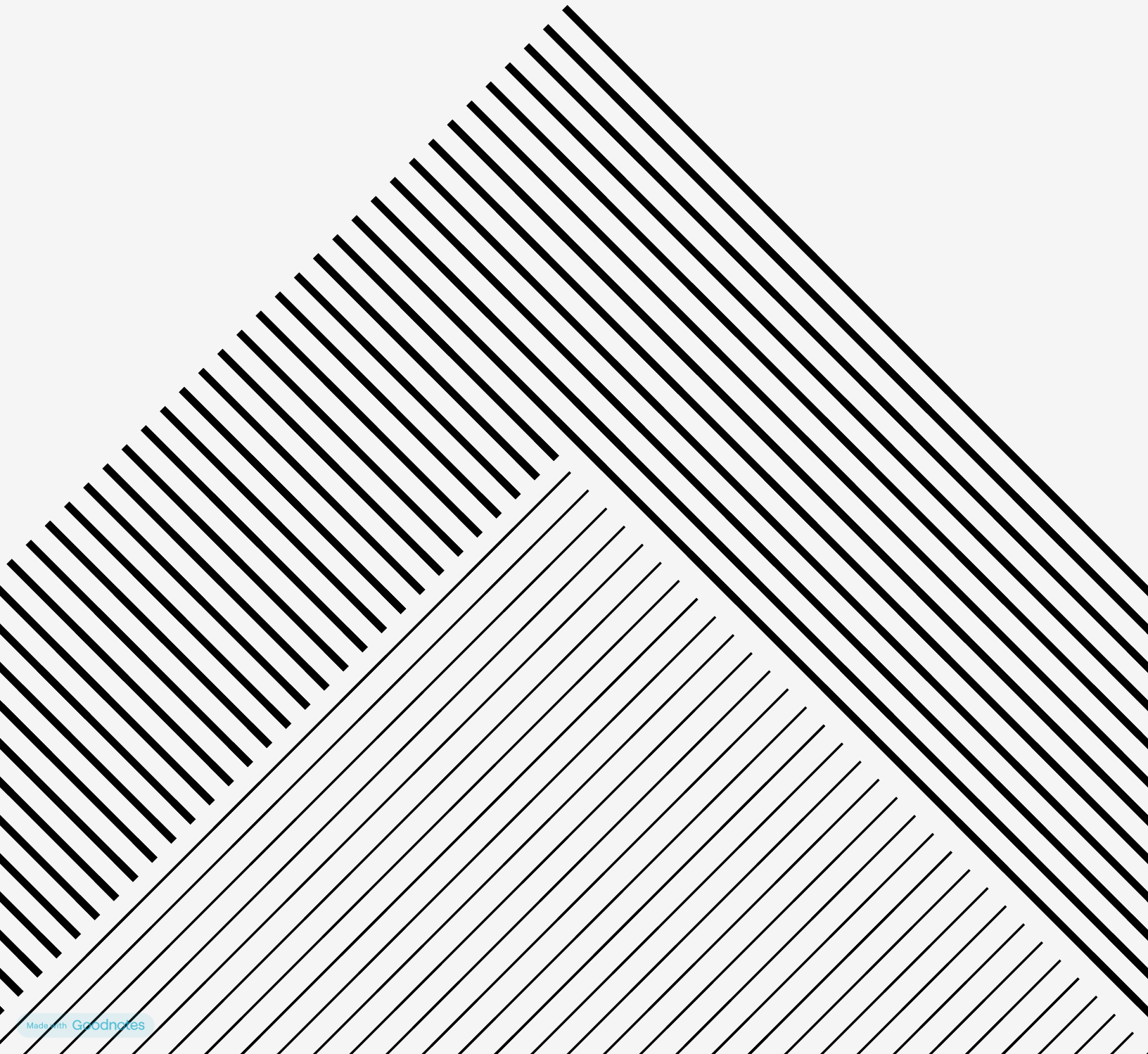
Statistics



STATISTICS

Definition:

Statistics is a science of Collecting, organizing and analyzing data

Data = ' facts or pieces of information

Eg:
- Heights of students in classroom
- IQ of students
- Weight of people

Entire statistics is devided into 2 parts:

Types of statistics

Descriptive Stats

defn:
It consists of organizing and summarizing data

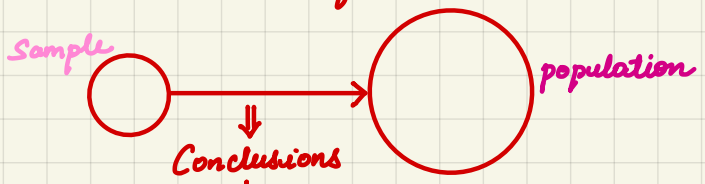
1. Measure of central Tendency {mean, median, mode}
2. Measure of Dispersion {variance, standard deviation}
3. Different types of distribution of data
   Eg: Histogram, pdf, pmf, cdf

CLT (Central Limit Theorem)

Eg: "what is the average height of students in the class"

Inferential Stats

defn:
It consists of data you have measured to form conclusion



Conclusions

Hypothesis Testing

1. Z test
2. t-test
3. Chi-Square test
4. ANOVA
5. F-Test

} Conclusions of sample on Population

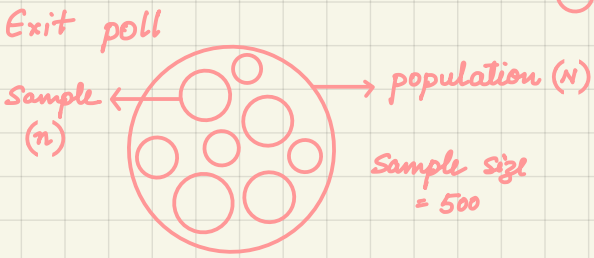
CI, p value, α

Eg: "Are the height of students in the class similar to what you expect in a college"

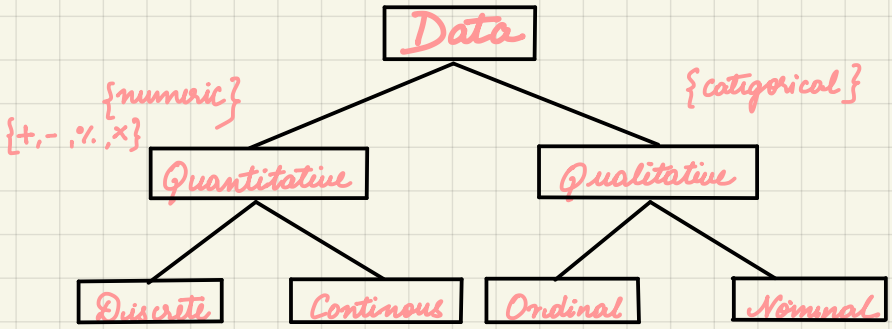
Population and Sample data

The best example for this would be exit polls

Here they take sample data from various regions and make conclusions on the population data



Types of data:



whole numbers with some range

Any value

We specifically talk about Ranks

No specific rank (one cannot be higher than another)

Eg: No of bank accounts
1, 2, 3

Eg: Weight, height
57.5, 5 feet 2 in
Temp, age

Eg:
Good (2), Better (3)
Best (1)

Eg: Gender (M, F)
Blood group
Color of hair

Based on the problem statement we can convert Nominal into ordinal

Eg: As an employee you might like weekends the best and dislike "Mondays"
So if I give a rank then

Friday (1), Sunday (2), Monday (3)

Scales of measurement of data

1. Nominal scale data
2. Ordinal scale data
3. Interval scale data
4. Ratio scale data

1. Nominal scale data

Here we mainly measure Qualitative / Categorical data

Eg: Colours, Gender, Labels.

Order does not matter

Eg: Favourite color.

Red – 5 – 50%.
Blue – 3 – 30%.
Orange – 2 – 20%.
10

2. Ordinal scale data

Categorical data

Ranking and order matters

Difference cannot be measured

Eg: Best → 1
Good → 2
Bad → 3

Race
1st – 3 min
2nd – 4 min
3rd – 5 min

We cannot measure based on only this data
Incase of race example only based on the rank i.e 1st, 2nd, 3rd we cannot measure it
Only when extra information is provided (time taken) we can measure it

3. Interval scale data

Order matters
Difference can be measured
Ratio cannot be measured
No zero starting point

Eg: Temperature

30 F
60 F
90 F
120 F

60:30 = 2:1

Room 1 – 30° F
Room 2 – 60° F

Here we can't say that Room 2 temperature is twice of Room 1

Farenheit can also have -ve values
Hence has no zero starting point

④ Ratio scale data

Order matters {we can sort this number}
Differences are measurable including ratio
Contains a zero starting point

Students marks in class
30, 45, 60, 90, 95, 99

Example:

Marital status [Nominal]
Favourite food based on gender [Nominal]
IQ measurement [Ratio scale data]
⇓ can convert
Ordinal

Discriptive statistics

① Measure of central tendency

① mean ② median ③ mode

① Mean :-

Population (N) Sample (n)

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

where
population size = (N)
sample size = (n)

Population mean $(\mu) = \sum_{i=1}^{n} \frac{x_i}{N}$ Sample mean $(\bar{x}) = \sum_{i=1}^{n} \frac{x_i}{n}$

$$\frac{1+2+2+3+3+4+5+5+6}{10} = \underline{\underline{3.2}}$$

② Median :

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

Steps:
1). Sort the random variable
2). Number of elements
3). if count == even
Take sum of center 2 elements and take the average
if count == odd
The middle element will be the median

[Note:
Median is used to find central tendency when Outlier is present]

Why median ?
- It is because of outliers

Ex: $X = \{1, 2, 3, 4, 5\}$
mean $\bar{x} = 3$
median = 3

$X = \{1, 2, 3, 4, 5, 100\}$
mean $(\bar{x}) = 19$
median = 3.5

[Note:
Mean is affected by Outliers]

③ **Mode**:

Maximum frequency occuring element

{2, 1, 1, 1, 4, 5, 7, 8, 9, 9, 10}

mode = 1

what if?

