INTRODUCTION TO STATISTICS

Defn :- Stats is the Science of Collecting, Organizing and analysing data.

Data :- facts or pieces of information

Eg :- Heights of students in classroom
Salary of people in society.

Types of Stats

① Descriptive Stats
② Inferential Stats.

Defn

It consists of Organizing, Summarizing, and visualizing data.

① Measure of Central tendency.
② Measure of dispersion (V, Sd, Z score)
② Diff types of distribution

Eg :- Histogram, pdf, pmf, Gaussian, log normal, Exponential, Binomial Bernoulli, Poisson

② Inferential Stats

Defn :- It consists of using data you have measured to form Conclusions.

① Z-test
② t-test
③ Chi square test
④ Anova test

Data

- Quantitative
- Qualitative

Sampling Techniques

① Random Sampling.

→ Most common in behavioral research
→ Equal chance of being selected in the sample.

② Stratified Sampling.

→ Dividing the population into subgroups or strata based on certain characteristics or attributes.
→ w.r.t age, gender, income etc.

③ Convenience Sampling.

→ Participants are selected based on availability and willingness to take part.

④ Cluster Sampling.

→ Divide a population into clusters.
→ eg: districts, schools and randomly selecting the clusters

⑤ Snowball Sampling.

→ Existing study subjects are used to recruit more subjects into the sample

⑥ Purposive Sampling.

→ Selecting samples based on the judgement of the survey taker or researcher.

⑦ Systematic Sampling

→ It is a probability sampling method where researchers select members at a regular interval.

Scale of Measurement

1) Nominal Scale data ⟶ Qualitative / categorical data.
   Ex Gender, Color, Labels.
   Order or Rank doesn't matter.
2) Ordinal Scale data
3) Interval Scale data
4) Ratio Scale data.

Ordinal Scale data

Rank is imp
Order Matters.
Difference cannot be measured unless we take more information.

Ex: Race, Score
Top colleges.

Interval Scale data

The Order will matter
Difference can measurable (Ratio)
Contains a '0' starting point.

Eg: Student marks

The Rank / order matters
Difference can be measured.
The Ratio cannot be measured.
No 0 Starting point.

Eg: Temperature Variables.
Intelligence Quotient.

Random Variables

Variable (x, y)

$x + 5 = 7$.

$x = 2$

$x = 2$

$y = 6$.

$S = y + x$.

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

ex :- Toss a coin.

$$X = \begin{cases} 0 & \text{if heads} \\ 1 & \text{if tails.} \end{cases}$$

Rolling a dice.

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

defn: A Random variable is a variable in Statistics that assigns numerical values to the outcomes of Sample Space. The possible values of a random variable depend on the outcomes of a random phenomenon.

Covariance & Correlation

It is the measure of the relationship b/w two random variables. It measures how much the variables change together. or the variance b/w them.

→ It measures the direction of a relationship b/w two variables.

$$Cov(x,y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N}$$

$\bar{y}, \bar{x}$ = mean

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

Advantages

→ +ve or -ve value

→ Shows the relationship b/w two variables positive / negative.

Disadvantages

Doesn't have a specific limit value.

② Pearson Correlation Co-efficient

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

PCC doesn't work well with non-linear data. So we will use Spearman Correlation.

③ Spearman's rank Correlation Coefficient

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

R = Rank. based on frequency / order / value

→ If 2 feature are highly correlated, it's ok to drop one of the feature.

Q

Probability Distribution function.

Probability density function

Continious value.

ex: Age, Height.

float

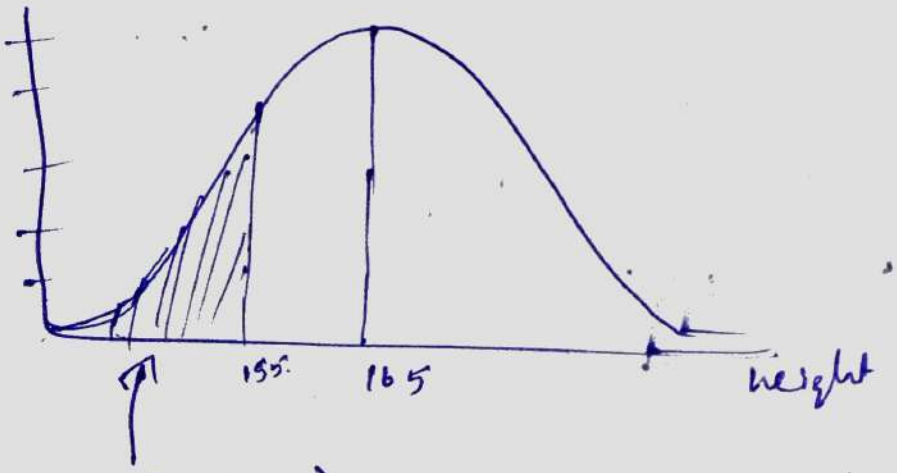
Probability mass function.

Discrete value.

ex: no of bank acc.

