PW SKILLS

Introduction to Statistics

Defn: Statistics is the science of collecting, organizing and analyzing data

⇓

Decision MAKing Process

Data: "facts or pieces of information"

Eg: Height of students in classroom

{175cm, 18 cm, 190cm, 160 cm ...}

IQ

{100, 90, 95, 80 ...}

② Types of Statistics

① Descriptive Statistics

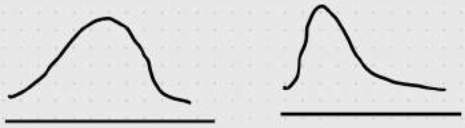
Defn: It consists of organizing and summarizing data

① Measure of Central Tendency [Mean, Median, Mode]

② Measure of Dispersion [Variance, Std]

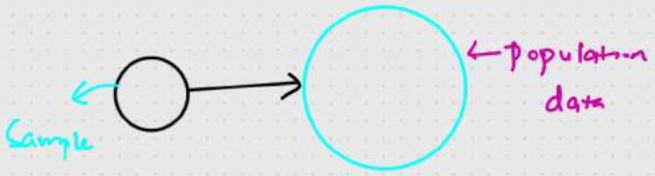
③ Different type of Distribution of data.

Eg: Histogram, pdf, pmf



② Inferential Statistics

Defn: It consists of using data you have measured to form conclusion



① Z-test
② t-test
③ CHI SQUARE

} Hypothesis Testing
Mo, M, p value, Significance value

Eg: Lets say there are 20 Statistics classes at your college. and you have collected the heights of students in the class.

Heights are recorded [175cm, 180cm, 140, 140, 135, 160, 135, 190cm]

$$\frac{175+180+140+140+135+160+135+190}{8} = \text{Avg Height}$$

Descriptive Question

"What is the average height of the entire classroom"

Inferential Question

Sample

"Are the height of the Students in classroom similar to what you expect in the entire college"

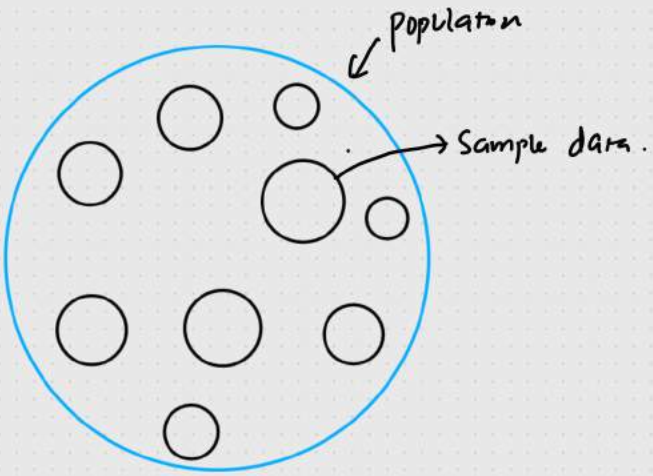
↳ population data

Population And Sample Data

Population: The group you are interested in studying

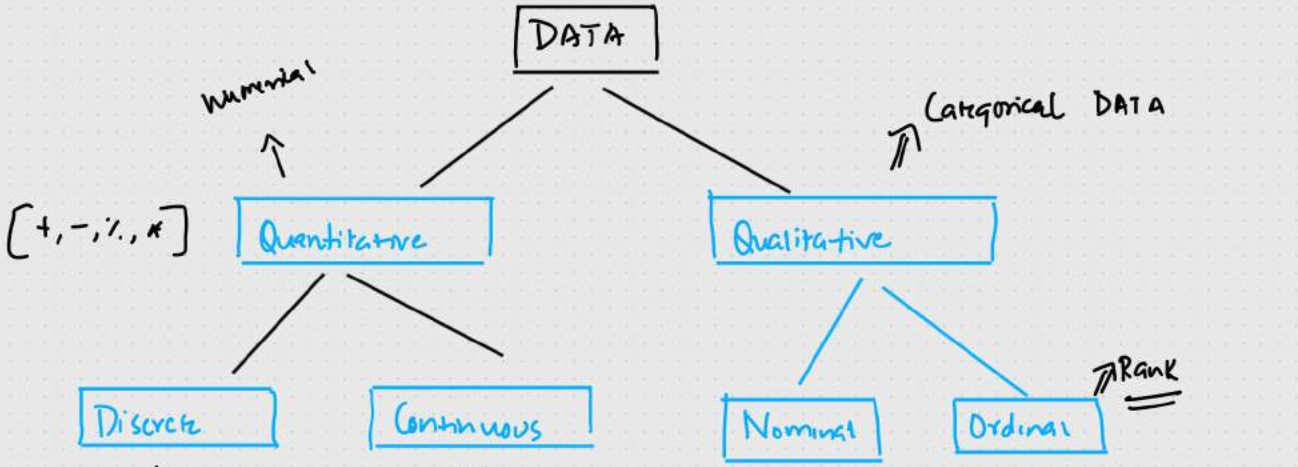
Sample : a subset of population

Eg: Exit Poll



③ Types of Data

DMC	DC	ISI	BUI	FWI	Classes	Region
3.4	7.6	1.3	3.4	0.5	not fire	0
4.1	7.6	1	3.9	0.4	not fire	0
2.5	7.1	0.3	2.7	0.1	not fire	0
1.3	6.9	0	1.7	0	not fire	0
3	14.2	1.2	3.9	0.5	not fire	0
5.8	22.2	3.1	7	2.5	fire	0
9.9	30.5	6.4	10.9	7.2	fire	0
12.1	38.3	5.6	13.5	7.1	fire	0



④ Scale of Measurement

1. Nominal Scale Data
2. Ordinal Scale Data
3. Interval Scale Data
4. Ratio. Scale Data.

① Nominal Scale Data

1. Qualitative / Categorical
2. Eg : Gender, Colors, Fabrics
3. Order does not matter

Eg : Favorite Color

Red → 5 → 50%

Blue → 3 → 30%

Orange → 2 → 20%

10

Gender

M

F

② Ordinal Scale Data

1. Ranking is important
2. Order matter
3. Difference cannot be measured

Eg :

1 → Best

2 → Good

3 → Bad

Working Professional

M

T

W

Th

← F

Race :

1st

2nd

3rd

4:20

5:30

6:00

OF

③ Interval Scale Data

① The order matter

② Difference can be measured

③ Ratio cannot be measured ✓

④ No "0" Starting Point

Eg : Temperature Variable.

-30° F

-15° F

30 F

60 F

90 F

120 F

60:30 = 2:1 →

90 - 60 = 30 F

120 - 90 = 30 F

④ Ratio Scale Data

① The order matter ✓

② Differences are measureable (including ratio)

③ Contains a "0" Starting point.

Eg : Students marks in a class

0, 90, 60, 30, 75, 40, 50

Asc = 30, 40, 50, 60, 75, 90

40 - 30 = 10

50 - 30 = 20

Ratio = $\frac{90}{30}$ = 3 : 1

Example

① length of Different Rivers In the World ?

② Favorite food based on Gender ?

③ Marital Status ?

④ IQ measurement ?

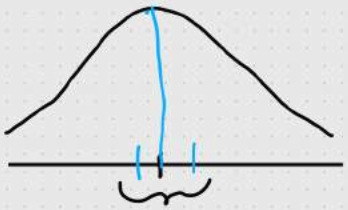
PW SKILLS

① Measure of Central Tendency

① Mean or Average

② Median

③ Mode



① Mean

Population (N) | Sample (n)

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

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

$= \left[\frac{1+1+2+2+3+3+4+5+5+6}{10}\right]$ =

$= \frac{32}{10} = 3.2$ //.

② Median

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

① Sort the Random Variable $\{1, 2, 2, 3, 4, 5\}$.

② No. of elements Count = 6

③ if Count == even

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

⇓

Median = $\frac{2+3}{2} = 2.5$

④ if Count is Odd

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

Median = 3

Why Median?

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

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

$\bar{x} = \frac{15}{5} = 3$

Outlier ⇓

$X = \{1, 2, 3, 4, 5, 100\}$

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

$\frac{115}{6} \approx 19.$

$X = \{1, 2, 3, 4, 5, 100\}$

Median = $\frac{3+4}{2} = 3.5$

Median is used to find the central Tendency when outliers is present.

③ Mode : Frequency Maximum

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

Mode = 1

EDA AND Feature Engineering

Age ⇓	Weight ⇓	Salary	Gender (Mode)	Degree (Mode)
24	70	40K	M	BE
25	80	70K	F	-
21	95	45K	F	-
→ 24	-	50K	M	PHD
→ 32	-	60K	-	B.E
→ -	60	-	-	Master
→ -	65	55K	-	Bsc
→ 40	72	-	M	B.E

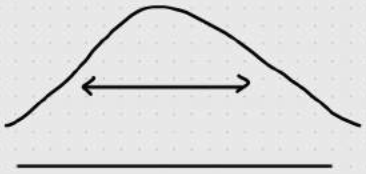
DATA IS MISSING

PW SKILLS

① Measure of Dispersion [Spread of the data]

① Variance

② Standard deviation



① Variance

Population Variance

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

$x_i \rightarrow$ DATA POINTS

$\mu \rightarrow$ Population mean

$N \rightarrow$ Population size

Sample Variance

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

$x_i \rightarrow$ DATA POINTS

$\bar{x} \rightarrow$ Sample Mean

$n \rightarrow$ Sample size

Why we divide Sample Variance by $n-1$? ⇑ Bessel's Correction

Ans) The sample variance is divide by $n-1$ so that we can create an unbiased estimator of the population variance