{1, 1, 2, 2} → In Python previously it used to throw errors

Now it will show 1 or (1, 2)

Mode is used in categorical values, whereas mean and median is used for numerical values

These mean, median and mode are specifically used in EDA and feature engineering

EDA: Exploratory data analysis

Age	Weight	Salary	Gender	Degree
24	70	40k	M	B.E
25	80	70k	F	—
27	95	45k	F	—
24	—	50k	M	PHD
32	—	60k	—	B.E
—	60	—	—	Masters
—	65	56k	—	Bsc
40	72	—	M	B.E

Steps
1. Check if outliers are present
2. if yes
   - median
3. else
   - mean
4. if non numeric (categorical)
   - mode

↑ mean/median

↑ mean/median

↑ Here M is most occuring so fill '-' with 'M'

↑ Similarly fill '-' with 'B.E'

② Measure of Dispersion [Spread of data]

① Variance (σ^2)
② Standard deviation (σ)

① **Variance**:

It is nothing but standard deviation square

Population variance

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

Sample variance

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

where
x_i = Data points
μ = Population mean
N = Population size
$\bar{x}$ = Sample mean
n = Sample size

(Note: Here we talk about the spread)

Question: Why we divide sample variance by n-1?

It is used to create an unbiased estimator of population variance

Ex: $\{1, 2, 3, 4, 5\}$

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

x_i	$\bar{x}$	$(x_i - \bar{x})^2$
1	3	4
2	3	1
3	3	0
4	3	1
5	3	4
$\bar{x} = 3$		$\sum (x_i - \bar{x})^2 = 10$

$\therefore S^2 = \frac{10}{4} = \underline{2.5}$

(Here x, y are random variables)

For example

$x = \{\ \}$ $\quad y = \{\ \}$

$S^2 = 2.5$ $\quad S^2 = 7.5$

The respective graph for these variances

When the variance is less then the spread is less

② Standard deviation:

It is nothing but root of variance

"How far the value is away from the mean"

Population standard deviation

$\sigma = \sqrt{\text{variance}} = \sqrt{\sigma^2} \Rightarrow \sigma$

Sample standard deviation

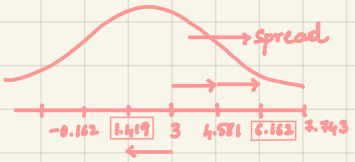
$S = \sqrt{S^2} = S$

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

$\bar{x} = 3$

$S / \sigma = 1.581$

How to show this



We can get how much far it is away from mean we use standard deviation

ex: 6.162 is 2 std deviation away from mean

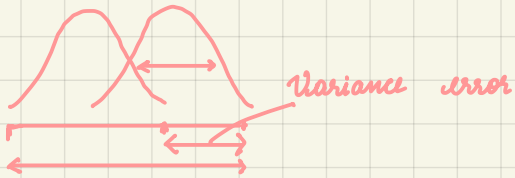
1.419 is one std deviation away from mean

Therefore variance is standard deviation square

In short we are talking about how well the data is spread

Note:

Variance error w.r.to one distribution to another



* Random variables

Linear algebra $\begin{cases} x+5=7 \\ 8=y+x \end{cases} \Rightarrow \begin{matrix} x=2 \\ y=6 \end{matrix}$ } variables

" Random variable is a process of mapping the output of a random process or experiment to a number "

Eg: Tossing a coin $\{\text{Head}, \text{Tail}\} \Rightarrow$ process

$X = \begin{cases} 0, \text{ if Head} \\ 1, \text{ if Tail} \end{cases} \Rightarrow$ "Outcome of a random process is converted to number"

"In every toss the value can change, i.e it is not fixed"

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

$y = \{$sum of rolling of dice 7 times$\}$

What can we do from this?

→ We can find probability of $(y \geq 15)$, $Pr(y < 10)$

* Histogram and Skewness.

Whenever we talk about [Frequency] the best visualization diagram we can use is Histogram

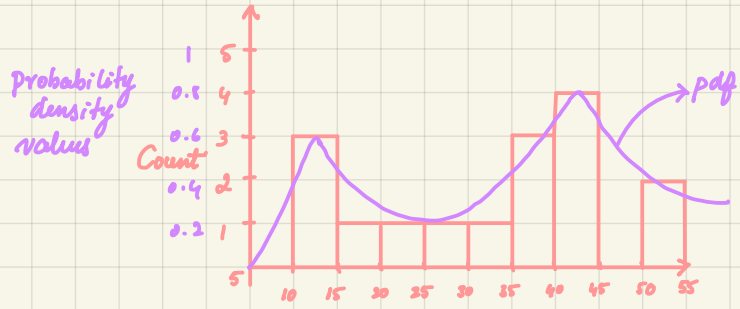
Ex:

Ages = $\{10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51\}$

If we want to map the frequency of the elements between the ranges and we want to visualize the diagrams.
Then we can specifically use Histograms

max range → 50 / Buckets → 10 = 5 → bin size

No of bins = 10 → Buckets

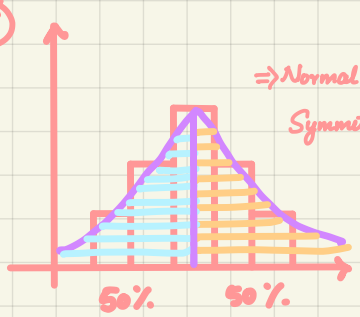


→ pdf (probability density function)

[Kernel density estimator] → is responsible for smoothening

* **Skewness** : It is a measure of the distortion of symmetrical distribution or assymmetry in a data set.

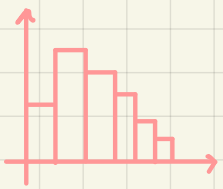
① ⇒ Normal distribution (Zero Skewed)



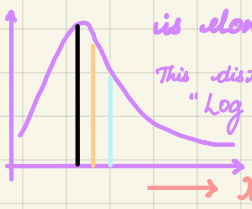
Symmetrical distribution

In this the center element is specifying
median = mean = mode
In Symmetrical distribution there is no skewness

② Right Skewed



Positive skewed (my RHS is elongated)

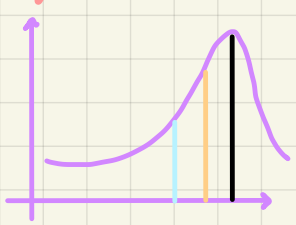


This distribution is called as "Log Normal Distribution"

mean > median > mode

③ Left Skewed:

Negative skewed



mode > median > mean

Statistics part - 2

Percentile and Quartile

Ex: GATE, CAT exams.

Percentage: 1, 2, 3, 4, 5, 6

% of numbers that are odd = $\frac{3}{6} \Rightarrow \frac{\text{No of odd numbers}}{\text{Total numbers}} \Rightarrow 50\%$

Percentiles:

A percentile is a value below which a certain percentage of data points lie

Ex:

$X = \{2, 3, 3, 4, 6, 6, 6, 7, 8, 8, 9, 9, \boxed{10}, 11, 12\}$

1. What is the percentile rank of 10?

$$\text{Percentile rank of } 10 = \frac{\text{Number of values below } 10}{n} \times 100$$

$$= \frac{12}{15} \times 100 \Rightarrow \frac{4}{5} \times 100 \Rightarrow 80 \text{ percentile}$$

This means that 80% of the distribution falls below the value of 10

2. What value exists at 25 percentile?

$$\text{value} = \frac{\text{Percentile}}{100} \times (n+1)$$

$$= \frac{25}{100} \times (15+1)$$

$$= \frac{1}{4} \times 16 \Rightarrow \underline{4^{th} \text{ element}}$$

The 4^{th} element i.e '4' is the value that is present at the 25^{th} percentile

What if we get it as 4.5^{th} element

→ Then we take average of 4^{th} and 5^{th} element

i.e $\frac{4+6}{2} = \underline{5}$

∴ The value at 25^{th} percentile is $\underline{5}$

Why are we learning this?

→ This will be used to calculate outliers

Quartiles:

① $Q_1 \longrightarrow 25^{th}$ percentile
$Q_2 \longrightarrow$ Median $\rightarrow 50^{th}$ percentile
$Q_3 \longrightarrow 75^{th}$ percentile

② 5 Number summary

(1) Minimum
(2) First Quartile (25 percentile) (Q_1)
(3) Median (Q_2)
(4) Third Quartile (75 percentile) (Q_3)
(5) Maximum (100 percentile)

Let's see how we can remove outliers using this technique

$$X = \{1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 29\}$$

Using 5 numbers summary we calculate 2 important things

[Lower fence ⟷ Higher fence]

This is a range (Border)

- Below this lower fence everything is an Outlier
- Above this Higher fence everything is an Outlier

Formula to calculate lower fence:

$$\text{Lower fence} = Q_1 - 1.5(IQR)$$

where

$IQR \longrightarrow$ Inter Quartile Range $= Q_3 - Q_1$

Formula to calculate Higher fence:

$$\text{Higher fence} = Q_3 + 1.5(IQR)$$

$$X = \{1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 29\}$$

$Q_1 = 25$ percentile $= \frac{25}{100} \times (19+1)$

$= 5^{th}$ element $\cong \boxed{3}$

$Q_3 = 75$ percentile $= \frac{75}{100} \times (19+1) \Rightarrow \frac{3}{4} \times 20^{5}$