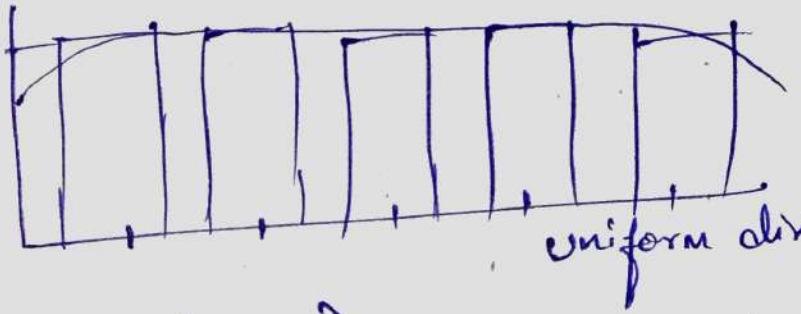
int

Rolling a dice.



$Pr(H <= 155)$

$Pr(H <= 155, H >= 175)$

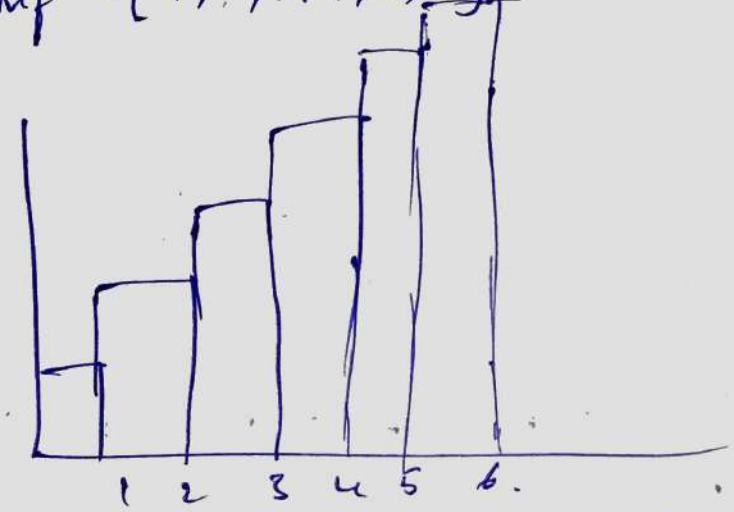


$Pr(X \leq 4) = Pr(x=1) + Pr(x=2) + P(x=3) + Pr(x=4)$

$= 1/6 + 1/6 + 1/6 + 1/6 = 4/6 = 2/3$

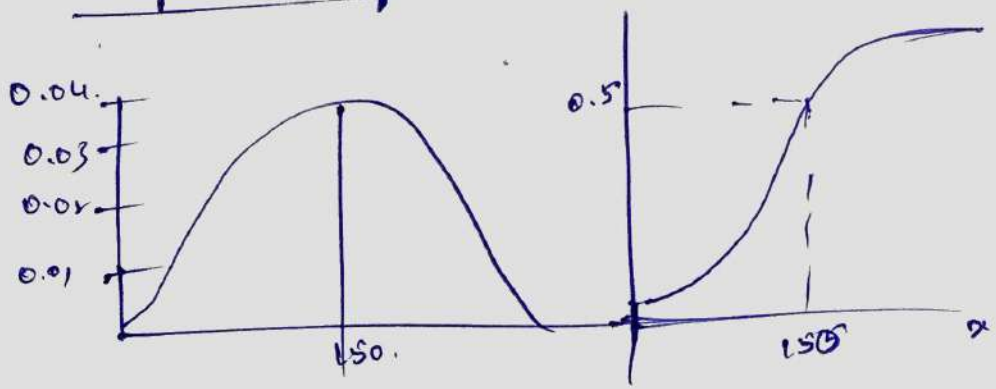
3) Cumulitive Distribution Function

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



→ Adding the previous value to the next value

4) pdf & cdf.



Probability Density of pdf => Gradient or derivative of cdf.

pdf is the derivative of cdf.

cdf is the integration of pdf.

Types of Probability function

1) Normal / Gaussian dist (pdf)
2) Bernoulli dist (pmf) (binary outcome)
3) Uniform dist. (pmf)
4) Poisson dist. (pmf)
5) Binomial dist (pmf)
6) Log Normal dist (pdf)

① Bernoulli Distribution

1) discrete Random Variable (pmf)
2) Outcomes are Binary

Ex:- Toming a coin {H, T}.

$Pr(H) = 0.5 = p$

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

whether a person will pass/fail.

$Pr(pass) = 0.7 = p$

$Pr(fail) = 1 - p = 0.3 = q$

② Binomial Distribution

* Discrete Random variable
→ Every Experiment outcom is Binary.
* Thus experint is performed for n trials

Eg :- Tossing a coin 10 times.

notation → $B(n, p)$

$Pr(H) = 0.5 = p$

$Pr(T) = 0.5 = q$

parameter : $n \in \{0, 1, 2, 3, \dots\}$ => no of trials.

$p \in \{0, 1\}$ → Success probability.

$q = 1 - p$.

Support = $k \in \{0, 1, 2, \dots n\}$ => no of success.

PMF :

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

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

Notes :- Refer wikipedia.