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

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

$n = 5$

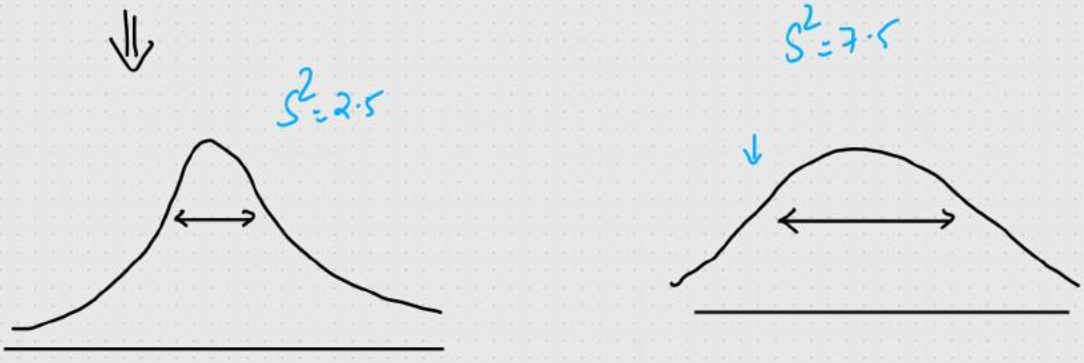
x	$\bar{x}$	$(x_i - \bar{x})^2$
1	3	4
2	3	1
3	3	0
4	3	1
5	3	4
3		10

$$S^2 = \frac{10}{4-2} = \frac{5}{2} = 2.5$$

$X = \{\quad\}$ $\quad$ $Y = \{\quad\}$

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

Dispersion or Spread



② Standard Deviation

Population Standard Deviation

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

Sample Std

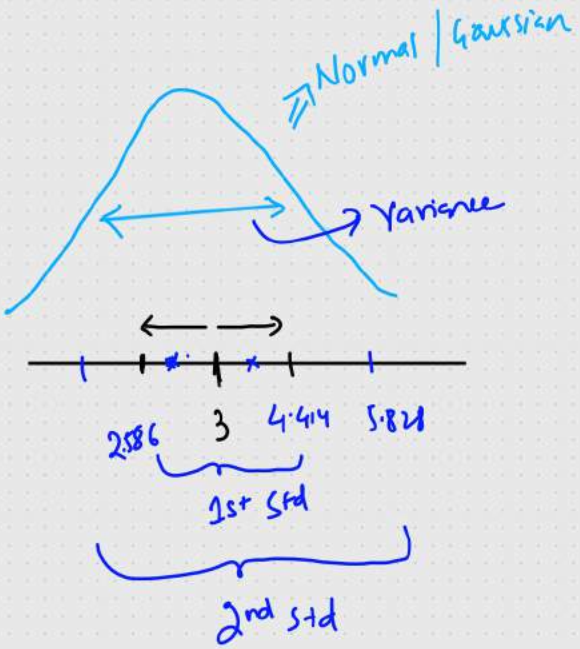
$$\text{Std} = \sqrt{s^2}$$

s^2 = Sample Variance

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

$$\bar{X} = 3$$

$$\sigma = 1.414$$



3.000
1.414
2.586

4.414
1.414
5.828

① Random Variable

x, y → variables

$x + 5 = 7$ $\quad x = 2$

$8 = y + x$ $\quad y = 6$

Random Variable is a process of mapping the output of a random process or experiments to a number.

Eg: Tossing a Coin

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

Rolling a dice

Measure the Temperature of the next day

$$Y = \begin{cases} \text{Sum of the rolling of dice 7 times} \\ \Downarrow \end{cases}$$

$P_r(Y \geq 15)$ $\qquad$ $P(Y < 10)$

$X = \{-, -, -, -, \}$ ⇒ Ages

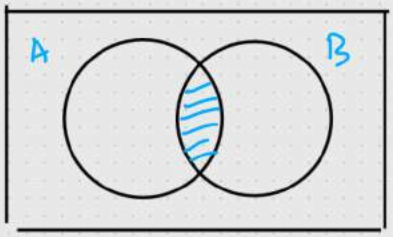
① Sets

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

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

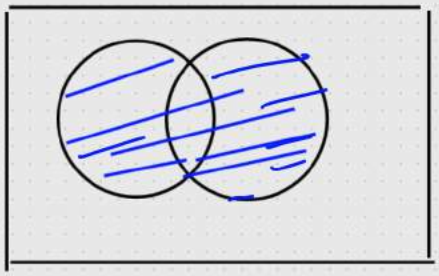
① Intersection

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



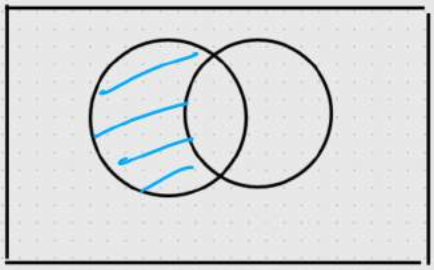
② Union

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



③ Difference

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



④ Subset

$A \longrightarrow B \Rightarrow$ False

$B \longrightarrow A \Rightarrow$ True

⑤ Superset

$A \to B \Rightarrow$ True

$B \to A \Rightarrow$ Fals

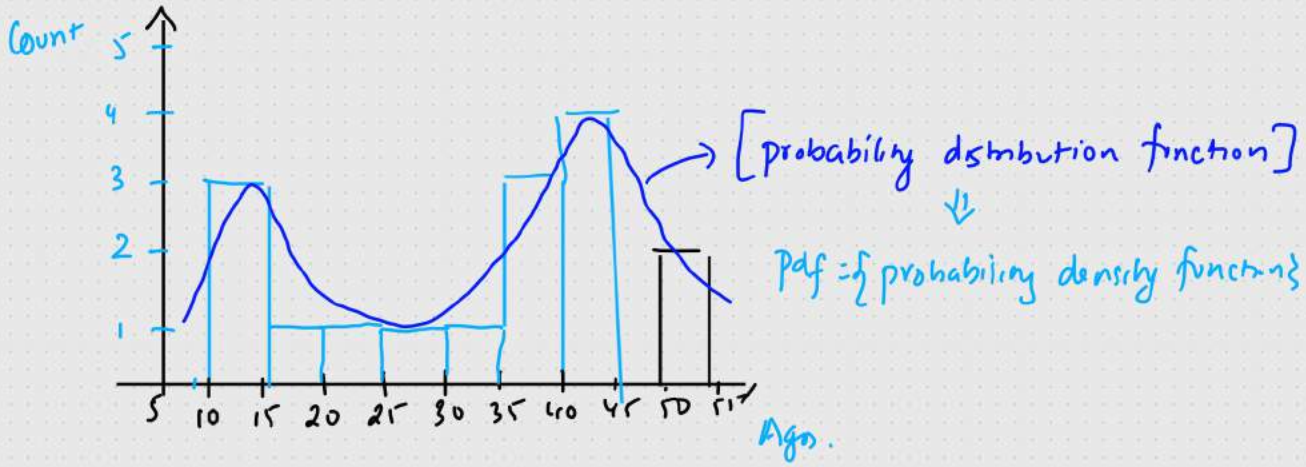
PW SKILLS

Histograms And Skewness → [Frequency]

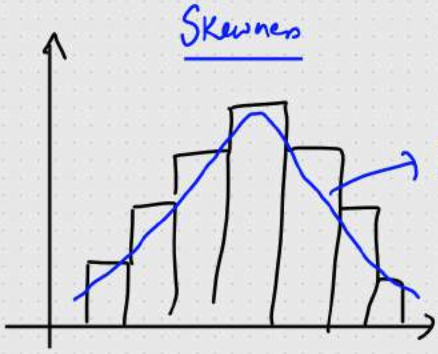
Ages = {10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 48, 50, 51}

$\frac{50}{10} = 5$ → bin size {No. of bins = 10}

$\frac{50}{20} = 2.5$ → bin size {No. of bins = 20}

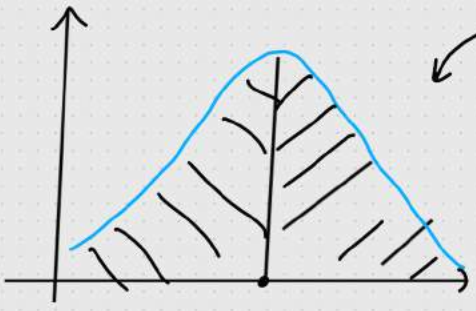


Skewness



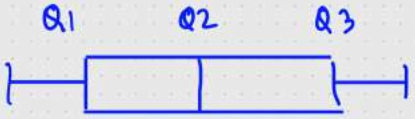
Normal / Gaussian Distribution

Symmetrical Distribution



① No skewness

Box plot

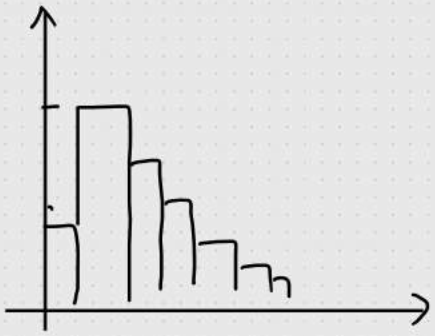


$Q_3 - Q_2 \approx Q_2 - Q_1$

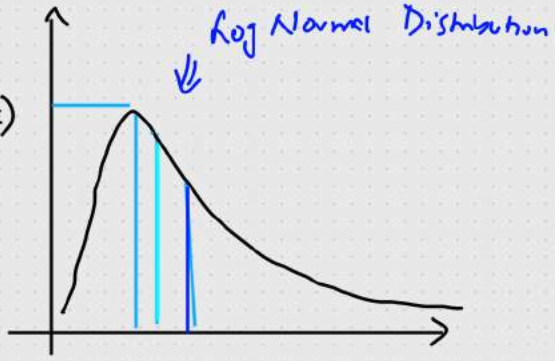
The mean, median, and mode all are perfectly at the center

mean = median = mode

② Right skewed



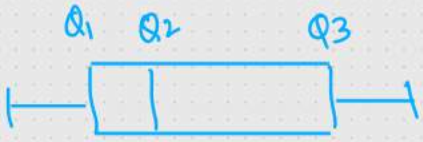
⇒ Positive Skewed ⇒



mean ≥ median ≥ mode

Relationship between Mean, Median, Mode

Box plot



$Q_3 - Q_2 \geq Q_2 - Q_1$

③ Left skewed Distribution

⇒ Negative Skewed

$Q_2 - Q_1 \geq Q_3 - Q_2$

Relationship : mean ≤ median ≤ mode

PW SKILLS

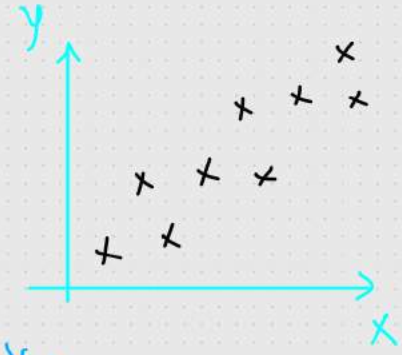
Covariance And Correlation

[Relationship between X and Y]

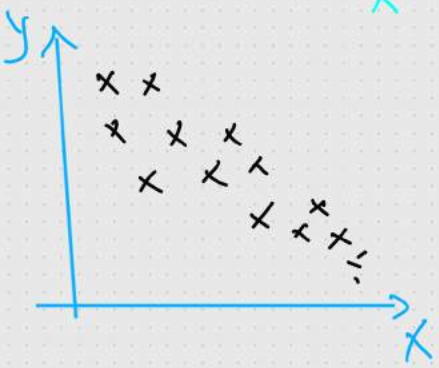
X	y
2	3
4	5
6	7
8	9

- $X\uparrow \; y\uparrow$
- $X\downarrow \; y\uparrow$
- $X\downarrow \; y\downarrow$
- $X\uparrow \; y\downarrow$

↓↑ Size of House → Price of House ↑↓ ↓o/p



- $X\uparrow \; y\uparrow$
- $X\downarrow \; y\downarrow$



- $X\uparrow \; y\downarrow$
- $X\downarrow \; y\uparrow$

Covariance

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

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

$= Cov(X,X) \Rightarrow$ Spread

$x_i \rightarrow$ Datapoints of x

$\bar{x} \rightarrow$ Sample mean of x

$y_i \rightarrow$ Datapoints of y

$\bar{y} \rightarrow$ Sample mean of y

$Cov(x,y)$

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

+ve Covariance

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

-ve Covariance

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

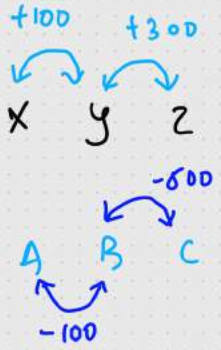
X	y
→ 2	3
→ 4	5
→ 6	7
$\bar{x} = 4$	$\bar{y} = 5$