$= 15^{th}$ element $\cong \boxed{7}$

$$IQR = Q_3 - Q_1 = 7 - 3 = 4$$

$7 + 1.5(4.5)$

$\text{Lower fence} = Q_1 - 1.5(IQR) \Rightarrow 3 - 1.5(4) = -3$

$\text{Higher fence} = Q_3 + 1.5(IQR) = 13$

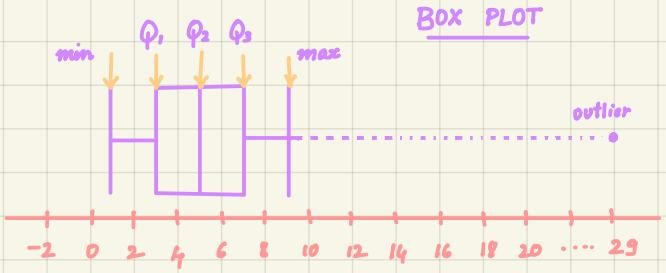
Lower fence and Higher fence are in the range between $[-3, 13]$

$\therefore$ 29 is the Outlier

Box Plot:

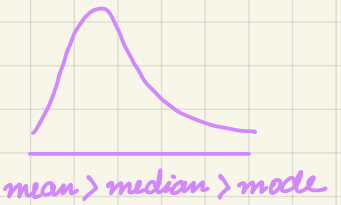
Box plot in seaborn is a diagram which will help us to visualize outliers

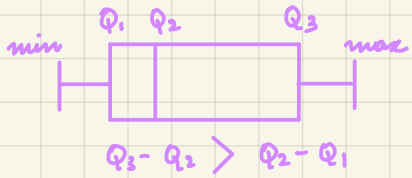
1. Minimum value = 1
2. $Q_1 = 3$
3. Median $Q_2 = 5$
4. $Q_3 = 7$
5. Maximum = 9 (donot consider outlier)



Interview Questions:

Draw a box plot for right skewed





Internal Assignment

$y = \{-13, -12, -6, -5, 3, 4, 5, 6, 7, 7, 8, 10, 10, 11, 24, 55\}$

Q_1 = 25 percentile $= \frac{25}{100} \times (17)$

$Q_1 = 4.25^{th}$ element (Hence take average of 4^{th} & 5^{th} element)

$Q_1 = -1$

Q_3 = 75 percentile $= \frac{75}{100} \times 17$

$= \frac{51}{4} = 12.7^{th}$ elem

$Q_3 = 10$

$IQR = Q_3 - Q_1$

$IQR = 11$

Lower fence $= Q_1 - 1.5(IQR)$

$= -1 - 1.5(11)$

$= -17.5$

Higher fence $= Q_3 + 1.5(IQR)$

$= 10 + 1.5(11)$

$= 26.5$

$[-17.5, 26.5]$

Hence "55" is an Outlier

$X = \{1, 2, 4, 6, 7, 12, 14, 18, 34, 66, 77, 99, 108\}$

$Q_1 = \frac{25}{100}(14) \Rightarrow 3.4^{th}$ element $\quad Q_3 = \frac{75}{100} \times 14 \Rightarrow \frac{3}{4} \times 14 \Rightarrow 10.5$ element

$Q_1 \Rightarrow \underline{5}$ $\quad Q_3 = \frac{66+77}{2} \Rightarrow \underline{71.5}$

$IQR = Q_3 - Q_1$

$= \underline{66.5}$

Lower fence $= 5 - 1.5(66.5)$

$= \underline{-94.75}$

Higher fence $= 71.5 + 1.5(66.5)$

$= \underline{171.25}$

$[-94.75, 171.25]$

Covariance and Co-relation

Ex:	X	Y
	2	3
	4	5
	6	7
	8	9

[Relationship between X and Y]

X ↑ Y ↑

X ↓ Y ↑

X ↑ Y ↓

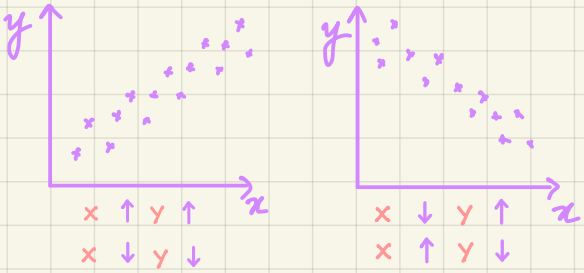
X ↓ Y ↓

These things help in feature selection

Ex: Size of House — Price of House

This follows: X ↑ — Y ↑ — directly proportional

X ↓ — Y ↓



If we want these in Mathematical formula if we want to get values from it then we use a technique called as covariance

① Covariance:

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

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

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

⇓ inversly we can write it as

$var(x) \Leftarrow Cov(x, x)$

What is the relationship between variance and co-variance?

Variance talks about spread of the data

Covariance talks about relationship between one data to another data

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

if ① | x ↑ | y ↑ | x ↓ | y ↓ | $\Rightarrow$ '+'ve Covariance

② | x ↓ | y ↑ | x ↑ | y ↓ | $\Rightarrow$ '-'ve Covariance

Ex:

x	y
2	3
4	5
6	7

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

$$Cov(x,y) = \frac{[(2-4)(3-5)+(4-4)(5-5)+(6-4)(7-5)]}{2}$$

$$= \frac{4+0+4}{2} \Rightarrow \frac{8}{2} \Rightarrow \underline{\underline{4}}$$ '+ve', The covariance is +ve and it is following ① property

x and y are having a positive covariance

Advantages

1. Relationship b/w x & y

Disadvantages

1. Covariance does not have a specific limit value (It can range anywhere from $-\infty$ to $+\infty$)

Ex: x —(+4)→ y —(+100)→ z

It also cannot tell us which is highly covariant between x, y and x, z

So in Covariance these are the disadvantages, so to overcome this we use Pearson co-relation co-efficient

② PEARSON CORRELATION CO-EFFICIENT

$$\rho_{x,y} = \frac{Cov(x,y)}{\sigma_x \cdot \sigma_y}$$

The outcome of Pearson correlation co-efficient will be ranging between [-1 to +1]

① The more the value towards +1 the more positive(+) correlated it is
② The more the value towards -1 the more negative(-) correlated it is

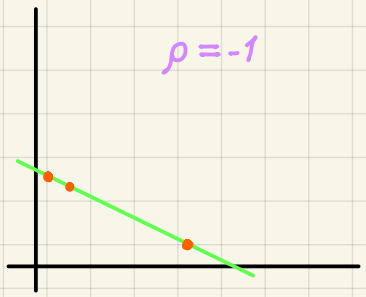
i.e x —(-1 to +1)→ y —(-1 to +1)→ z

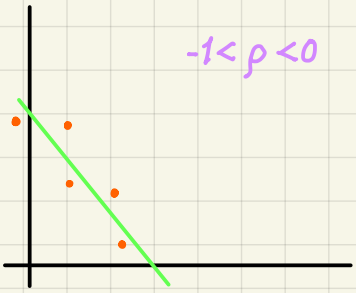
Ex:

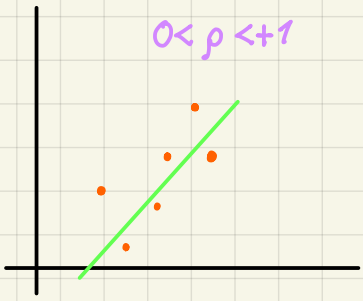
x	y	0.6
x	z	0.7

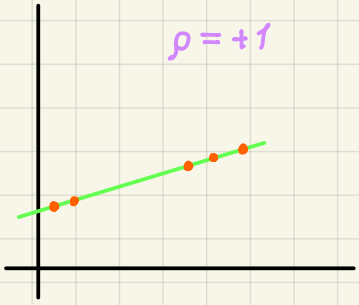
Now we can say that x and z are highly co-related than x and y
By using Pearson correlation coefficient

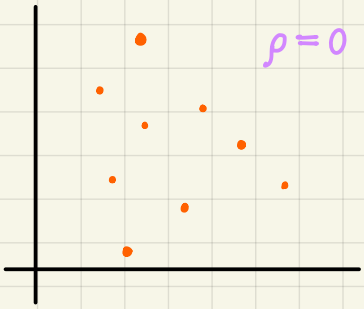
Note: Only when we restrict it we will be able to compare it











Examples of scatter diagrams with different values of correlation co-efficient

Disadvantages:

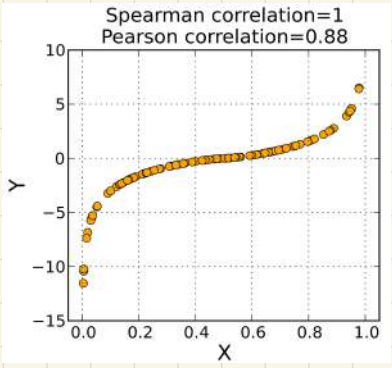
With respect to pearson correlation it can only capture the linear properties

③ SPEARMAN'S RANK CORRELATION COEFFICIENT

A Non linear relationship can be captured by Spearman's

$$r_s = \frac{Cov(R(x), R(y))}{\sigma_{R(x)} * \sigma_{R(y)}}$$

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



Lets see where we specifically use this

Feature Selection

Ex: with respect to house price

Size of House	No of rooms	location	No of ppl staying (≈0)	Haunted	Price ↑
+ve	+ve	+ve	not a very big role (hence drop the feature)	-ve	

Probability Distribution Function

① Probability density function (PDF)
② Probability mass function (PMF)
③ Cumulative distribution function (CDF)