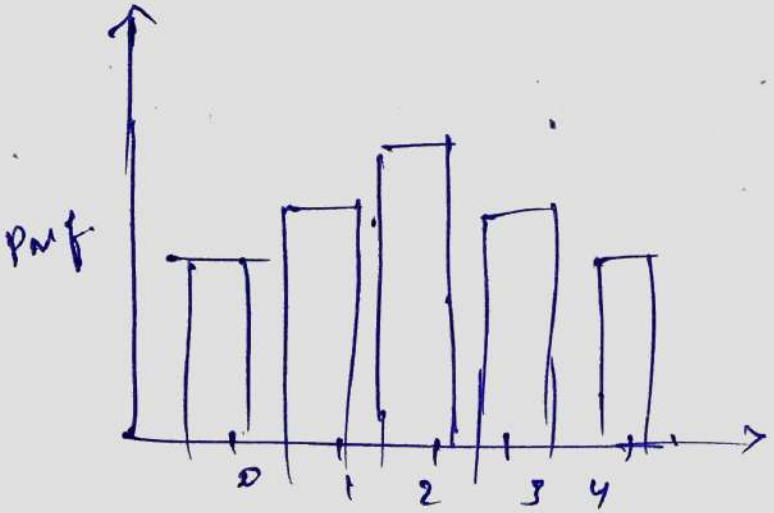
Poisson Distribution

→ Discrete Random Variable (pmf)

→ Describes the no of events occuring in a fixed time interval.

Eg:- no of people visiting hospital every hour.

no of people visiting banks. at 11 am.



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

what is the propability of no of people visiting at 3 pm. => Pr(x=3)

$$= \frac{e^{-3} 3^3}{5!} : 0.401 < 10\%$$

10% of the people visit the bank at 3 pm

Empherical Rule of Normal dist

68 — 1st sd

95 — 2nd sd

99.7 % — 3rd sd.

iv)

UNIFORM DISTRIBUTION

i) Continious Uniform Distribution.

* If a Continious uniform dist is a probability dist that takes value within a specified range. It is defined by two parameters. a, b when a is the lower limit and b is the upper limit.

Notation = $U(a,b)$

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

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

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

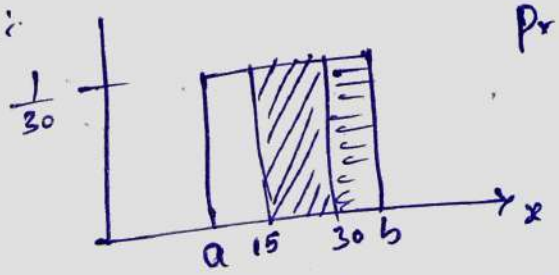
mean $= \frac{a+b}{2}$ median $= \frac{a+b}{2}$

Example: The no of candies sold at a shop is uniformly distributed with a max of 40 and a min of 10.

1→ Probability of daily sale to fall b/w 15 and 30. $a=10$, $b=40$

Ans:

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



$(x_2 - x_1) * \frac{1}{b-a}$

$= 30 - 15 * \frac{1}{40-10}$

$= 15 \times \frac{1}{30} = 0.5$ //

2→ $Pr(x \ge 30)$

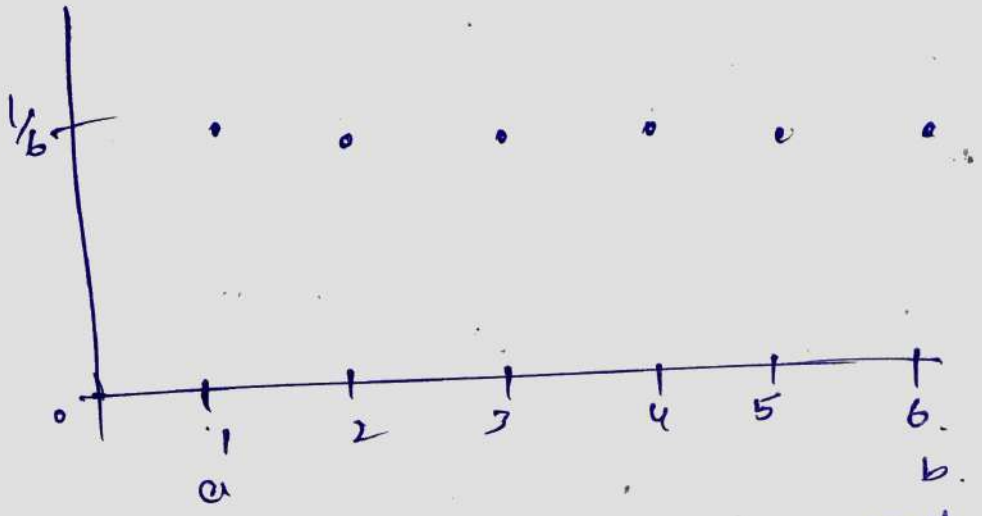
$(x_2 - x_1) \frac{1}{b-a}$

$40 - 30 \times \frac{1}{30}$

$\frac{10}{30} = 0.33$

② Discrete Uniform Distribution (pmf)

Ex → Rolling a Dice



Notation: $U(a,b)$

$n = 6$ (no of outcomes)

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

$Pr(\cdot) = 1/n = 1/6.$

parameter a, b $\boxed{b > a}$

pmf $= 1/n$

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

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

Definition: A discrete uniform distribution is a statistical distribution where the probability of outcomes is equally likely and with finite values. In a discrete uniform dist, every one of a n value has equal probability $1/n$.