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

$$= \frac{4+0+4}{2} = \frac{8}{2} = 4 \text{ +ve value}$$

⇓

Positive Covariance



X & y are having a positive Covariance

Advantages

① Relationship between X and y

+ve or -ve value

Disadvantages

① Covariance does not have a specific limit value

② Pearson Correlation Coefficient [-1 to 1]

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

① The more the value towards +1 the more +ve Correlated it is (x, y)

② The more the value towards -1 the more -ve correlated it is (x, y)

③ Spearman Rank Correlation [-1 to 1]

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

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

Feature Selection

+ve	+ve	+ve	≈0	-ve	
Size of House ↑	No. of Rooms ↑	Location ↑	No. of people staying	Maintd	Price ↑

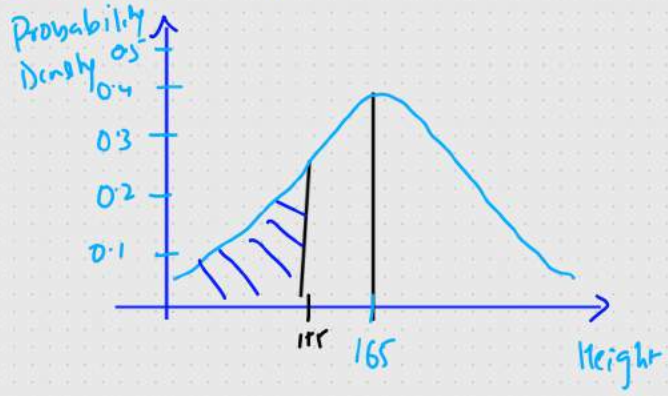
PW SKILLS

Probability Distribution Function / Density Function (pdf) (pmf)

① Probability Density Function (pdf).

① Continuous Random Variable

Eg: Heights of students in classroom [0 - 1]



$Pr(X \leq 155)$ = Area under the curve

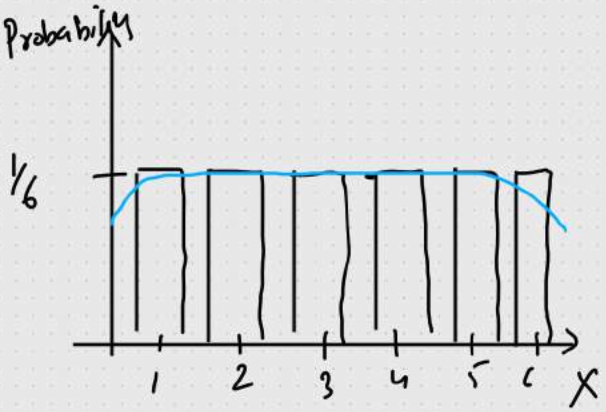
② Probability Mass Function (pmf)

Variable ⟶ Discrete Random Variable

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

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

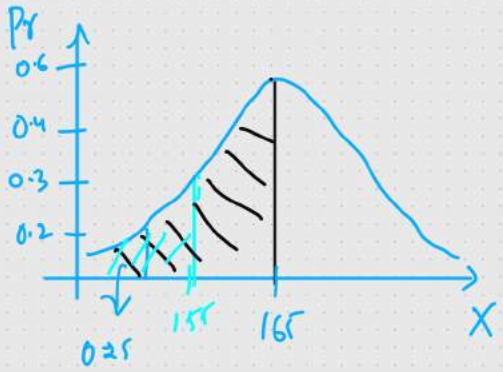
$Pr(2) = \frac{1}{6}$

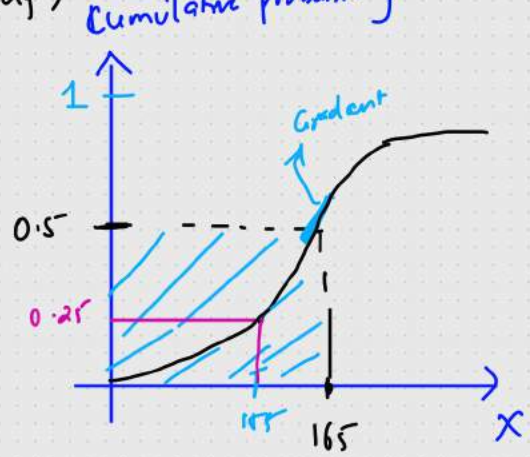


$Pr(X \leq 4) = Pr(X=1) + Pr(X=2) + P(X=3) + Pr(X=4)$

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

③ Cumulative Distribution Function (cdf)





Probability Density Function & Probability Mass Function

Cumulative Distribution Function

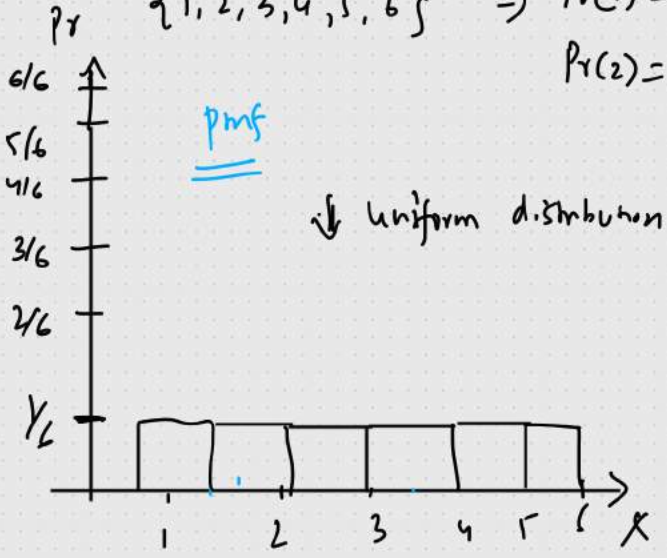
① PMF

① Discrete Random Variable

Eg: Rolling a dice

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

$Pr(2) = 1/6$

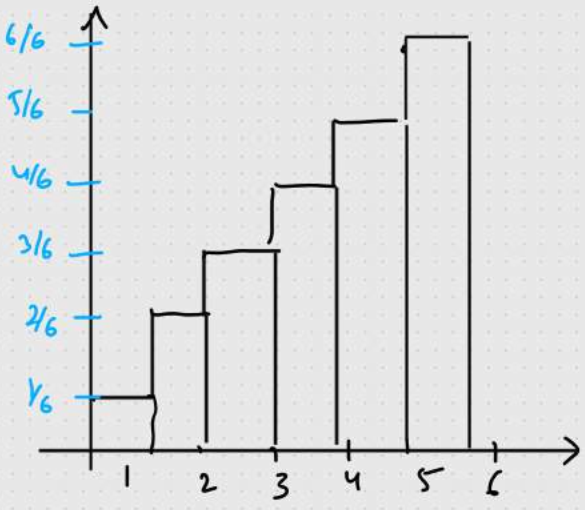


pmf

↓ uniform distribution

CDF

Cumulative probability



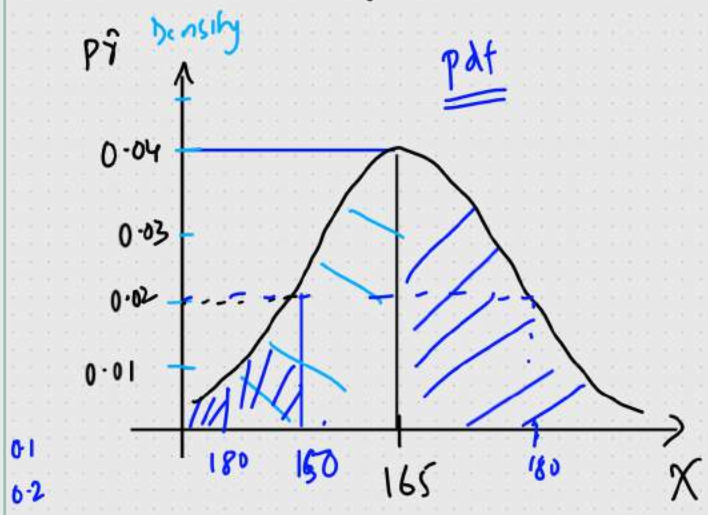
$Pr(X \leq 2) = Pr(X=1) + Pr(X=2)$

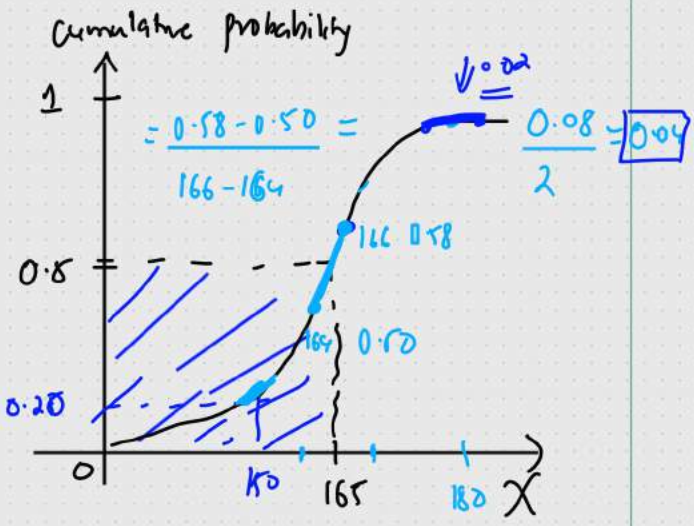
$= \frac{1}{6} + \frac{1}{6} = 2/6$

$Pr(X \leq 6) = 1$

(2) Probability Density Function (pdf)

(1) Distribution of Continuous Random Variable

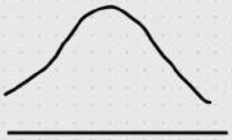




Probability Density → Gradient of Cumulative Curve

PW SKILLS

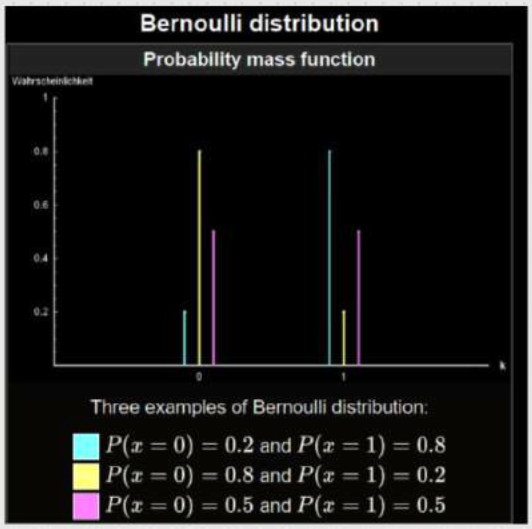
Types of Probability Distribution



1. Normal / Gaussian Distribution (pdf)
2. Bernoulli Distribution (pmf) → Outcomes are binary only
3. Uniform Distribution (pmf)
4. Poisson Distribution (pmf)
5. Log Normal Distribution (pdf)
6. Binomial Distribution (pmf)

Bernoulli distribution

In probability theory and statistics, the Bernoulli distribution, named after Swiss mathematician Jacob Bernoulli,is the discrete probability distribution of a random variable which takes the value 1 with probability p and the value 0 with probability q=1-p. Less formally, it can be thought of as a model for the set of possible outcomes of any single experiment that asks a yes–no question.