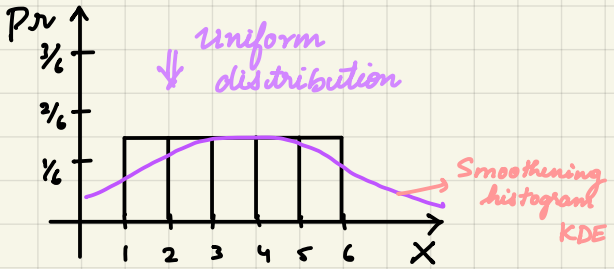
① PMF

1. Discrete random variable

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

Probability of getting 1 = 1/6
Probability of getting 2 = 1/6
Probability of getting 3 = 1/6

Probability of getting 1 or 2 = $1/6 + 1/6 \Rightarrow 2/6 \Rightarrow 1/3$

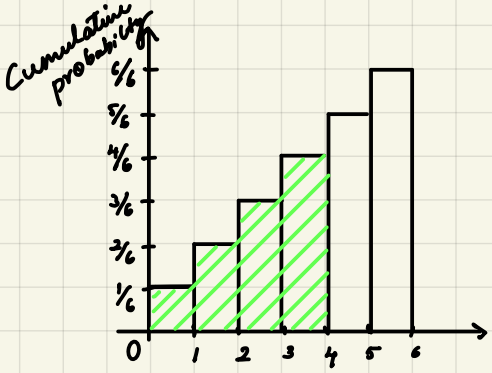


Ex: Bernoulli's Distribution

Lets see the relationship between PMF and CDF (Cumulative Distribution function)

② Cumulative Distribution function

CDF is adding up with previous entries

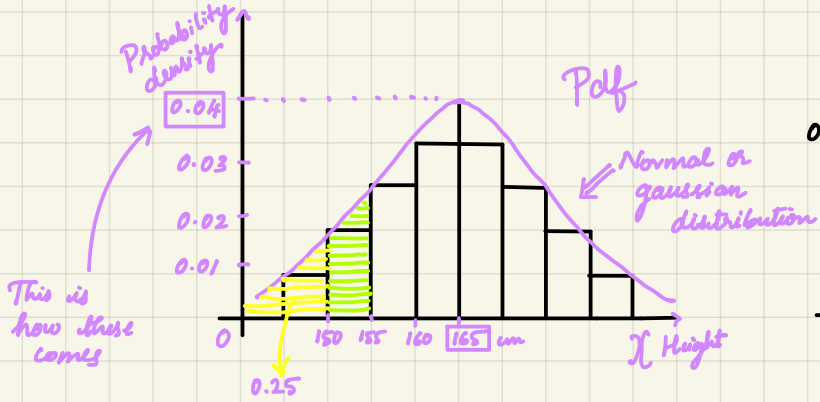


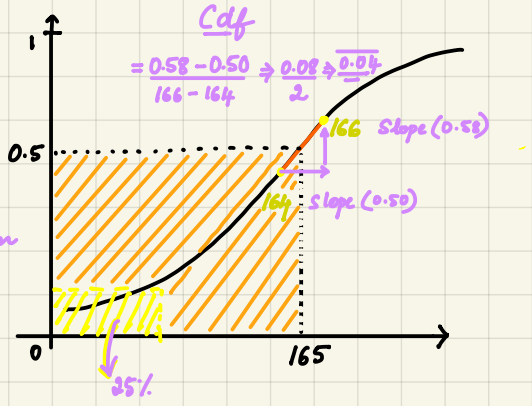
$Pr(x \leq 4) = Pr(x=1) + Pr(x=2) + Pr(x=3) + Pr(x=4)$
= Output

③ Probability density function

1. Distribution of Continous Random Variable

Eg: Age, Height, Weight, Temperature, IQ





$= \frac{0.58 - 0.50}{166 - 164} \Rightarrow \frac{0.08}{2} \Rightarrow 0.04$

Therefore

Probability Density $\Rightarrow$ Gradient of cumulative Curve

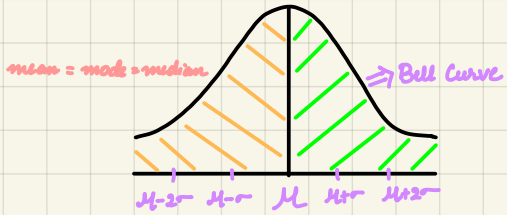
Different types of distribution:

1. Normal / Gaussian distribution → pdf
2. Standard Normal distribution → pdf
3. Log Normal distribution → pdf
4. Power Law distribution → pdf
5. Bernoulli distribution → pmf
6. Binomial distribution → pmf
7. Poisson distribution → pmf
8. Uniform distribution → pdf, pmf
9. Exponential distribution → pdf
10. CHI SQUARE distribution → pdf
11. F distribution → pdf

Why study these Distribution?

Note: The datasets we will be working on, will be following one or the other distribution

① Normal / Gaussian distribution

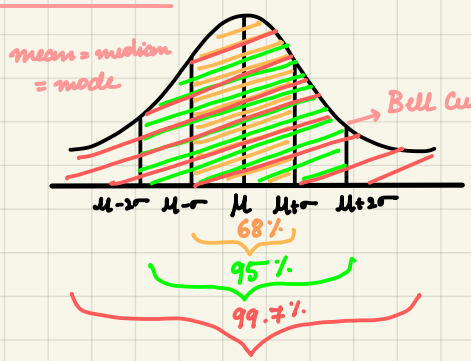


X = Continous Random Variable

Eg: Height, Weight, Age, IRIS distribution

Note: Most of the datasets that are available in the universe follow gaussian distribution

Empirical Rule — 68 – 95 – 99.7% Rule



Whenever you find a distribution that follows a gaussian distribution this rule is assumed 100%

Normal distribution

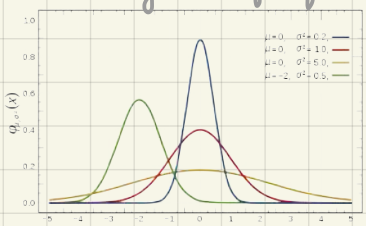
$X \approx N(\mu, \sigma^2)$

Support parameters

μ = mean

σ^2 = variance

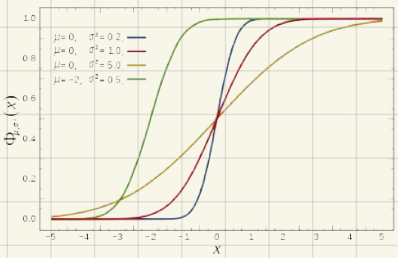
Probability density function



The red curve is standard Normal Distribution

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

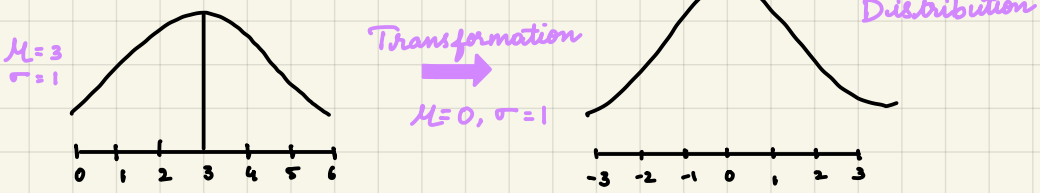
Cumulative distribution function



$$CDF = \frac{1}{2}\left[1 + erf\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$$

② Standard Normal distribution

Lets consider a random variable X
Let $\mu = 3$ and $\sigma = 1$



↙ Standard Normal Distribution

↳ This can be converted into standard normal distribution by **Z-score**

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

$= \frac{1-3}{1} = -2$

Similarly 0, 1, 2, 3

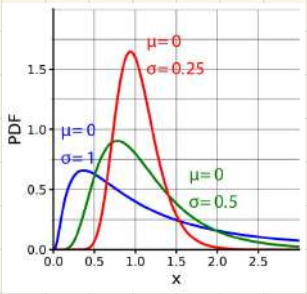
$= \frac{1-2}{1} = -1$

(Z score tells us about a value, how many standard deviations it is away from the mean)

③ Log Normal Distribution {continous random variable}

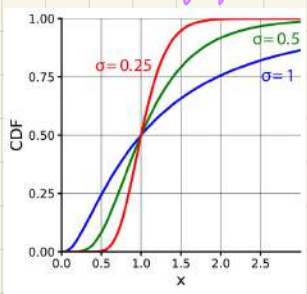
Log-Normal

Probability Density Function



In probability theory, a log normal distribution is a continous probability distribution of a random variable whose logarithm is normally distributed.
Thus if the random variable X is log normally distributed, then $y = \ln(x)$ has a normal distribution.

Cumulative density function

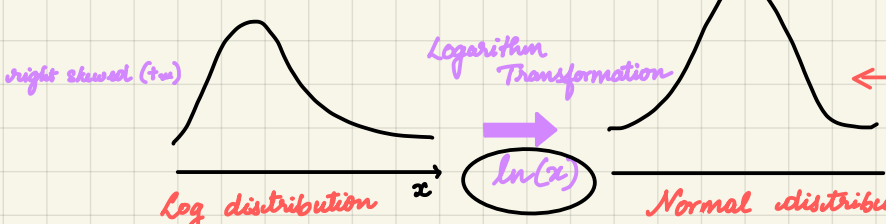


Equivalently if y has a normal distribution, then the exponential function of Y, $x = \exp(y)$, has a log normal distribution

$X \approx$ Log normal (μ, σ^2)

$y = \ln(x) \Rightarrow$ Normal Distribution (μ, σ^2)

ln = Natural log (log$_e$)

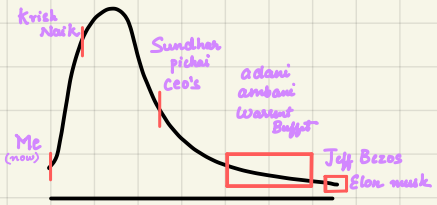


Why do we need such Transformations ? Because whenever our data follows THIS distribution, my model will get trained efficiently

(NOTE: For vice versa i.e Normal → Log Normal)
$X = \exp(y)$

(NOTE: In linear regression it says that your independent feature should follow gaussian / normal distribution.)