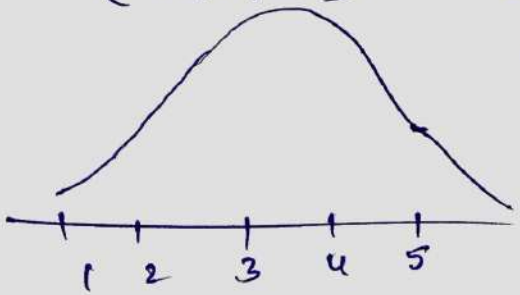
Ex:- Rolling a die
Selecting a card from deck of cards
Taming a fir coin.

STANDARD NORMAL DISTRIBUTION

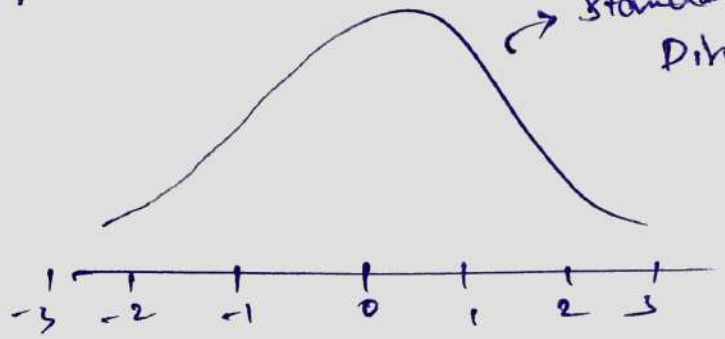
Z-SCORE

Z-Stats → Normally Distributed.

$X = \{1, 2, 3, 4, 5\}$ $\mu = 3$ $\sigma = 1.414$

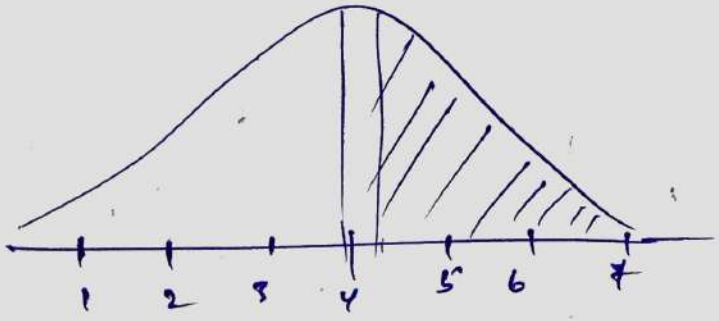


if $\mu = 0$ & $\sigma = 1$ → Standard Normal Distribution.



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

Z table - we use to find out the area under the curve



what % of the data is falling down 4.5

$$Z\text{Score} = \frac{x_i - \mu}{\sigma} = \frac{4.5 - 4}{1} = 0.5$$

$\Rightarrow$ 0.6915

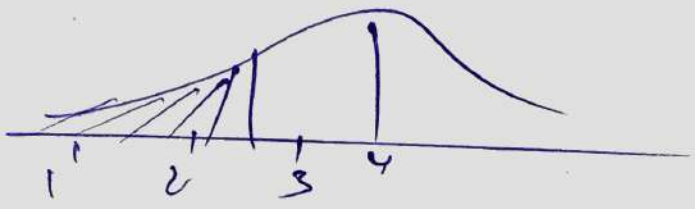
Area under the curve $= 1 - 0.6915 = 0.3085 = 30.85\%$

② percent of data falling below 2.5 ?

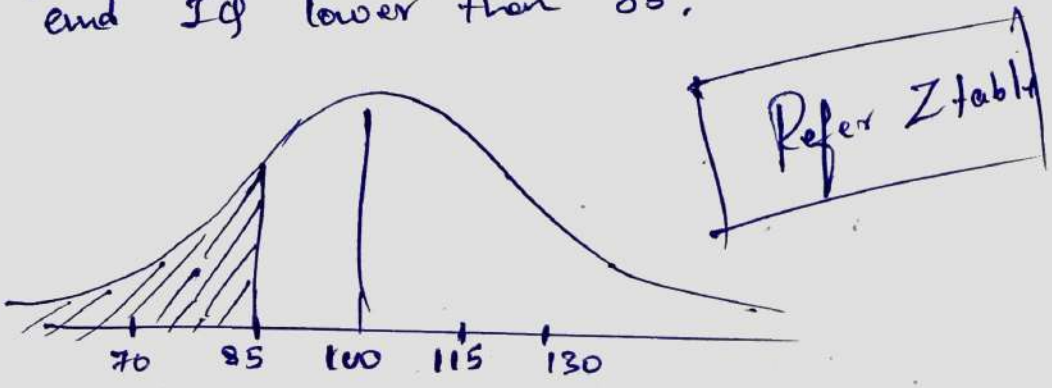
$$Z\text{Score} = \frac{2.5 - 4}{1} = -1.5$$

$= 0.0668$

$= \boxed{6.68\%}$



In india the avg IQ is 100, with σ of 15. what is the percentage of the population would you expect to have and IQ lower than 85.



Refer Z table

$$Z_{score} = -\frac{85-100}{15} = \frac{-15}{15} = -1 = 0.1587$$

$= 15.87\%$

~~$= 0.1602$~~

~~$\neq 1602\%$~~

Area under the curve > 85.