① Discrete Random Variable {pmf}

② Outcomes are Binary

Eg: ① Tossing a coin {H, T}

$Pr(H) = 0.5 = p$

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

② Whether the person will Pass / Fail

$Pr(Pass) = 0.7 = p$

$Pr(Fail) = 1 - 0.7 = 0.3 = q$.

Parameters

$0 \leq p \leq 1$

$q = 1-p$

$K = \{0, 1\} \rightarrow 2$ outcomes

① PMF

$$PMF = p^k * (1-p)^{1-K} \quad K \in \{0,1\}$$

if $K = 1$

$$Pr(K=1) = p^1 (1-p)^{1-1}$$

$$= p \;//$$

Simplified

$$pmf \begin{cases} q = 1-p & \text{if } K=0 \\ p & \text{if } K=1 \end{cases}$$

if $K = 0$

$$Pr(K=0) = p^{0} * (1-p)^{1-0}$$

$$= (1-p) = q \;//$$

② Mean of Bernoulli Distribution

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

$Pr(K=1) = 0.6 \Rightarrow p$

$Pr(K=0) = 0.4 \Rightarrow 1-p$

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

$$= 0.6 = p \;//$$

③ Median of Bernoulli Distribution

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

④ Variance

$$\text{Var} = p * (1-p)$$

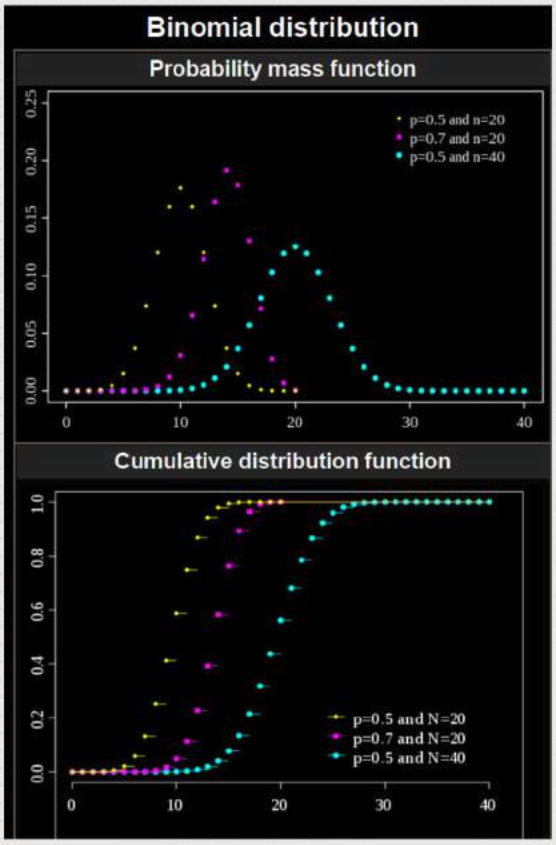
$$= pq \;//$$

Std

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

Binomial Distribution

In probability theory and statistics, the binomial distribution with parameters n and p is the discrete probability distribution of the number of successes in a sequence of n independent experiments, each asking a yes–no question, and each with its own Boolean-valued outcome: success (with probability p) or failure q=1-p. A single success/failure experiment is also called a Bernoulli trial or Bernoulli experiment, and a sequence of outcomes is called a Bernoulli process; for a single trial, i.e., n = 1, the binomial distribution is a Bernoulli distribution. The binomial distribution is the basis for the popular binomial test of statistical significance.



① Discrete Random Variable.

① Every experiment outcome is binary

② These experiment is performed for n trials.

Eg: Tossing a coin 10 times

Notation : $B(n,p)$

Parameters : $n \in \{0,1,2,\cdots\}$ → no. of trials

$p \in [0,1]$ → success probability for each trial

$q = 1-p$

Support : $K \in \{0,1,2,\cdots n\}$ → Number of succes

PMF :

$$Pr(K, n, p) = {}^{n}C_{K}\, p^{K} (1-p)^{n-K}$$

for $K = 0, 1, 2, \ldots n$ where

$$\boxed{{}^{n}C_{K} = \frac{n!}{K!(n-K)!}}$$

Mean

$$\text{Mean} = np$$

Variance

$$\text{Var} = npq$$

Std $= \sqrt{npq}$.

Poisson Distribution

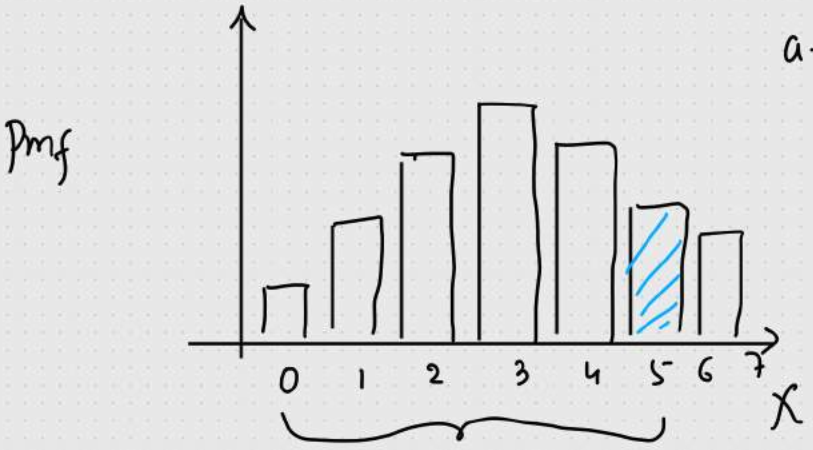
① Discrete Random Variable (pmf)

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

Eg: No. of people visiting hospital every hour

No. of people visiting banks every hour

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



PMF

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

$\lambda = 3$

$$= \frac{e^{-3} 3^{5}}{5!} = 0.101 = 10.1\%$$

λ

$P(x=4) + P_{x}(x=5)$

Mean of Poisson Distribution

λ = Expected No of events to occur at every time interval

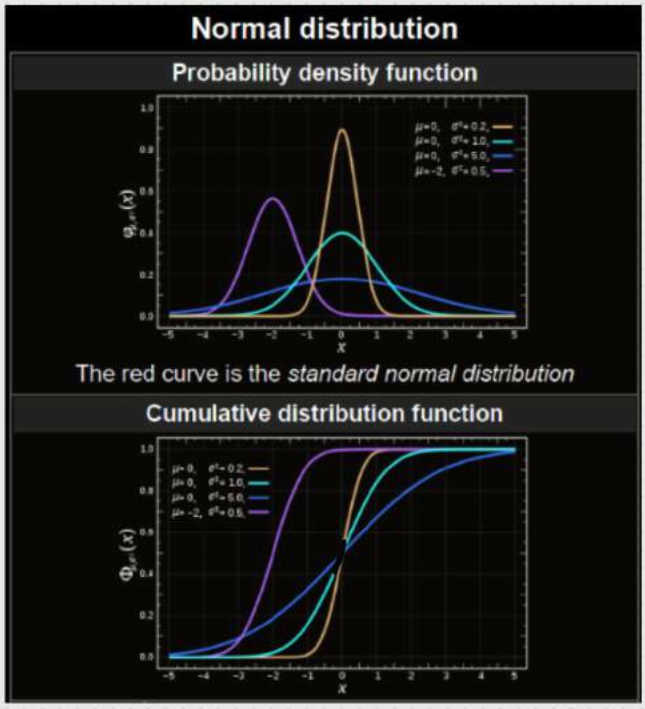
t = Time interval

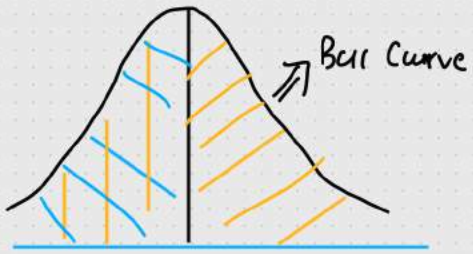
$$\text{Mean} = E(x) = \mu = \lambda * t$$

Variance = 

Normal / Gaussian Distribution (pdf)

In statistics, a normal distribution or Gaussian distribution is a type of continuous probability distribution for a real-valued random variable.





μ = median = mode

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

Parameters : $\mu \in R$ = mean

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

$x \in R$

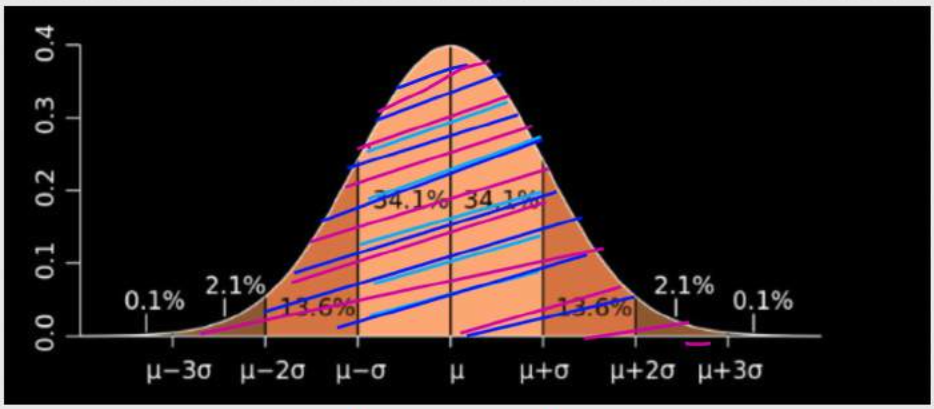
$$\text{PDF} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2}$$

Mean of Normal Distribution

Mean = μ = Average

Variance & Std $\quad Var = \sigma^2 \quad \sigma = \sqrt{Var}$

Empirical Rule of Normal Distribution



68-95-99.7%

100

$X = \{$ $\}$

Probability

$$Pr(\mu - \sigma \le X \le \mu + \sigma) \approx 68\%$$

$$Pr(\mu - 2\sigma \le X \le \mu + 2\sigma) \approx 95\%$$

$$Pr(\mu - 3\sigma \le X \le \mu + 3\sigma) \approx 99.7\%$$

Examples : ① Weight of the student in the class

② Height of the " " "

③ IRIS DATASET {Sepal width}

Q-Q plot { Quantile Quantile Plot }

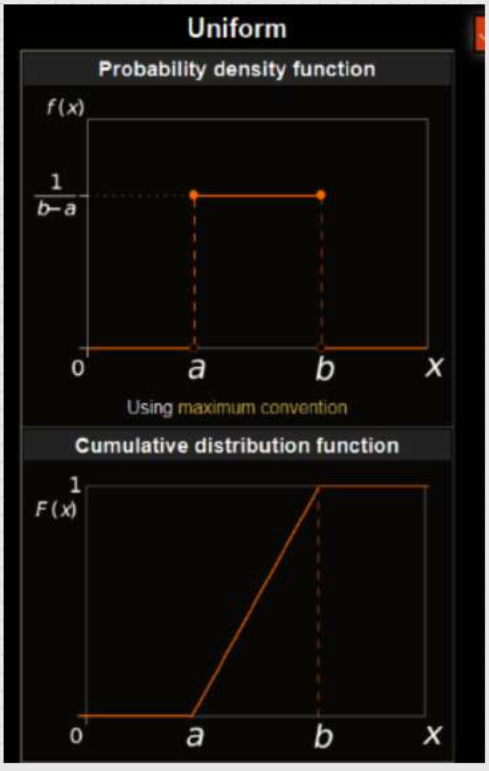
Uniform Distribution

① Continuous Uniform Distribution (pdf)

② Discrete Uniform Distribution (pmf)

① Continuous Uniform Distribution [Continuous Random Variable]

In probability theory and statistics, the continuous uniform distribution or rectangular distribution is a family of symmetric probability distributions. The distribution describes an experiment where there is an arbitrary outcome that lies between certain bounds. The bounds are defined by the parameters, a and b, which are the minimum and maximum values



Notation : $U(a,b)$

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

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

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

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

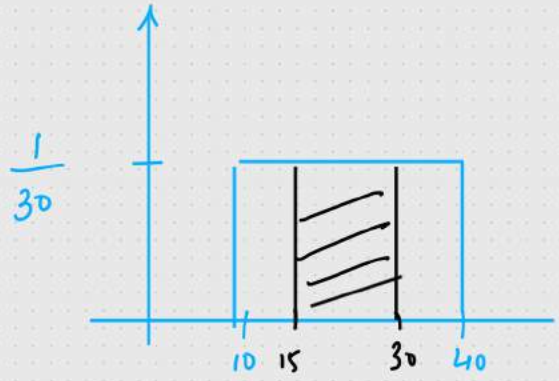
Variance $= \frac{1}{12}(b-a)^2$

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

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

i) Probability of daily sales to fall between 15 and 30?