Ex: Wealth distribution



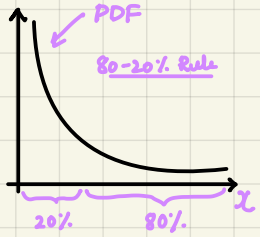
DATA
① Wealth distribution of the world
② Salary distribution in a company
③ Length of comments in youtube

Notation
Lognormal (μ, σ^2)

Parameters
$\mu \in (-\infty, +\infty)$
$\sigma > 0$

PDF : $\frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right)$

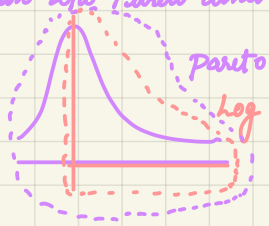
④ Power Law Distribution {Continous random variable}



In statistics, a power law is a functional relationship between two quantities, where a relative change in one quantity results in a proportional relative change in the other quantity, independent of the initial size of those quantities: one quantity varies as a power of another.
For instance,
consider the area of a square in terms of the length of its side, if the length is doubled, the area is multiplied by a factor of four

Eg: IPL games
RCB
① 20% of team is responsible for winning 80% of the match
② 80% of sales in amazon is derived from 20% of the products
③ 80% of wealth is distributed among 20% of the people
④ 80% of project completion by 20% of the team.

Q. Can you convert log normal → Pareto distribution
Q. Relation b/w Pareto and log normal distribution?

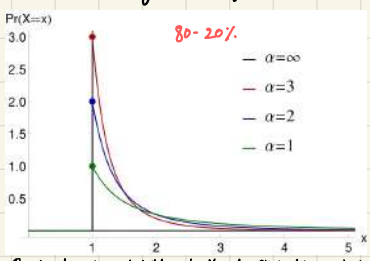


Types of Power Law Distribution
① Pareto Distribution
② Exponential Distribution

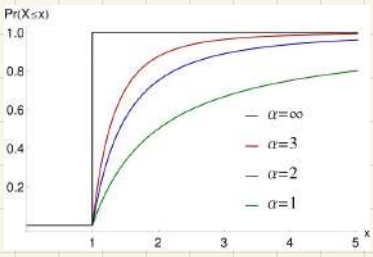
① Pareto distribution {Continous Random variable}
It follows power law distribution

Pareto Type 1

Probability Density function



Cumulative Distribution function



Pareto type 1 probability density functions for various α with $x_m = 1$. As $\alpha \to \infty$, the distribution approaches $\delta(x - x_m)$, where δ is the Dirac delta function

Parameter: $x > 0$ (real)
$\alpha > 0$ (real)

PDF = $\frac{\alpha x_m^\alpha}{x^{\alpha+1}}$

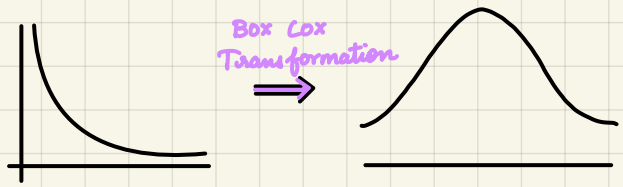
CDF = $1 - \left(\frac{x_m}{x}\right)^\alpha$

Box Cox transformation : $\begin{cases} y = \frac{x^\alpha - 1}{\alpha} & \text{where } \alpha \neq 0 \\ y = \ln(x) & \text{where } \alpha = 0 \end{cases}$

If X is a random variable with a Pareto (Type I) distribution, then the probability that X is greater than some number x, i.e. the survival function (also called tail function), is given by

$$\bar{F}(x) = Pr(X > x) = \begin{cases} \left(\frac{x_m}{x}\right)^{\alpha} & x \geq x_m, \\ 1 & x < x_m, \end{cases}$$

Q. Can we convert pareto distribution to Normal distribution?



② Exponential distribution {continous random variable}
80-20% rule

Height is varied by λ (lambda)

$$PDF = \lambda e^{-\lambda x}$$
$$CDF = 1 - e^{-\lambda x}$$
$$PDF \Rightarrow f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$
$$CDF \Rightarrow F(x;\lambda) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

⑤ Bernoulli distribution {discrete random variable}

Note: Tossing a coin discrete random variable

Outcome of the process is binary {1, 0}, {Succes, failure}

Ex: Tossing a coin

$Pr(H) = 0.5 \Rightarrow p$ $\quad \therefore p + q = 1$

$Pr(T) = 0.5 \Rightarrow 1 - p \Rightarrow q$

$$PMF = \begin{cases} q = 1-p & \text{if } k = 0 \\ p & \text{if } k = 1 \end{cases} \qquad CDF = \begin{cases} 0 & \text{if } k < 0 \\ 1-p & \text{if } 0 \leq k < 1 \\ 1 & \text{if } k \geq 1 \end{cases}$$

It can also be derived by

$pmf = p^k (1-p)^{1-k}$ where $k = \{1, 0\}$

⑥ Binomial distribution

Binomial distribution is n-times of Bernoulli distribution
(Combination of multiple Bernoulli distribution)

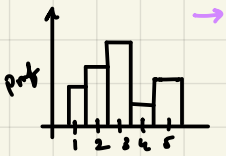
⑦ Poisson distribution

① discrete random variable (pmf)

② describes the number of events occuring in a fixed interval of time

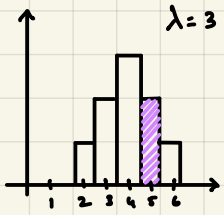
Here we take a parameter λ (lambda)

Ex: Number of people visiting bank every hour



→ $\lambda = 3 \Rightarrow$ Expected number of people to come at that specific time interval

pmf $Pr(x = 5) = \frac{e^{\lambda} \lambda^{x}}{x!}$

Ex: 

$Pr(x=5) = \frac{e^{-\lambda}\lambda^x}{x!} \Rightarrow \frac{e^{-3}3^5}{5!}$

$= 0.101 \Rightarrow \underline{10\%}$

Note: If $Pr(x=5 \text{ or } x=6) = Pr(x=5) + Pr(x=6)$

$= \frac{e^{-\lambda}\lambda^5}{5!} + \frac{e^{-\lambda}\lambda^6}{6!}$

* Uniform Distribution

① Continuous Uniform Distribution (pdf)
② Discrete Uniform Distribution (pmf)

① Continous Uniform Distribution {Continous random variable}

Eg: No of candies sold daily at a shop is uniformly distributed

Here we will be able to find a range i.e [15-20]

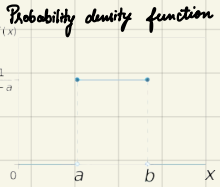
So whenever we have these kind of datasets, which will have some specific range, that will be called as Continous Uniform Distribution

Here we will have a min and max value [min, max] ⇒ Interval

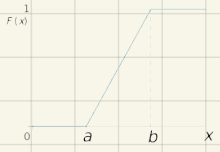
Notation : $U(a,b)$

Parameters : $-\infty < a < b < \infty$ (b >> a)

Probability density function



Cumulative distribution function

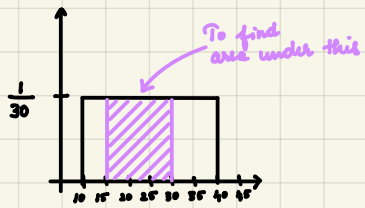


$$Pdf = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$

Eg: The number of candies sold daily at a shop is uniformly distributed with a maximum of 40 and minimum of 10

i) Probability of daily sales to fall below 15 and 30.

→



$x_1 = 15$
$x_2 = 30$

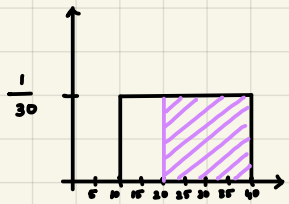
$Pr(15 \le x \le 30)$

= We should find area under square / rectangle (l × b)

$= (30-15) \times \frac{1}{b-a}$

$= \frac{15}{1} \times \frac{1}{30_2} \Rightarrow \underline{1/2} \Rightarrow \underline{0.5} \Rightarrow 50\%$

ii) Probability of sales above 20



$x_1 = 20$
$x_2 = 40$

$Pr(20 \le x \le 40)$

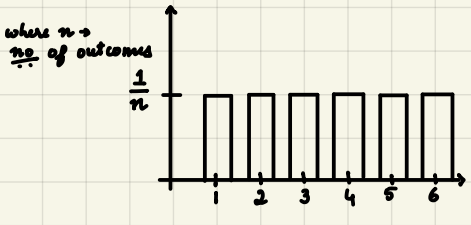
$= (40-20) \times \frac{1}{b-a}$

$= 20 \times \frac{1}{30} = \underline{0.66} \Rightarrow \underline{66\%}$

② Discrete Uniform Distribution

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

$Pr(1) = 1/6$, $Pr(2) = 1/6$



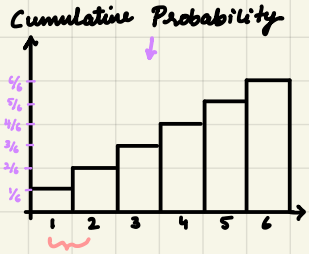
$n = b - a + 1$

$n = 6 - 1 + 1 \Rightarrow \underline{6}$

Notation : $U(a, b)$

Parameters : a, b with $b > a$

$Pmf = 1/n$

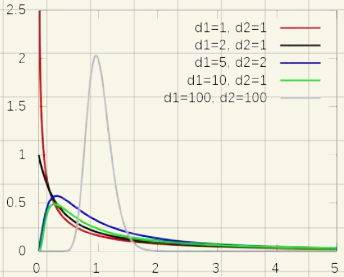


⇒ If we ask what is the $Pr(X = 1 \text{ or } 2)$

* F- Distribution :