$1 - 0.1587$

$= 0.8413 = 84.13\%$

Area between 85 & 100

$0.5 - 0.1587$

$= 0.3413 = 34.13$

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

(13) # CENTRAL LIMIT THEOREM.

After taking multiple samples from the population

Sample distr of mean $\bar{X} \simeq N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

n can be any value.

Mean will be approx to μ

Standard dev will be $\boxed{\frac{\sigma}{\sqrt{n}}}$ Standard error

* QQ plots → to check whether the dist is normal or not -

Defn:- CLT is a statistical theory that states that when a large sample size has a finite variance, the samples will be normally distributed.

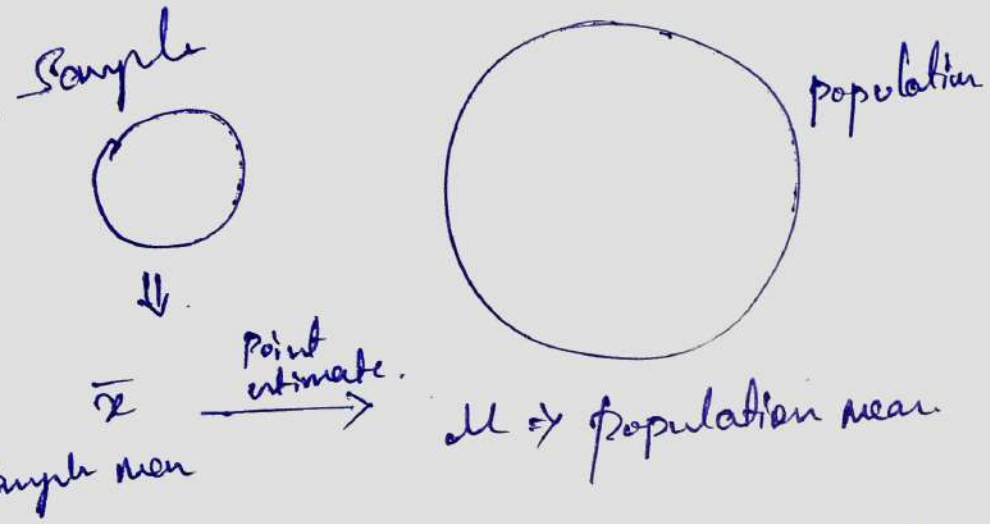
→ The theorem also states that the mean of the samples will be approximately equal to the mean of the whole population.

→ CLT assumes that all samples are identical in size, and regardless of the population's actual distribution

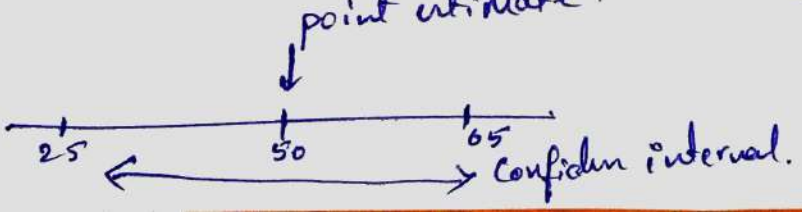
→ The theorem holds true regardless of whether the source population is normal or skewed, provided the sample size is sufficiently larger than n=30

INFERENTIAL STATISTICS

1) **Point Estimate** :- It is an observed numerical value used to estimate an unknown population parameter.
(Single numerical value)

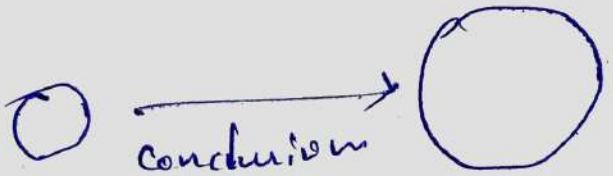


2) **Interval Estimate** → Range of value used to estimate the unknown population parameter.



Hypothesis & Hypothesis Testing Mechanism

Infrential Stats → Conclusions



↳ Hypothesis testing.

Hypothesis testing mechanism

Person Crime →.

① **Null hypothesis** (H_0)
⇒ The person is not guilty.
→ Assumptions you are beginning with.

② **Alternate Hypothesis** (H_1) The person is guilty.
→ ~~Hy~~ opp of H_0.
→ **Evidence** · Experiments
Statistical Analysis. ()
Z-test, t test, Annova test etc.

(14)

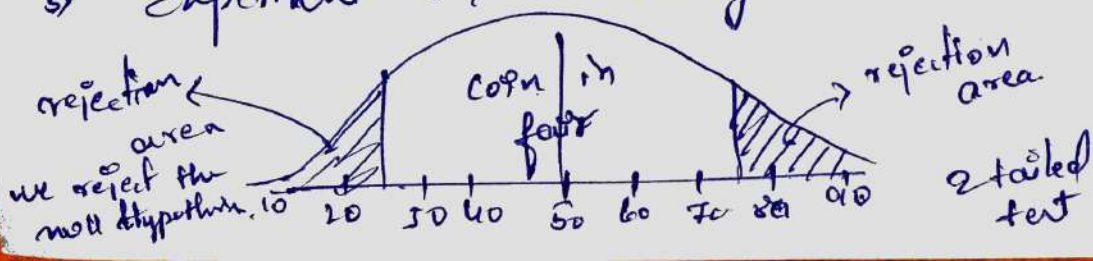
P value

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

Example: Coin is fair or not
{H, T}.