Ans)



$x_1 = 15$

$x_2 = 30$

$$Pr(15 \leq X \leq 30) = (x_2 - x_1) * \frac{1}{b-a}$$

$$= 15 * \frac{1}{30}$$

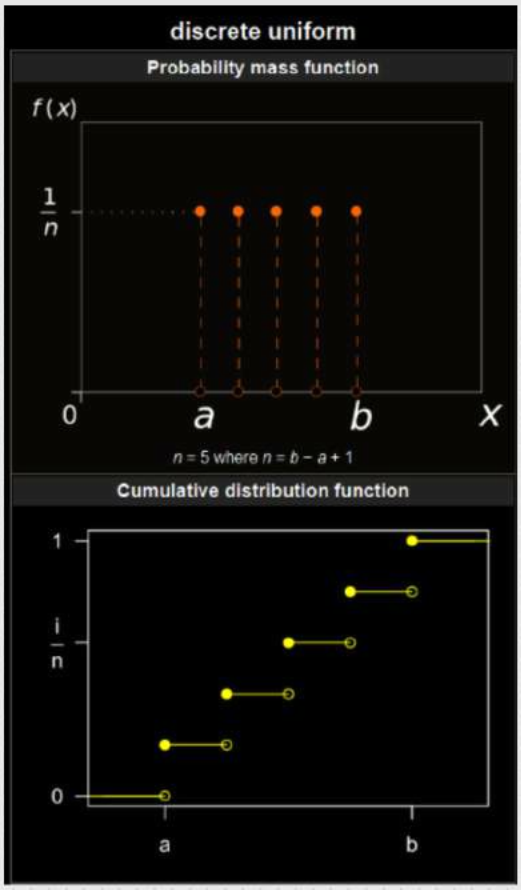
$$= 0.5$$

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

$$= 0.66 = 66\%$$

② Discrete Uniform Distribution (pmf)

In probability theory and statistics, the discrete uniform distribution is a symmetric probability distribution wherein a finite number of values are equally likely to be observed; every one of n values has equal probability 1/n. Another way of saying "discrete uniform distribution" would be "a known, finite number of outcomes equally likely to happen".



Eg: Rolling a dice

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

$a = 1 \qquad b = 6$

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

$Pr(2) = \frac{1}{6}$

$Pr(3) = \frac{1}{6}$

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

Notation $U(a, b)$

Parameters a, b with $b \geq a$

PMF $\frac{1}{n}$

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

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

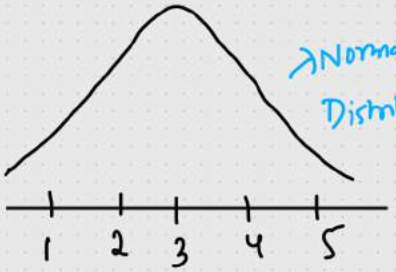
Standard Normal Distribution And Z-Score

Z-stats

Normally Distributed

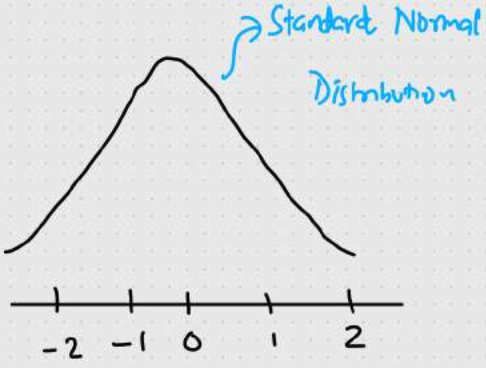
$X = \{1, 2, 3, 4, 5\}$ $\mu = 3$

$\sigma = 1.414 \approx 1$



$\mu = 0$

$\sigma = 1$



$X \sim SND(\mu=0, \sigma=1)$

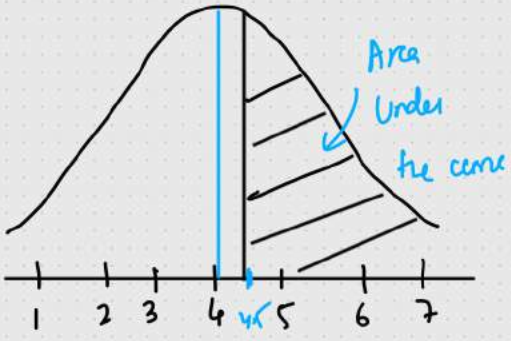
$$Z\text{-score} = \frac{x_i - \mu}{\sigma}$$

$$= \frac{1-3}{1} = -2 \qquad \frac{4-3}{1} = \boxed{1}$$

$$= \frac{2-3}{1} = -1 \qquad \frac{5-3}{1} = 2$$

$$= \frac{3-3}{1} = 0$$

①



$\mu = 4$

$\sigma = 1$

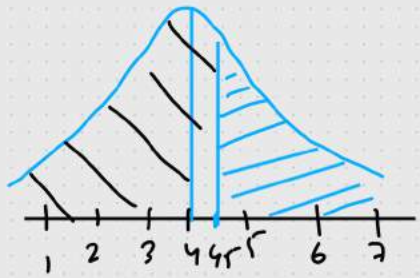
① How many standard deviation 4.5 is away from mean?

$x_i = 4.5$

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

Question : $\mu = 4 \quad \sigma = 1$

What percentage of data is falling above 4.5?



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

1.0000
0.69146
0.30854

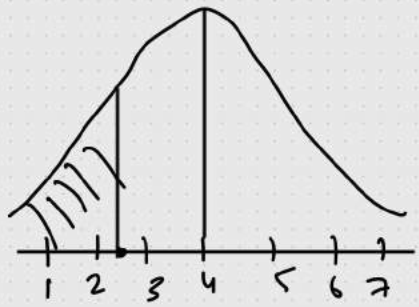
0.69146 ⇒ 69.146

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

$= 0.30854$

$= 30.85\%$

④ What percentage of data is falling below 2.5?



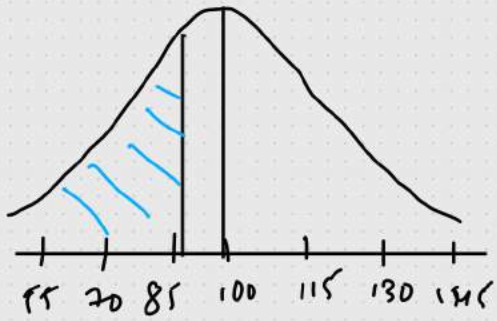
$$Z\_score = \frac{2.5 - 4}{1} = -1.5$$

Area under the curve (≤ 2.5) = 0.00621

= 6.6%

(Prob) 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?

Ans) $\mu = 100 \quad \sigma = 15 \quad x_i = 85$



① Z score = $\frac{x_i - \mu}{\sigma}$

= $\frac{85-100}{15}$

= -1 //.

Area under the curve = 0.15866

= 15.866

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

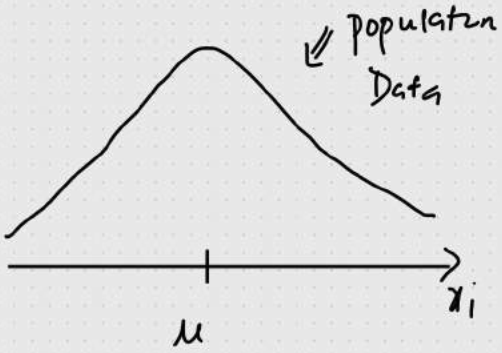
= ≈ 84%

Central limit Theorem (CLT)

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.

$n = 20$

① $X \tilde{N}(\mu, \sigma)$



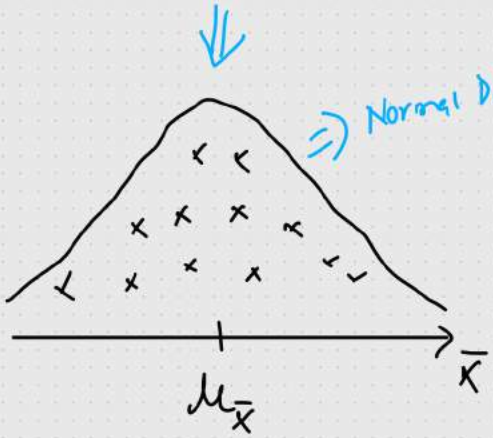
$S_1 = \{x_1, x_2, x_3 \cdots x_{20}\} = \bar{x}_1$

$S_2 = \{x_2, x_3, x_5 \cdots x_{20}\} = \bar{x}_2$

$S_3 = \{\quad\} = \bar{x}_3$

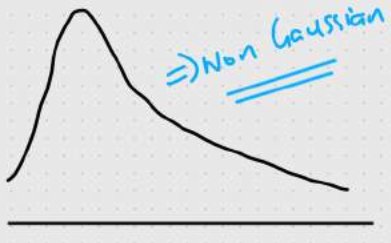
$S_m = \{\ldots\ldots\} \quad \bar{x}_m$

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



→ $n \geq 30$ ←

② $X \not\tilde{N}(\mu, \sigma)$



⇒

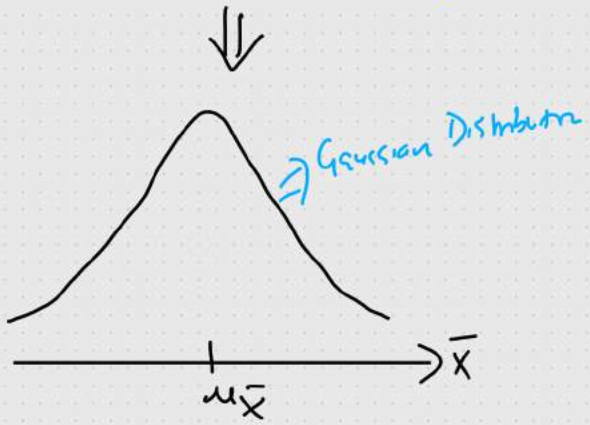
$S_1 = \{x_1, x_2 \cdots x_{30}\} = \bar{x}_1$