It is used to compare the variance between 2 groups

PDF



Here d_1 and d_2 are degrees of freedom for 2 distribution

d_1 – 1st and d_2 – 2nd distribution

Here there is unique relationship

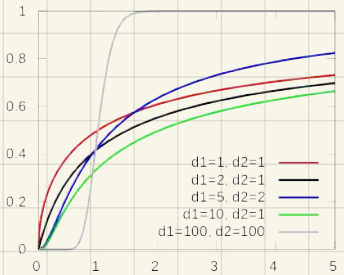
As degree of freedom is increasing it is becoming normal distribution

initially if degree of freedom is less we have exponential distribution

then as it increases it becomes log normal

then when d_1 and d_2 are very high it becomes Normal distribution

CDF



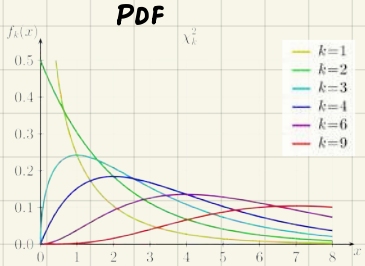
Parameters : $d_1, d_2 > 0$ deg. of freedom

NOTE: In ANOVA test we use this F- distribution table

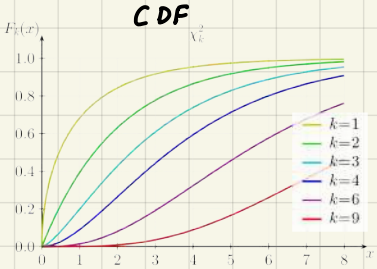
* Chi - Square distribution

This is also a distribution used in Inferential Statistics

PDF



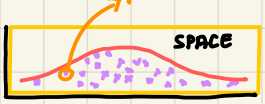
CDF



We can even say that exponential, pareto and Log normal distribution can be a Chi-square distribution

* Hypothesis Testing [Inferential Stats]

① P value



→ Out of 100 touches to the space bar 10 times touching in that area

Hypothesis testing:

Eg: Person → Crime

① Null Hypothesis (H_0): Person has not committed crime

Alternate Hypothesis (H_1): Person has committed crime

② Experiments: Proofs, DNA, fingerprints, evidences ⇒ Judge

Person has committed the crime

Person has not committed crime

③ Reject the null hypothesis

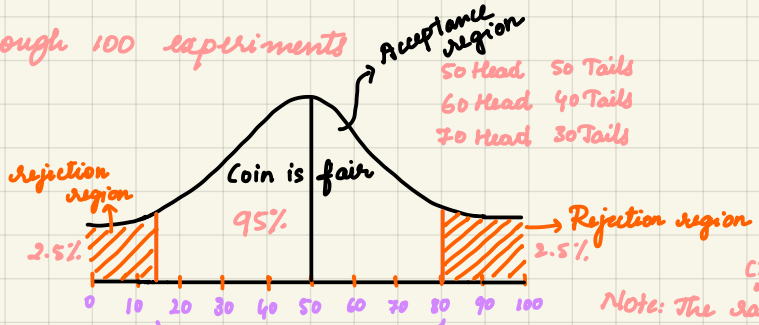
OR

Failed to reject null hypothesis

Eg: Coin is fair or Not through 100 experiments

H_0 = Coin is fair

H_1 = Coin is not fair



Experiment :-

Confidence interval (C.I) = 95%

Significance level (α) = 1 − C.I

= 0.05

P value < 0.05 This means that it is falling in extreme region

Note: The CI range is decided by domain expert "We want 95% confidence interval"

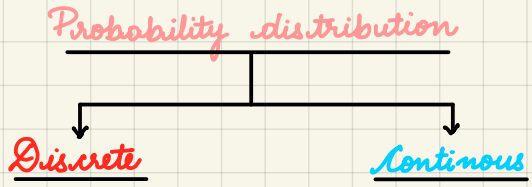
P value is basically used to compare my data point is in acceptance region or rejection region

(p value we compare with significance level)

P value ⇒ possibility value for the null hypothesis to be true

Assignment: Explore and understand other distribution

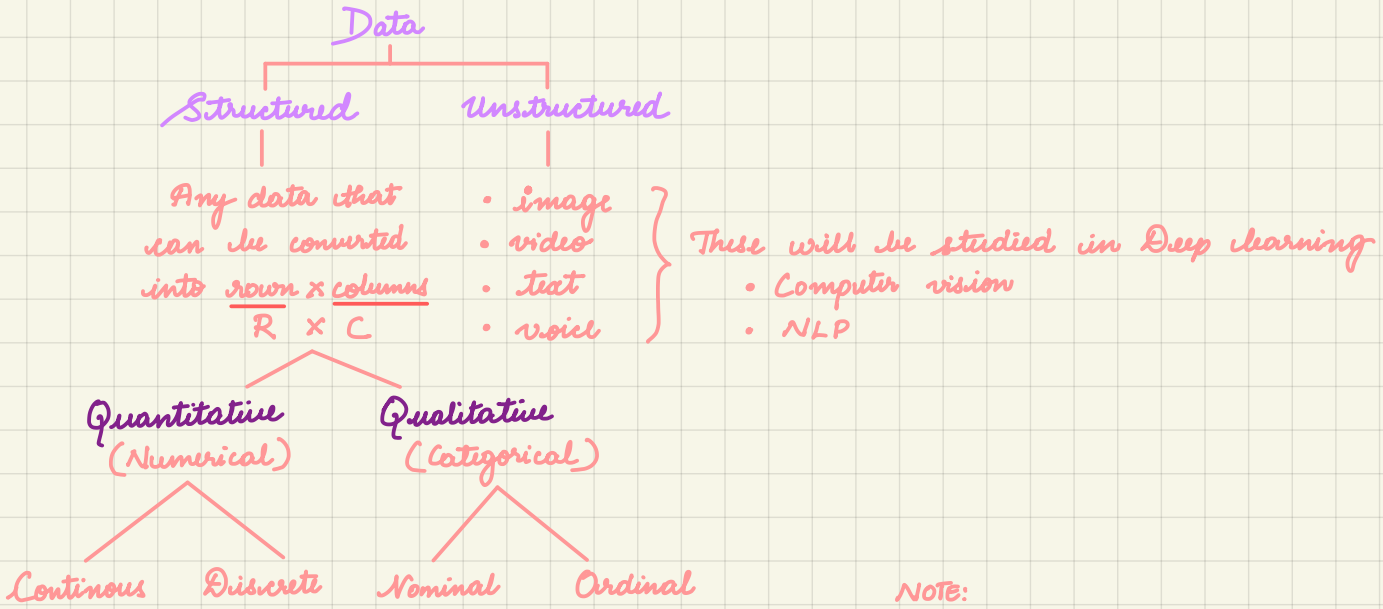
Summary:



- Uniform
- Bernoulli
- Binomial

- Uniform
- Normal
- Log Normal
- Chi-Square
- T- distribution
- F- distribution
- Gamma distribution

Complete Data Hierarchy



NOTE:
data = [23, 24, 32, 45, 12, 43, 67, 45, 32, 56, 32]

Different libraries used in Python for statistics

- Pandas
- numpy {mean, median} Ex: np.mean(data)
- matplotlib and seaborn
- scipy {mode} Ex: from scipy import stats as st | st.mode(data)
- statsmodel
- statistics {mode} Ex: statistics.mode(data)

To find percentile value using python!

→ Ex: 25th percentile, 50th percentile, 75th percentile, 100th percentile

np.percentile(data, [25, 50, 75, 100])

→ array([28., 32., 45., 67.]) where these are q1, q2, q3, q4 values

(NOTE: IQR = q3 - q1)

Note:

lower fence = q1 - IQR * 1.5

Upper fence = q3 + IQR * 1.5

NOTE:

We take n+1 in case of odd numbers

To plot a box plot using python.

→ import seaborn as sns

sns.boxplot(data)

To find the variance of a data in python.

→ np.var(data)

→ 226.23140495867773

Similarly for standard deviation

→ np.std(data) → 15.04099

Sample / population

There are different techniques of sampling

Eg: Random sampling

np.random.choice(data, size=3)

→ array([56, 23, 32])

Assignment:

Find out atleast 5 techniques of sampling and implement it with the help of python

If we want a sample data from the dataframe (15 rows)

data.sample(n=15)

Finding sample and population variance using statistics library.