Hypothesis Testing

1) Null hypothesis H_0 → Coin is fair
2) Alternate hypothesis H_1 → Coin is not fair.
3) Experiment → Testing

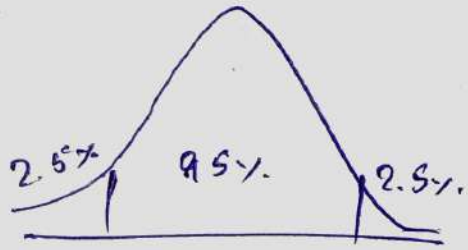


Confidence interval = 95%.

Significance value = α 1 - CI
= 1 - 0.95 = 0.05.

if p < Significance value.
we reject the null hypothesis
else:
we fail to reject the null hypothesis

Confidence Interval & Margin of Error



CI = 95
α = 0.05

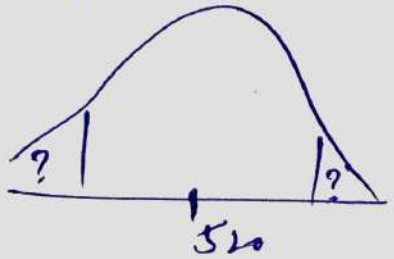
We construct a confidence to help estimate what is the actual value of unknown population mean

↓

Point estimate ± Margin of error.

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

i) In the verbal section of cat exam, the s.d in known to be 100, A sample of 25 test taken has a mean of 520 Construct a 95% of CI. about the mean.



$\sigma = 100$

$n = 25$

$\bar{x} = 520$

CI = 95%.

$\alpha = 0.05$.

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

$Z_{0.025} = -1.96$

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

$= 480.8$

Higher CI $= 520 + 1.96 \times \frac{100}{\sqrt{25}} = 559.2$

I am 95% Confident that the mean cat score lies b/w 480.8 and 559.2.

⑤ Hypothesis testing & Statistical Analysis.

① Z test } avg value.
② t test }
③ Chi square. — categorial
④ Anova. → variance

① Z test

The avg. heights of all residents in a city is 168 cm with a $\sigma = 3.9$. A doctor believes the mean to be different. He measured the height of 36 individuals and found the avg. height to be 169.5 cm.

① State null & alternate hypothesis
② At a 95% Confidence level, is there enough evidence to reject null hypothesis

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

whenever population sd is given, then we should definitely use z test.

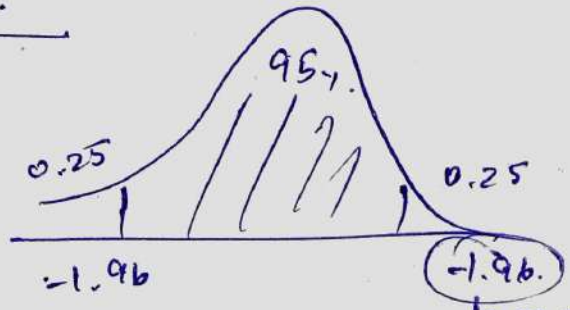
(5) a) Null hypothesis - H_0 - $\mu = 168$ cm.

b) Alternate hypothesis - H_1 $\mu \neq 168$ cm

CI = 0.95

$\alpha = 0.05$.

Decision boundary.



Statistical Analysis

$$Z_{stat} = \frac{\bar{x} - \mu}{\boxed{\sigma/\sqrt{n}}} \Rightarrow \text{Standard error}$$

$$Z_{score} = \frac{x_i - \mu}{\sigma}$$

$n = 1$

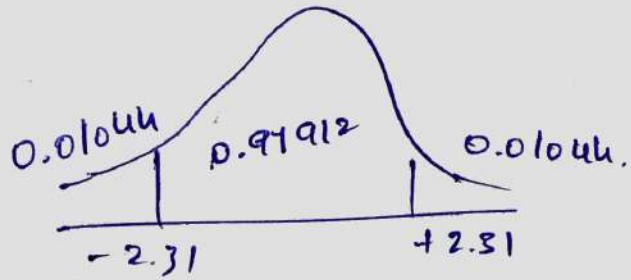
$$= \frac{169.5 - 168}{3.9/\sqrt{36}} = \boxed{2.31}$$

Conclusion => If Z.test value is less than -1.96 or greater than 1.96, we reject the null hypothesis.

$2.31 > 1.96 \Rightarrow$ Reject the null hypothesis

The doctor is absolutely right.