$S_2 = \{x_2\ x_3 \cdots x_{30}\} = \bar{x}_2$

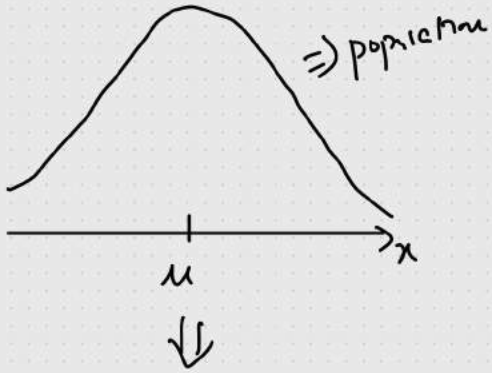
$S_3 = \quad\quad = \bar{x}_3$

S_4

s_m ⋮ $\bar{x}_m$



⑧ Important for Interview



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

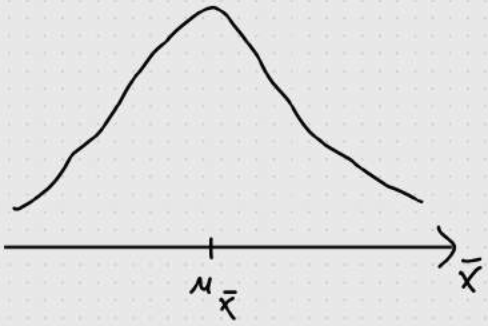
σ = population std

μ = population mean

n = sample size

Sampling distribution of mean

CLT



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

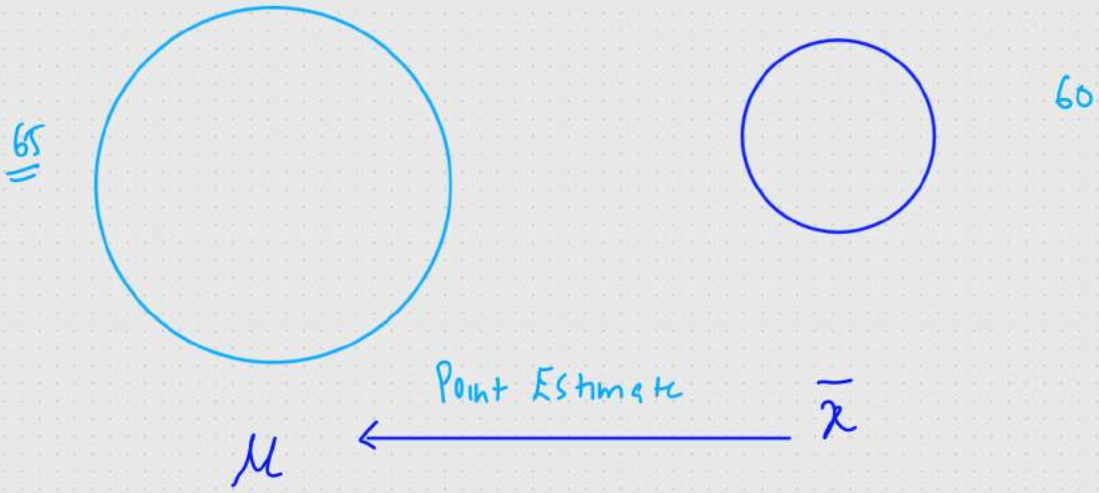
n can be any value

PW SKILLS

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

(1) **Point Estimate** : Single numerical value used to estimate the unknown population parameter

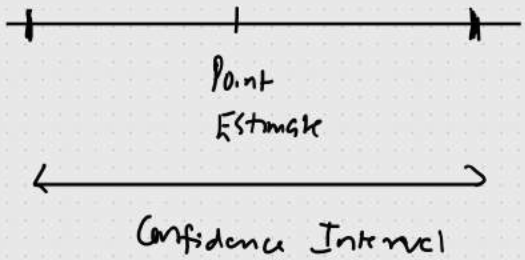
Sample mean is a point estimate of a population mean



(2) **Interval Estimate** : Range of values used to estimate the unknown population parameters

Interval estimates of population parameters are called Confidence Intervals

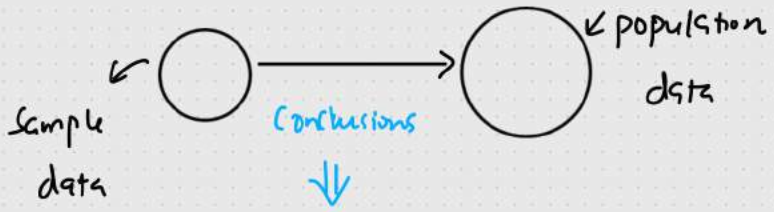
55 – 65



PW SKILLS

A Hypothesis And Hypothesis Testing Mechanism

Inferential Stats : Conclusion or Inferences



{unknown population parameter}.

Hypothesis Testing

Hypothesis Testing Mechanism

Person Crime → Court

① Null Hypothesis (H_0) = The Person is not Guilty
- The assumption you are beginning with

② Alternate Hypothesis (H_1) = The Person is guilty.
- Opposite of Null Hypothesis

{p value}.

③ Experiments → Proof collect (DNA Test, Finger Print) ⇒ Statistical Analysis

④ Accept the Null Hypothesis or Reject the Null Hypothesis

Eg: Colleges at District A states its average passed percentages of Students are 85%. A new college opened in the district and it was found that a sample of student 100 have a pass percentage of 90%. with a standard deviation of 4%.

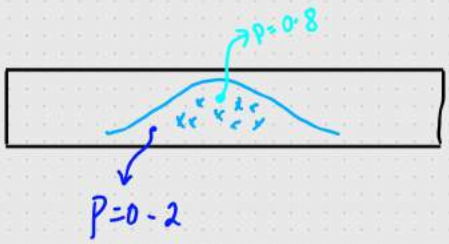
Does this school have a different passed percentage.

Ans) Null Hypothesis $H_0 = \mu = 85\%$

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

Pvalue

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



Out of 100 touches in this Keep the probability of touching in this region 20

Hypothesis Testing

Exp = Coin is fair or Not {100 Tosses}

{H, T}

① Null Hypothesis H_0 : Coin is fair

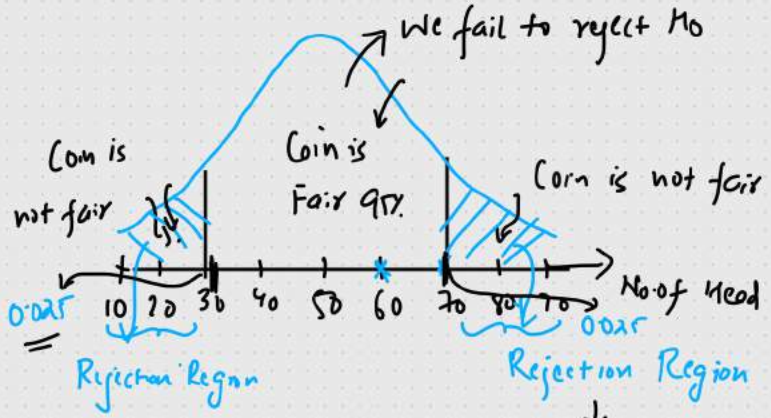
P(H) = 0.5 P(T) = 0.5

② Alternate Hypothesis H_1 – Coin is not fair

P(H) = 0.6 P(T) = 0.4

→ P(H) = 0.7 P(T) = 0.3

③ Experiment



We Reject the Null Hypothesis

We reject the Null Hypothesis

Significance value : $\alpha = 0.05$ C.I = 1 − 0.05 = 0.95

Conclusion

P = 0.01 < Significance

We reject the Null Hypothesis

c) c

We fail to reject Null hypothesis

Hypothesis Testing And Statistical Analysis

① Z Test
② t Test
} ⇒ Average ⇒ Z table / t table → Z score And p value

③ CHI SQUARE ⇒ Categorical Data

④ ANNOVA ⇒ Variance

Z test : i) Population std ii) $n \geq 30$

With a $\sigma = 3.9$

1) The average heights of all residents in a city is 168cm. A doctor believes the mean to be diffuent. 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.

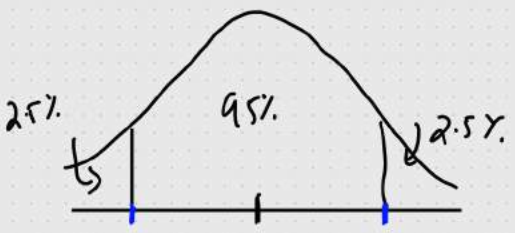
Ans) $\mu = 168cm$, $\sigma = 3.9$, $n = 36$ $\bar{x} = 169.5cm$

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

b) Alternate Hypothesis H_1 $\mu \neq 168cm$ { 2 Tail Test }

c) C.I = 0.95 $\alpha = 1 - 0.95 = 0.05$

Decision Boundary



① Z test
② p value

$1 - 0.25 = 0.9750$

−1.96 168 +1.96

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

d) Statistical Analysis

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

$\sigma/\sqrt{n}$

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

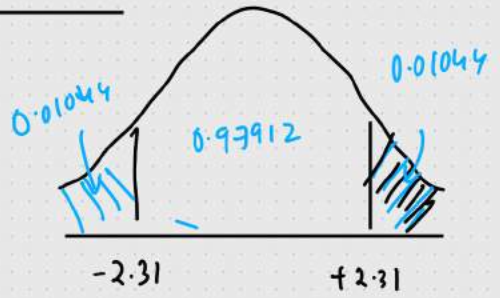
If Z-test value is less than −1.96 or greater +1.96 We Reject the Null Hypothesis

Else

We Accept Null Hypothesis

$2.31 > +1.96$ {We Reject the Null Hypothesis}.

② P-value



1 − Area under the curve of 2.31

1 − 0.98956

=

p value = 0.01044 + 0.01044 = 0.02088//.

if P-value < Significance

0.02088 < 0.05 {Reject the Null Hypothesis}

② A factory manufactures bulbs with a average warranty of 5 years with 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 find 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?

Ans) $\mu = 5 \quad \sigma = 0.50 \quad n = 40 \quad \bar{x} = 4.8$.

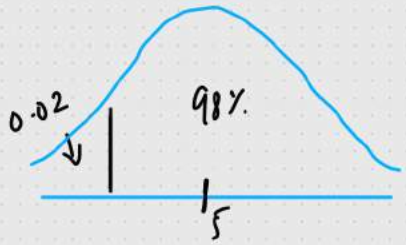
① Null Hypothesis $H_0 : \mu = 5$

② Alternate Hypothesis $H_1 : \mu < 5$ {1 Tail Test}

③ Decision Boundary

C.I = 0.98

$\alpha = 1 - 0.98 = 0.02$.