Sample:

statistics.variance(data)

→ 248.85454

Population:

statistics.pvariance(data)

→ 226.2314

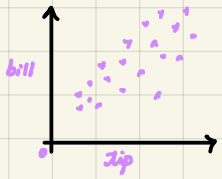
Finding standard deviation using math library

math.sqrt(statistics.variance(data))

→ 15.775

Correlation:

```
import seaborn as sns
df = sns.load_dataset('tips')
df.corr()
sns.scatterplot(x=df['tip'], y=df['total_bill'])
```



Covariance:

df.cov()

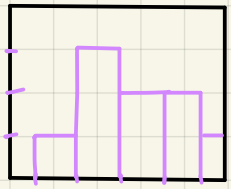
Note: In histogram we can see frequency of a variable (How many times a variable has occured)

Probability density function

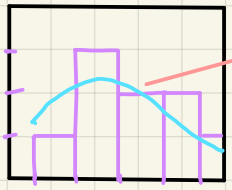
Using histogram in python

① sns.histplot(data)

→



② sns.histplot(data, kde=True)



Kde → Kernel density (Smoothening data)

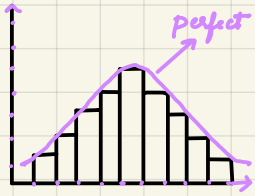
In numpy we have a function which will generate a random data point which will use normal distribution

np.random.normal(mean, standard deviation, points to be generated)

S = np.random.normal(0.5, 0.2, 1000)

→ " — " — "

sns.histplot(S, kde=True)



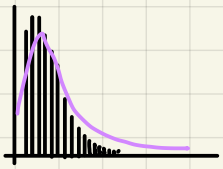
Similarly

mu, sigma = 3.0, 1.0

p = np.random.lognormal(mu, sigma, 1000)

sns.histplot(p, kde=True)

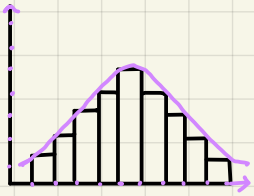
→



Converting log normal into normal

sns.histplot(np.log(p), kde=True)

→



Note: Learn about Pandas profiling, qq plot (quartile - quartile)

1. Central Limit Theorem
2. Z-score and Z-stats
3. Z-test, t-test

} Inferential Stats / Hypothesis testing

1 Central Limit Theorem

The central limit theorem relies on the concept of a sampling distribution which is the probability distribution of a statistics for a large number of samples taken from a population

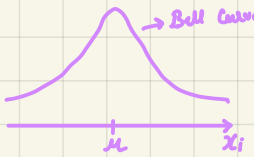
The central limit theorem says that the sampling distribution of the mean will always be normally distributed, as long as the sample size is large enough. Regardless of whether the population has a normal, Poisson, binomial, or any other distribution, the sampling distribution of the mean will be normal

1 $X \approx N(\mu, \sigma)$ (gaussian distribution)

If we have an X random variable following Normal/Gaussian distribution and if we take multiple samples from it ($S_1, S_2, S_3 \cdots S_m$) of some sample size (n) and if we calculate all the sample means and plot it Then all the sample mean will also follow a normal/gaussian distribution

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

n = Sample size ⇒ any value



population distribution

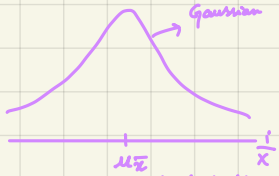
$S_1 = \{x_1, x_2, x_3 \cdots x_n\} = \bar{x}_1$

$S_2 = \{x_5, x_6, x_3 \cdots x_n\} = \bar{x}_2$

$S_3 = \{x_2, x_4, x_7, \cdots x_n\} = \bar{x}_3$

$S_m = \{\quad\} = \bar{x}_4$

⇒

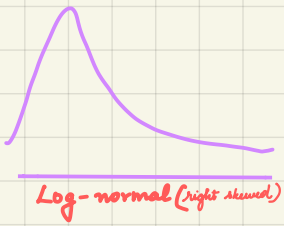


Sampling distribution of the mean

2 $X \not\approx N(\mu, \sigma)$ (Non gaussian distribution)

If we have X random variable not following Normal/gaussian distribution having some mean (μ) and standard deviation (σ) (like log normal)

If we take this we should consider n to be greater than or equal to 30,



Log-normal (right skewed)

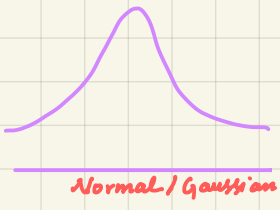
$S_1 = \{$
$S_2 = \{$
$S_3 = \{$

n = Sample size

$n \geq 30$

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

⇒



Normal/Gaussian

Then if I plot this sample mean we will see according to central limit theorem

It will be following a gaussian or Normal distribution

Note:

Population Distribution

$X \approx N(\mu, \sigma)$ where σ = population standard deviation

μ = population mean

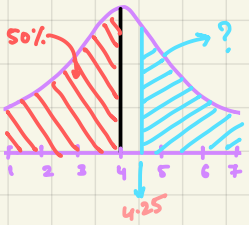
Sampling Distribution of the mean

$\bar{X} \approx N\left(\mu, \left\{\frac{\sigma}{\sqrt{n}}\right\}\right)$ → Standard error

n = sample size

② Z - Score

Consider $X \approx N(4, 1)$

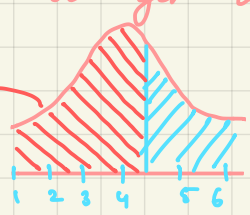


$x_i = 4.25$

$Z\text{-Score} = \frac{4.25 - 4}{\sigma} = \frac{0.25}{1}$ (i.e it is 0.25 std deviation away from mean)

In Z table search for positive 0.2 and 0.05 the value we get will be .59871

i.e 59.87%



→ So this will be

$1 - 0.59871$

$\rightarrow 0.4013$

$\Rightarrow 40.13\%$

Q: What percentage of score lies between 3.5 to 4.5?

Z-table values

$Z\text{-Score}_{(min)} = \frac{3.5 - 4}{\sigma} = \frac{-0.5}{1} \Rightarrow 0.3085$

$Z\text{-Score}_{(max)} = \frac{4.5 - 4}{\sigma} = \frac{0.5}{1} \Rightarrow 0.6915$

Therefore the percentage of score that lies between 3.5 to 4.5 is

$0.6915 - 0.3085$

$= 0.383$

i.e 38.3%

Q: In India the average IQ is 100, with a standard deviation of 15. What is the percentage of the population, would you expect to have an IQ lower than 85?

→ $X \approx N(100, 15)$ $\quad x_i = 85$

$$I\text{-Score} = \frac{85 - 100}{15} \Rightarrow -1 \Rightarrow 0.1587$$

Lower than 85% i.e 15.87%

Greater than 85% ie $1 - 0.1587 \Rightarrow 0.8413 \Rightarrow 84.13\%$

② $75 \leq IQ \leq 100$

$x_i = 75$

→ $I\text{-Score(min)} = \frac{75 - 100}{15} \Rightarrow -1.66 \Rightarrow 0.0485 \Rightarrow 4.85\%$

Therefore percentage of population between 75 to 100 will be

$50\% - 4.85\%$

$\Rightarrow 45.15\%$

Hypothesis Testing and Statistical Analysis

1. Z-test
2. t-test
3. Chi-square
4. ANNOVA

① Z-test:

1. The average heights of all residents in a city is 168 cm with a σ of 13.9. A doctor believes the mean to be different. He measured the height of 36 individuals and found the average height to be 169.5 cm.

a) State null and Alternate hypothesis
b) At a 95% confidence level, is there enough evidence to reject the null hypothesis.

→ Given

$\mu = 168$ cm $\quad \bar{x} = 169.5$ cm

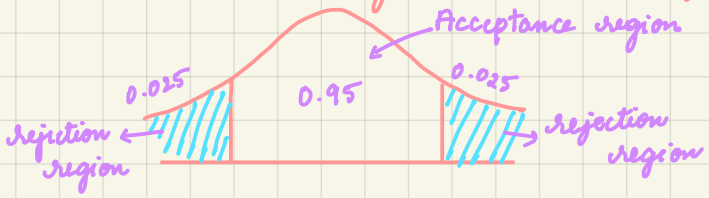
$\sigma = 3.9 \quad$ C.I = 0.95

$n = 36 \quad \alpha = 1 - C.I$

$= 0.05$ //

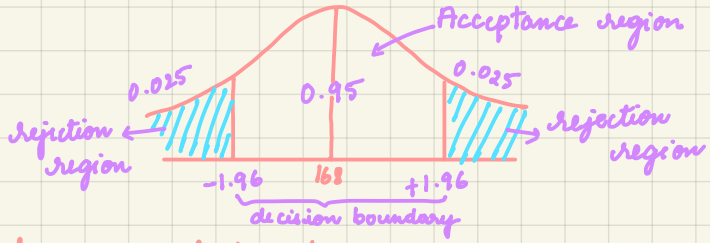
① Null Hypothesis $H_0 : \mu = 168\,cm$
Alternate Hypothesis $H_1 : \mu \neq 168\,cm$

② Decision boundary based on confidence interval



We are checking alternate hypothesis for $\neq 168\,cm$
So our value can be greater than 168 or less than 168 also
That means it can come in left / right extreme ends

So this types of tests are called as "Two tailed tests"
This z-test follows gaussian distribution



Now how can we findout how much standard deviation these lines are away from mean

Total area = 1
Now $1 - 0.025$
$= \overline{0.9750}$ See Z table to find out the standard deviation using the area
we get 1.96
-1.96 and $+1.96$ as both are symmetrical

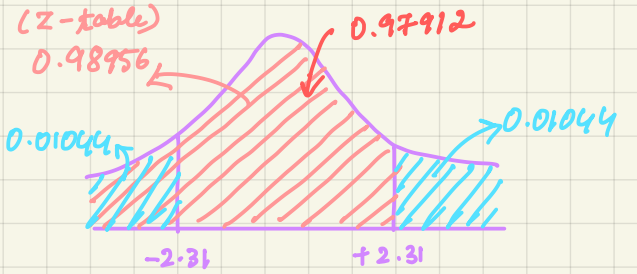
If Z-test is less than -1.96 or greater than $+1.96$, Reject the Null Hypothesis

Whenever we take sampling, our Z-test formula will be

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

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

Conclusion:
As $2.31 > 1.96$ Reject the Null Hypothesis
Therefore the doctor is right about the mean being different