& P.value



P. value = 0.01044 + 0.01044 = 0.02088.

if p-value < significan value

0.02088 < 0.05 => Reject the null hypothesis

(6) A factory manufactures bulbs with an avg warranty of 5 years with standard deviation of 0.50. A worker believes that the bulb will manufacture in less than 5 years. He tests a sample of 40 bulbs and find the avg time to be 4.8 years.

a) State null & Alternate hypothesis

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

~~$\mu = 168$~~
~~$\sigma = 3.9$~~
~~$n = 36$~~
~~$\bar{x} = 167$~~

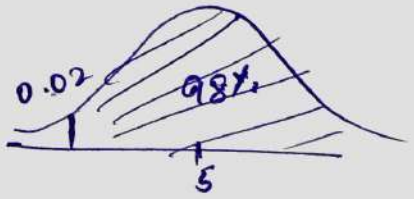
$\mu = 5$
$\bar{x} = 4.8$
$n = 40$
$\sigma = 0.50$

i) Null hypothesis $H_0 = 5 = \mu$.

ii) Alternate hypothesis $H_1 < 5$.

iii) Decision boundary C.I. = 0.98

$\alpha = 1 - 0.98 = 0.02$

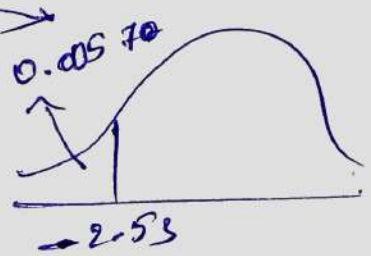


iv) Z. test

$$= \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{4.8 - 5}{0.50/\sqrt{40}} = -2.53.$$

Z. test < -2.05 ⇒ Reject the null hypothesis.

v) P. - value



p value < significan value.
we reject the H_0

T test

① In the population, the avg IQ is 100. A team of researches want to test a new medication to see if it has either a +ve or -ve 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?

CI = 95%.

⇒ If population s.d is not given, use t. test.

$\mu = 100$, $n = 30$ $\bar{x} = 140$ $S = 20$, CI = 0.95
$\alpha = 0.05$.

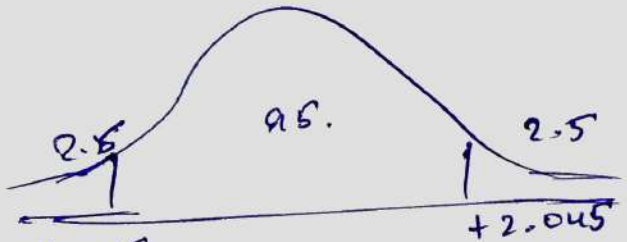
① Null hypothesis - $H_0 : \mu = 100$.
$H_1 : \mu \neq 100$. (2 tailed)

② $\alpha = 0.05$ CI = 0.95.

③ Degree of freedom (dof) = $n - 1 = 30 - 1 = 29$

(16)

④ Decision Rule.



$\alpha = 0.05$

$= \pm 2.045$

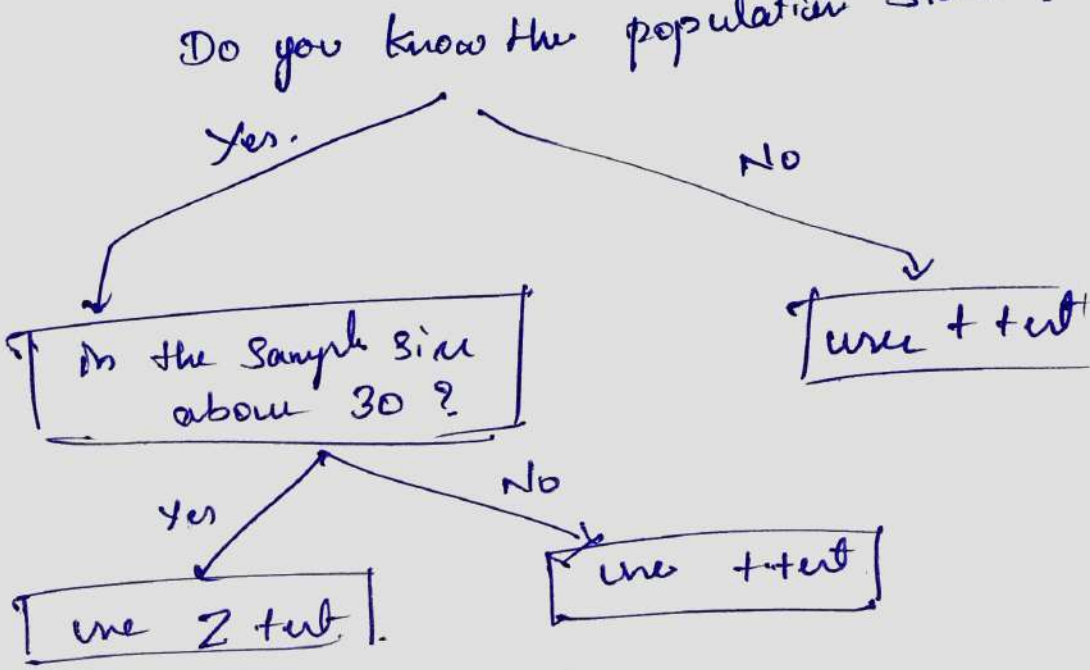
Conclusion: the t-test is less than -2.045 and greater than 2.045. Reject the null hypothesis.

⑤ T-test Statistics.

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

10.96 > 2.045 so we reject the null hypothesis.

When to use T-tests & Z tests



CHI SQUARED TEST

The chi squared test for Goodness of fit claim about population proportion (categorical test).

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

① In a student class of 100 students, 30 are right handed. Does this class fit the theory 12% of people are right handed.

	observed	E	
Right handed = 30		12	theory categorical distribution
left handed = 70		88	

Chi square for goodness of fit

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

<50 kg	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 differences of weights has changed in the last 10 years.