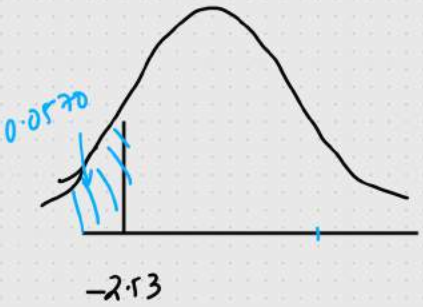


④ P value

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

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

$Z\text{-test} = -2.53$



Area Under the Curve of -2.53

Z value is = 0.00570

P-value : 0.00570.

if p-value < Significance

0.0570 < 0.02 => False

{We accept the Null Hypothesis}

Conclusion : The Warranty needs to be revised.

Student t distribution

Statistical Analysis using Z-score we need population standard deviation (σ)

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

Soln: Student t distribution

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

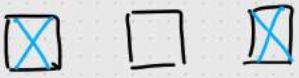
$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

s = sample std

Degree of freedom

$$dof = n - 1$$

n = sample size

3 people 

https://www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf

PW SKILLS

T-stats ÷ J 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 sample of 30 participants who have taken the medication has a mean of 140 with a standard deviation of 20. Did the medication affect intelligence? C.I = 95%

Ans) $\mu = 100$ $n = 30$ $\bar{x} = 140$ $\sigma = 20$ C.I = 0.95

$\alpha = 0.05$

① Null Hypothesis $H_0 : \mu = 100$

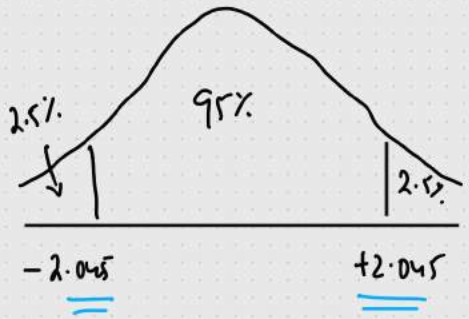
Alternate " $H_1 : \mu \neq 100$ {2 Tail Test}

② $\alpha = 0.05$

③ Degree of freedom

$df = n - 1 = 30 - 1 = 29$

④ Decision Rule



If t test is less than -2.045 and greater than 2.045, Reject the Null Hypothesis

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 Conclusion

Decision Rule : If t is less than $-2.045^{<2}$ and greater than $2.045^{<2}$, reject the Null Hypothesis

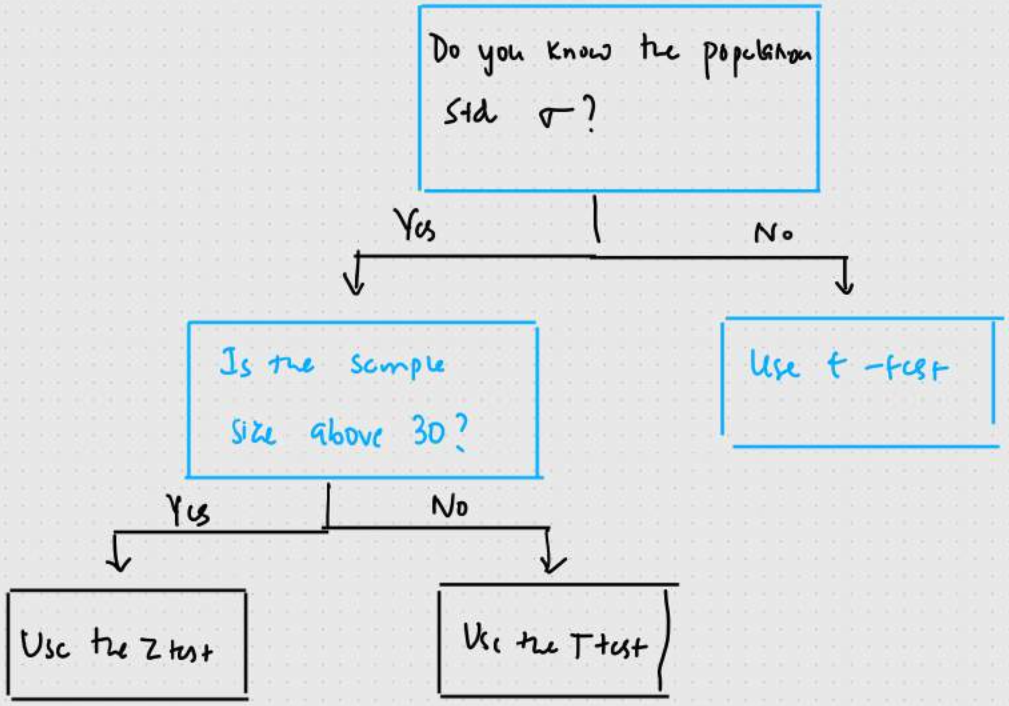
$t = 10.96 > 2.045^{<2} \Rightarrow$ Rejecting the Null Hypothesis

Answer

Conclusion : Medication has increased the Intelligence.

PW SKILLS

When To Use T-tests Vs Z-test



PW SKILLS

Type 1 and Type 2 Errors

Reality : Null Hypothesis is True or Null Hypothesis is False

Decision : Null Hypothesis is True or Null Hypothesis is False

⇓

Conclusion

Outcome 1 : We reject the Null Hypothesis when in reality it is False ⟶ Good

Outcome 2 : We reject the Null Hypothesis when in reality it is True ⟶ Type 1 Error

Outcome 3 : We retain the Null Hypothesis, when in reality it is False ⟶ Type 2 Error

Outcome 4 : We retain the Null Hypothesis when in reality it is True ⟶ Good

PW SKILLS

Bayes Statistics (Bayes Theorem)

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

Baye's Theorem

Probability → Independent Events
→ Dependent Events

① Independent Events

Eg: Rolling a Dice

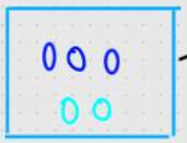
$\{1, 2, 3, 4, 5, 6\}$

$Pr(1) = \frac{1}{6}$ $Pr(2) = \frac{1}{6}$ ...

Tossing a Coin

$Pr(H) = 0.5$ $Pr(T) = 0.5$

② Dependent Events



$Pr(R) = \frac{2}{5}$

$Pr(Y) = \frac{3}{4}$

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

$$= \frac{2}{5} * \frac{3}{4} = \frac{6}{20}$$

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

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

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

Bayes theorem

Conditional Probability

⇓

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

A, B = events

$Pr(A/B)$ = probability of A given B is true

$Pr(B/A)$ = probability of B given A is true

$Pr(A), Pr(B)$ = The independent probability of A and B.

Dataset

Size of house	No. of Room	Location	Price
x_1	x_2	x_3	y

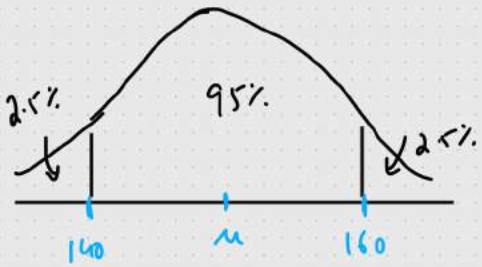
$$Pr(y/x_1, x_2, x_3) = \frac{Pr(y) * Pr(x_1, x_2, x_3/y)}{Pr(x_1, x_2, x_3)}$$

⇓

Bayes theorem

Confidence Intervals and Margin of Error

$\mu = 160$



C.I = 95%

140 ⟷ 160

Point Estimate : A value of any statistics that estimates the value of an unknown population parameter is called Point Estimate

$\bar{x} \longrightarrow \mu$

$\bar{x} = 2.95$ $\quad\quad$ $\mu = 3$

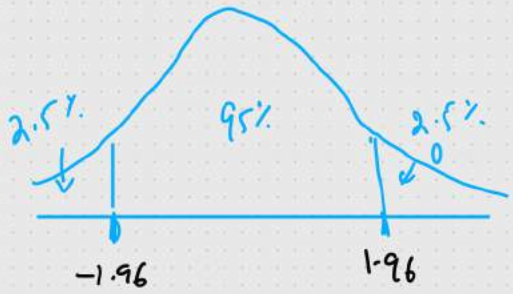
Confidence Interval

We construct a confidence interval to help estimate what the actual value of the unknown population mean is.

Point Estimate $\pm$ Margin of Error

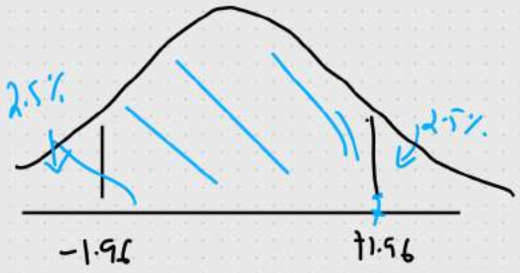
Z test $\quad\quad \bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ $\quad\quad \alpha = 0.05$

$Z_{0.05/2} \Rightarrow Z_{0.025}$



① On the verbal section of CAT Exam, the standard deviation is known to be 100. A sample of 25 test takers has a mean of 520. Construct a 95% C.I about the mean?

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



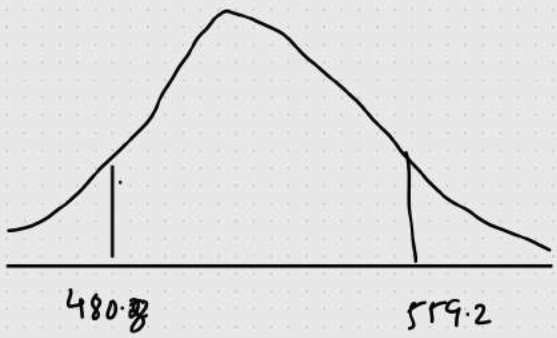
$1 - 0.25 = 0.9750$

$=$ $=$

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

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

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



I am 95% confident that the mean CAT score lies between 480.8 and 559.2.

PW SKILLS

CHI SQUARE TEST

The Chi Square Test for Goodness of fit test claims about population proportions. {Categorical variable}

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

Eg: There is a population of Male who likes different Color of bikes

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

⇓ Theory categorical distribution

↳ Observed categorical distribution.

(K) Goodness of fit.