$1 - 0.98956$
$\Rightarrow \underline{0.01044}$

① p value = $0.01044 + 0.01044$
$= \underline{0.02088}$

W.K.T If p value $< 0.05 \Rightarrow$ Reject the Null Hypothesis

② A factory manufactures bulbs with an average warranty of 5 years, with a standard deviation of 0.50. A worker believes that the bulb will malfunction in less than 5 years.
He tests a sample of 40 bulbs and finds the average time to be 4.8 years
a). State null and alternate hypothesis
b). At a 2% significance level, is there enough evidence to support the idea that the warranty should be revised?
(Take C.I as 99%)

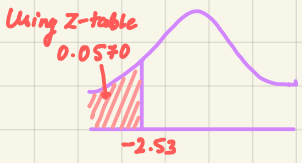
⟶ Given
$\mu = 5$, $\sigma = 0.50$, $n = 40$, $\bar{x} = 4.8$

a) Null Hypothesis $H_0 : \mu = 5$
Alternate Hypothesis $H_1 : \mu < 5$ {One tailed test}

b) Decision Boundary

If z test is < -2.05 Reject the Null Hypothesis

Z-test $= \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \dfrac{4.8 - 5}{0.50/\sqrt{40}} = \underline{-2.53}$



p-value = 0.0570
$\therefore$ p-value < 0.02 (Significance value)
$\Rightarrow$ False
Reject the null hypothesis

$-2.53 < -2.05$ True
$\therefore$ Reject the null hypothesis

T-Test

major diff b/w T-Test & Z-test is in Z-test we get population S.D whereas in T-test we get Sample Standard Deviation

② 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 sample of 30 participants who have taken a medication has a mean of 140 with a standard duriation of 20. Did the medication effect Intelligence? Take C.I = 95%, $\alpha = 0.05$

2 tailed

⟶ Given:

$\mu = 100$, $n = 30$, $\bar{x} = 140$, $S = 20$, CI = 95%, $\alpha = 0.05$

① Null Hypothesis (H_0) = $\mu = 100$

Alternate hypothesis (H_1) = $\mu \neq 100$

②. $\alpha = 0.05$

Whenever we need to solve T-Test here we need to compute Degree of freedom

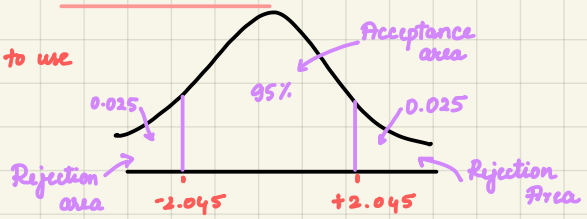
$dof = n - 1$

$= 30 - 1$

$= 29$ //

③ Decision Rule

→ We need to use t-table



If t-test is less than -2.045 and greater than 2.045, Reject the null hypothesis

* T-test statistics

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

Therefore

$t > 2.045$ Reject the null hypothesis

Conclusion: Medication has a positive effect on intelligence

Agenda:

1. Point Estimate
2. Range of Confidence Interval (C.I)
3. Chi Square distribution
4. F distribution → F Test
5. [ANOVA] → Assignment

① Point estimate

The value of any statistics that estimates the value of a parameter is called point estimate

point estimate

Ex: $\bar{x}$ ⟶ μ ← parameter

$\bar{x} = 2.95$ $\quad \mu = 3.00$

$\bar{x} = 3.5$ $\quad \mu = 3.00$

Here we can say that ($\bar{x}$) sample is point estimate of the population (μ)

(point estimate can be less than population mean or greater than PM)

Hence

We rarely know if our point estimate is correct because it is an estimation of actual value

We construct Confidence Interval (C.I) to help estimate what the actual value of unknown population mean is

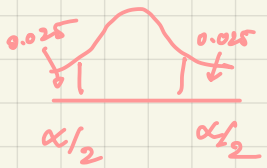
Point estimate ± margin of error

Lower Range CI = Point estimate - margin of Error

Higher Range CI = Point estimate + margin of Error

① On the verbal section of CAT exam, a sample of 25 test takers has a mean of 520, with a standard deviation of 80. Construct a 95% C.I about the mean?

(T-test)



$t_{df/2}$ → T table using $n = 24$, $\alpha = 0.05$

⟶ $\bar{x} = 520$, $n = 25$, $S = 80$, $CI = 0.95$, $\alpha = 0.05$

C.I = Point Estimate ± Margin of Error

$= \bar{x} \pm t_{df/2}\left(\frac{S}{\sqrt{n}}\right)$ ⟶ Standard error

whenever we do t-test, we find degree of freedom

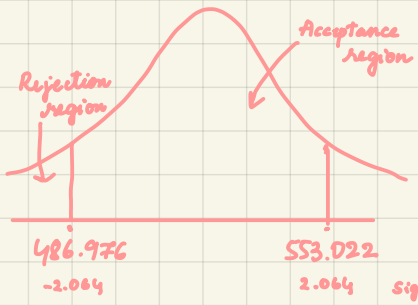
$df = n - 1$

$= 24$

$= 520 \pm 2.064\left(\frac{80}{\sqrt{25}}\right)$

$\therefore$ Lower (C.I) $= 520 - 2.064\left(\frac{80}{\sqrt{25}}\right) \Rightarrow \underline{486.976}$

$\therefore$ Higher (C.I) $= 520 + 2.064\left(\frac{80}{\sqrt{25}}\right) \Rightarrow \underline{553.022}$



Chi Squared Test:

The chisquare test for goodness of fit tests claims about population proportion
[categorical variables]

It is a non parametric test that is performed on categorical data
[ordinal, nominal data]

Eg: There is a population of males that like different colors of bike

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

Theory ⇓ Theoretical Categorical distribution

Sample ⇒ Observed categorical distribution

Goodness of fit: Sample info i am trying to fit in population info

NOTE: According to chiquare what we have to do is that w.r.to Sample data we need to come to a conclusion if Theory is true or not

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

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

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

<50kg	50-75	>75
140	160	200

⇒ (observed)

Using $\alpha = 0.05$, would you conclude that the population differences of weights has changed in last 10 years? C.I = 95%

→ In 2010 expected

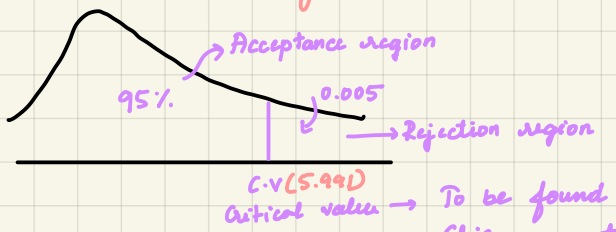
<50kg	50-75	>75
500×0.2 = 100	500×0.3 = 150	500×0.5 = 250

① Null hypothesis H_0 : The data meets the expectation
Alternate hypothesis H_1: The data does not meet the expectation

② $\alpha = 0.05$, CI = 95%.

③ Degree of freedom: dof = No of category − 1 ⇒ 3 − 1 ⇒ 2

④ Decision boundary



→ To be found out using Chisquare table

using dof and α value
2 0.05

C.V = 5.991
critical value

If Chisquare test $X^2 > 5.991$ {Reject the Null Hypothesis}

⑤ Chi square test statistics

$$X^2 = \sum \frac{(O-E)^2}{E}$$ Summation of Observed minus Expected whole square devided by Expected

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

$$X^2 = 26.67$$

Therefore
$X^2 > 5.991$ {Reject the Null hypothesis}

F-Distribution

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

$$X = \frac{S_1/d_1}{S_2/d_2}$$

where S_1 → Indipendent random variable
S_2 → Indipendent random variable
} follows Chi-Square distribution
d_1 → Degree of freedom (S_1)
d_2 → Degree of freedom (S_2)

This will be important for another test called F test

F-test: Variance Ratio Test {Comparing variance between 2 groups}

Q. The following data shows the number of bulbs produced daily for some days by 2 workers A and B.

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

Can we Consider based on the data that worker 'B' is more stable or not and efficient
$\alpha = 0.05$, C.I = 95%

→ Null Hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ (variance of 1st sample is equal to variance of 2nd sample)

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

Now let's calculate variance

A

X_1	$\bar{x}$	$(X_i - \bar{x})^2$
40	37	9
30	37	49
38	37	1
41	37	16
38	37	1
35	37	4

$\bar{X_1} = 37$ $\quad \sum (x_i - \bar{x})^2 = 80$

$$S_1^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} = \frac{80}{5}$$

$$S_1^2 = \underline{\underline{16}}$$

B

X_2	$\bar{x}$	$(X_i - \bar{x})^2$
39	37	4
38	37	1
41	37	16
32	37	25
33	37	16
39	37	4
40	37	9
34	37	9

$\bar{X_2} = 37$ $\quad \sum (x_i - \bar{x})^2 = 84$

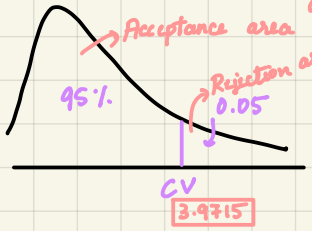
$$S_2^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} = \frac{84}{7}$$

$$S_2^2 = \underline{\underline{12}}$$

* Calculating Variance Ratio [F-Test]

$$F = \frac{S_1^2}{S_2^2} \Rightarrow \frac{16}{12} \Rightarrow \underline{\underline{1.33}}$$

* Decision Boundary [F distribution]



$df_1 = 6 - 1 \Rightarrow \underline{5}$

$df_2 = 8 - 1 \Rightarrow \underline{7}$

for critical value refer F table for α value 0.05 ($df_1 \rightarrow$, $df_2 \downarrow$)
(Column 5, row 7)

we get 3.9715 → Critical value

Therefore

If F test > 3.9715 {Reject the Null Hypothesis}

As 1.33 is not greater than 3.9715 (False)
We fail to Reject the null hypothesis

Worker A ≈ Worker B (in term of efficiency and work)