Ans

Expected value

<50 kg	50-75	>75
100	150	250

$H_0 \rightarrow$ The data meets the expectation.

$H_1 \rightarrow$ The data does not meet the expectation

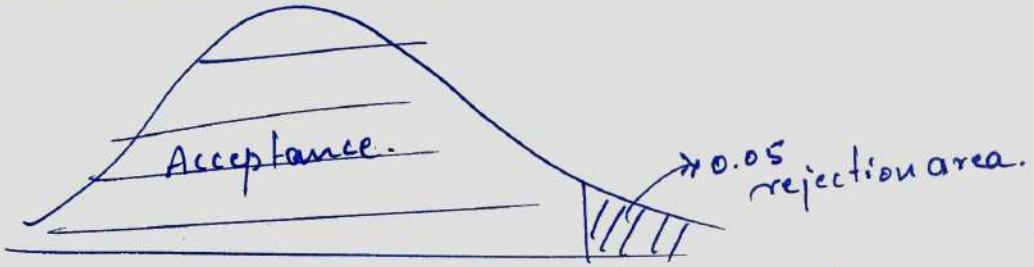
(7)

$\alpha = 0.05$ CI = 95%.

degree of freedom.

$df = k - 1 = 3 - 1 = 2$

Decision boundary → chi squared test



distribution is always right skewed

Critical value = 5.991

If x^2 is > 5.991, we reject the H_0

Calculate chi square test statistics

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

Table: Chi-Square Probabilities

The areas given across the top are the areas to the right of the critical value. To look up an area on the left, subtract it from one, and then look it up (ie: 0.05 on the left is 0.95 on the right)

df	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997

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

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

$$= 16 + 0.66 + 10$$

$$x^2 = 26.66.$$

26.66 > 5.991, Reject H_0.

(18) # Analysis of Variance (ANOVA)

→ It is a statistical method used to compare the means of 2 or more groups.

① Factors (variables)
② Levels

Eg Medicine. → factors

Levels → 5 mg 10 mg 20 mg [dosage]

Mode of payment — factors

Gpay phonepe Imps NEFT [levels]

Types of ANNOVA

① One Way Annova — one factor with atleast 2 levels, then levels are independent.

Eg:- Doctor want to test a new Medication to decrease headache. they split the participants into 3 Condition wrt to dosages [10, 20, 30].

→ Doctors ask the patients to rate the headache (1-10)

Medication → Factor

10 mg	20 mg	30 mg → levels.
5	7	2
9	8	7
—	—	—
—	—	—

② Repeated Measures Anova

one factors with atleast 2 levels, levels are dependent.

Running → factor.

Day 1	Day 2	Day 3
8 kms	5 kms	6 kms

③ Factorial Anova.

Two or more factors (each of which with atleast 2 levels), levels can be either independent or dependent

Running → Factor

	Day 1	Day 2	Day 3
Gender	8	5	6
Male	7	4	3
Female	6	5	4
	3	2	1

dependent

Hypothesis Testing in Anova.

Null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_k$.

Alternate hypothesis H_1 : Atleast one of the mean is not equal.

F test Statistics

$$F = \frac{\text{Variation b/w samples}}{\text{Variation within Samples.}}$$

① Doctors want to test a new medication which reduces headache. They split the participants into 3 condition [15, 30 45] later on the doctor ask the patient to rate the headache between [1 – 10] are there any difference between the 3 conditions using alpha = 0.05 ?

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

(19) ① Define null hypothesis

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

② Alternate hypothesis:

Atleast one mean is not equal.

③ State Significance Value

$\alpha = 0.05 \Rightarrow CI = 95\%$

④ calculate degree of freedom

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

df between $= a - 1 = 2$

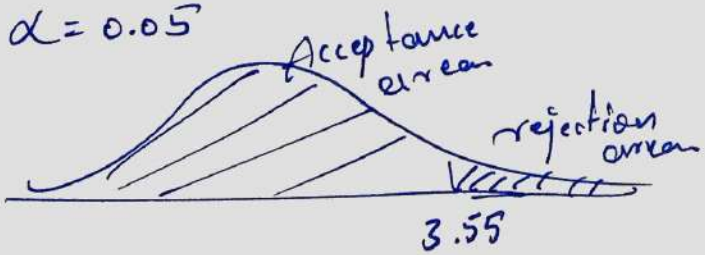
df within $= N - a = 18$

$(2, 18) =$ F test table

df total $= N - 1 = 21 - 1 = 20$

Decision boundary

$\alpha = 0.05$



if f test is > 3.55, we reject the Ho.

⑤ Calculate F test Statistics.

	SS	df	MS	F
between	98.67	2	49.34	86.56
within	10.29	18	0.54	
total.	108.95	20		

⑱ $SS_{b/w} = \frac{\sum(\sum a_i)^2}{n} - \frac{T^2}{N}$

15 mg = 9 + 8 + 7 + 8 + 8 + 9 * 8 = 57

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

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

} T^2

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

$= \boxed{98.67}$

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

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

Σy^2 = Sum of squares of each example

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

$$= \boxed{10.29}$$

---

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

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

④ MS.

$$\frac{\text{Mean Sq between}}{\text{Mean Sq within}}$$

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

$86.56 > 3.556$, Reject the null hypothesis