Goodness of fit test

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

	O	E	$\frac{12}{100} \times 100 = 12$
Right handed	30	12	⇒ Theory categorical distribution
Left handed	70	88	
	100	100	


Sample

PW SKILLS

CHI SQUARE For Goodnces of Fit

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 results

<50	50-75	>75
140	160	200

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

Ans)

2010
Expected

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

2020
n=500
Observed

<50	50-75	>75
140	160	200

Expcted

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

① Null Hypothesis : H_0 : The data meets the expectation

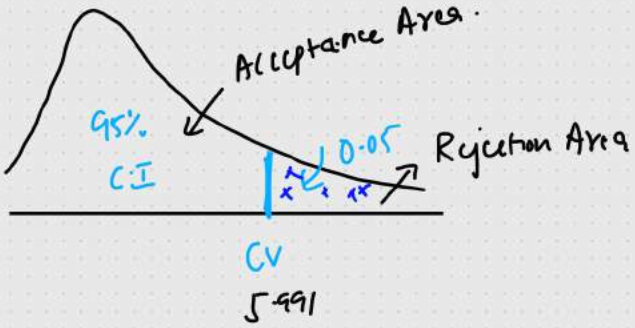
Alternate Hyp : H_1 : The data does not meet the expectation

② $\alpha = 0.05$ C.I = 95%

③ Degree of freedom

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

④ Decision Boundary



If χ^2 is greater than 5.99, Reject H_0

else

We fail to reject the Null Hypothesis

⑤ Calculate Chi Square Test Statistic

2020

n=500

	<50	50-75	>75
Observed	140	160	200

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

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

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

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

$$= 16 + 0.66 + 10$$

$$= 26.66$$

$$\chi^2 = 26.66$$

If χ^2 is greater than 5.99, Reject Ho

else

We fail to reject the Null Hypothesis

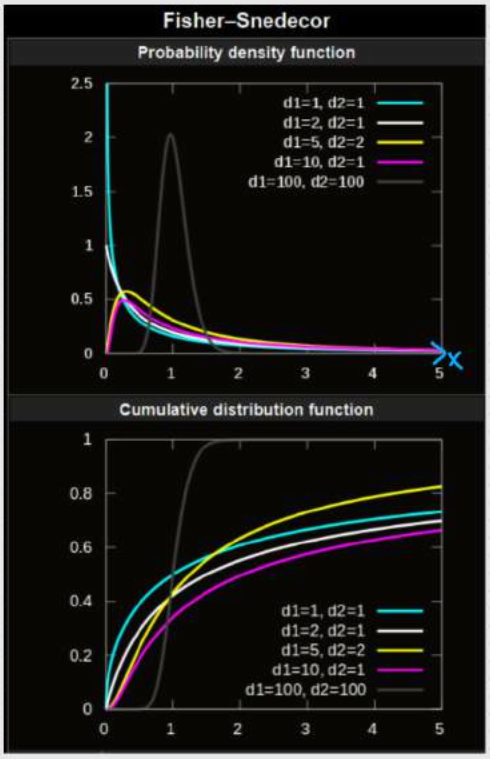
$26.66 > 5.99$, Reject Ho

Answer

The weights of 2020 population are different than those expected in the 2010 population

F distribution

In probability theory and statistics, the F-distribution or F-ratio, also known as Snedecor's F distribution or the Fisher–Snedecor distribution (after Ronald Fisher and George W. Snedecor) is a continuous probability distribution that arises frequently as the null distribution of a test statistic, most notably in the analysis of variance (ANOVA) and other F-tests.



Parameters : $d_1, d_2 > 0$ degree of freedom

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

F table

$\alpha = 0.05$ d_1 d_2

pdf =

$$f(x; d_1, d_2) = \frac{\sqrt{\frac{(d_1 x)^{d_1}\, d_2^{d_2}}{(d_1 x + d_2)^{d_1 + d_2}}}}{x\,\mathrm{B}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)}$$

Beta function

$$\mathrm{B}(m, n) = \frac{(m-1)!\,(n-1)!}{(m+n-1)!} = \frac{m+n}{mn} \Big/ \binom{m+n}{m}.$$

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

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

S_1 = Independent random variables { Chi square distribution }.

S_2 = " " "

(S_1) d_1 = Degree of freedom

(S_2) d_2 = " " "

F Test : Variance Ratio Test

F-Test [Variance Ratio Test]

① The following data shows the no. of bulbs produced daily for some 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

$\alpha = 0.05$

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

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

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

② Calculation of Variance

A

X_1	$\bar{x}$	$(x_1 - \bar{x})^2$
40	37	9
30	37	49
38	37	1
41	37	16
38	37	1

B

X_2	$\bar{x_2}$	$(x_2 - \bar{x})^2$
39	37	4
38	37	1
41	37	36
33	37	16
32	37	25
39	37	4
40	37	9

35	37	4
$\bar{X}_1 = 37$		$\sum (x-\bar{x})^2 = 80$

34	37	9
$\bar{X}_2 = 37$		$\sum (x_2-\bar{x})^2 = 84$.

$$S_1^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n_1 - 1}$$

$$S_1^2 = \frac{80}{6-1} = \frac{80}{5} = 13$$

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

$$= \frac{84}{7} = 12$$

→ Calculation of Variance Ratio F-test

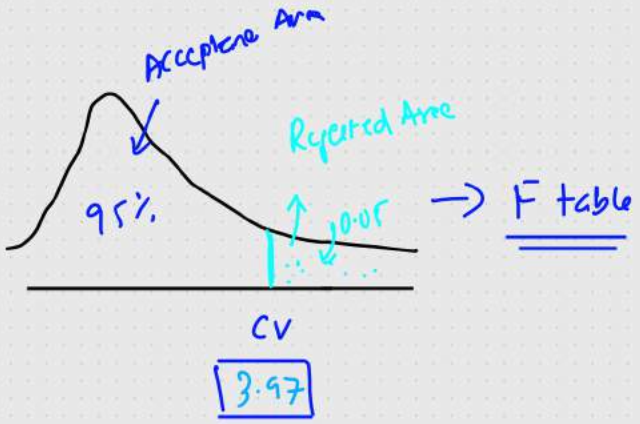
$$F = \frac{S_1^2}{S_2^2} = \frac{13}{12} = 1.33$$

③ Decision Rule

$df1 = 6-1 = 5$

$df2 = 8-1 = 7$

$\alpha = 0.05$



→ F table

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

1.33 < 3.97, we fail to Reject the Null Hypothesis

Conclusion

Worker B is not efficient when worked to worker A.

Analysis Of Variance (ANOVA)

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

Anova

1. Factors (variables)
2. Levels

Eg : Factor = Medicines

Levels = 5mg, 10mg, 20mg [Dosage]

Mode of Payment → Factor

GPAY, PhonePE, IMPS, NEFT [Levels]

PW SKILLS

Analysis Of Variance (ANOVA)

Assumptions in ANOVA

① Normality of Sampling Distribution of Means

The distribution of sample mean is normaly distributed.

② Absence of Outliers

Outlying score need to be removed from dataset.

③ Homogenity Of Variance

Each one of the population has same variance

$$[\sigma_1^2 = \sigma_2^2 = \sigma_3^2]$$

Population variance in different levels of each independent variable are equal

④ Samples are independent and random.

PW SKILLS

Analysis Of Variance (ANOVA)

Types Of ANOVA

① One Way ANOVA : One factor with atleast 2 levels, these levels are independent

Eg: Doctor want to test a new medication to decrease headache. They split the participant in 3 conditions [10mg, 20mg, 30mg]. Doctor ask the participants to rate the headache [1-10]

Medication → Factor

10mg	20mg	30mg
5	7	2
9	8	7
_	_	_
_	_	_

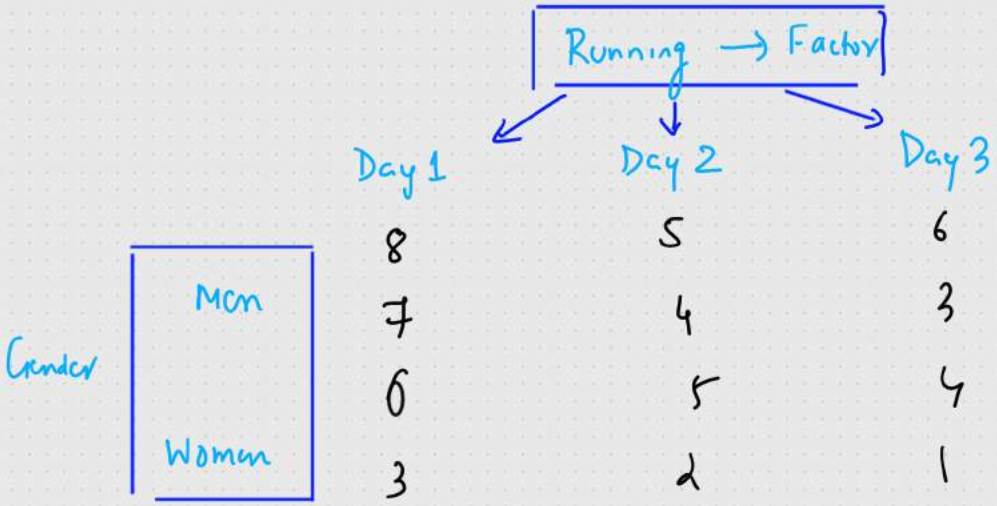
② Repeated Measures Anova : One factor with atleast 2 levels, levels are dependents

Running → Factor

Day 1	Day 2	Day 3
8	5	6
7	4	3

9 8 7

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



PW SKILLS

Analysis Of Variance (ANOVA)

Hypothesis Testing In ANOVA

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

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

$\times$ $\mu_1 \neq \mu_2 \neq \mu_3 \cdots \neq \mu_K$

Test Statistics

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

F test

Variance between sample

σ^2

	X_1	X_2	X_3
Variation within Sample	1	6	5
	2	7	6
	4	3	3
	5	2	2
	3	1	4
	$\sum X_1 = 15$	$\sum X_2 = 19$	$\sum X_3 = 20$
	$\bar{X}_1 = 3$	$\bar{X}_2 = 19/5$	$\bar{X}_3 = 4$

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

H_1 : Atleast one sample mean is not equal

One Way ANOVA

One factor with atleast 2 levels, levels are independent

① Doctors want to test a new medication which reduces headache. They split the participant into 3 condition [15mg, 30mg, 45mg]. Later on the doctor ask the patient to rate the headache between [1-10]. Are there any differences between the 3 conditions using alpha = 0.05?

Ans)

15 mg	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 Null and Alternate hypothesis?

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

H_1 : not all μ's are equal

② State Significance value

$\alpha = 0.05 \Rightarrow C.I = 0.95$

③ Calculate Degree of freedom

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

$df_{between} = a - 1 = 3 - 1 = 2$

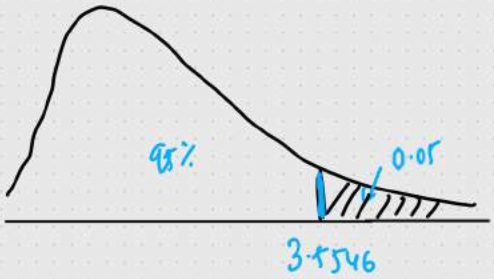
$df_{within} = N - a = 21 - 3 = 18$

$df_{total} = N - 1 = 21 - 1 = 20$

$(2, 18)$

⇓

F table



Critical value = 3.5546.

④ State Decision Rule

If F is greater than 3.5546, reject the null hypothesis.

⑤ Calculate Test Statistics

	SS	df	MS	F
Between				
Within				
Total				

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

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

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

15mg : $9+8+7+8+8+9+8 = 57$

30mg : $7+6+6+7+8+7+6 = 47$

45mg : $4+3+2+3+4+3+2 = 21$

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

$= \boxed{98.67}$

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

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

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

$= 853.$

$= 853 - \left[\frac{57^2+47^2+21^2}{7}\right]$

$= \boxed{10.29}$

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

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

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

$$F = \frac{\text{Variation between samples}}{\text{Variation within samples.}}$$

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

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

If F is greater than 3.5546, reject the Null hypothesis.

$86.56 > 3.5546$, Reject the Null hypothesis