Statistics →

- Descriptive
- Inferential (predictive)

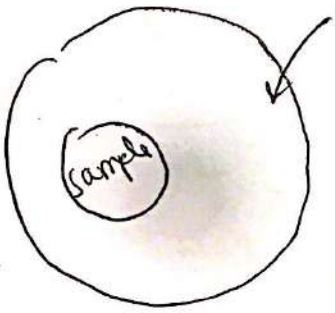
Descriptive Statistics →

Central tendencies of Data Set

1) mean
2) median
3) mode

} you should know how to calculate

Sample vs population mean :- →



suppose I want to calculate average height of people of america, Then, don't take complete population instead takes the mean of sample.

μ = population mean = $\frac{\sum_{n=1}^{N} x_n}{N}$

$\bar{x}$ = sample mean = $\sum_{n=1}^{n} x_n$

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N} = \text{Population mean}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \text{Sample mean}$$

⟶ If the distribution is skewed then median is the best center of measure.

Dispersion :- (variance)

2, 3, 3, 3

$$\mu = \frac{2+2+3+3}{4} = 2.5$$

0, 0, 5, 5

$$\mu = \frac{0+0+5+5}{4} = 2$$

They are more dispersed because numbers are farthest away from mean

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

(variance)

2, 2, 3, 3

	x_i	μ	$x_i - \mu$	$(x_i - \mu)^2$
1	2	2.5	-0.5	0.25
2	2	2.5	-0.5	0.25
3	3	2.5	-0.5	0.25
4	3	2.5	-0.5	0.25

$$\sigma^2 = \frac{0.25 + 0.25 + 0.25 + 0.25}{4} = \frac{1}{4} = 0.25$$

0, 0, 5, 5

$$\sigma^2 = (-2.5)^2 + \cdots$$

$$\sigma^2 = 6.25 + 6.$$

$$\sigma^2 = 6.2$$

We have two data sets, mean of both data sets are equal but variance of data sets differ.

$\sigma^2 = 0.25$ $\sigma^2 = 6.25$

→ larger variance means numbers in the data sets are much farther away than the mean than the number in the second data set which have less variance.

Sample variance :-

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

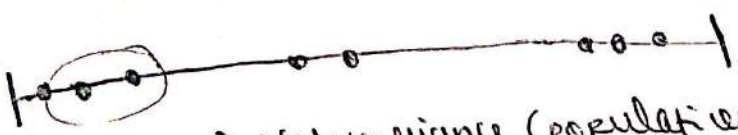
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

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

($\bar{x}$ ↓ sample mean)

↓ sample variance

→ sometimes it may be possible that we choose data set which is very skewed / दूर तक फैला हुआ



→ Net variance (population) lies in between but if you choose / for suppose, first three data set

population variance. population variance
नहीं बीच में होता । but if हमारे sample
ही बोहत skewed सा select वाला है तो
variance of sample & variance of population
काफी फरक हो सकता है, then in that
case we use.

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

s^2 ↓ Sample variance

] → unbiased estimator

Standard deviation :-

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

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

Alternate variance formulas :-

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

$$= \sum_{i=1}^{N} x_i^2 - 2\mu \sum_{i=1}^{N} x_i + \mu^2 \sum_{i=1}^{N} 1$$

$$= \frac{\sum_{i=1}^{N} x_i^2 - 2\mu \sum_{i=1}^{N} x_i + \mu^2 N}{N}$$

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

$$= \frac{\sum_{i=1}^{N} x_i^2}{N} - 2\mu^2 + \mu^2$$

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

σ^2 → variance.

Nominal variable :- categorical variable.

ordinary variable :- They are also categorical variable.

Inter Quartile range $= (Q_3 - Q_1)$

NOTE :- Q_2 is median

Random Variables (X)

→ we can't solve the random variable.

$$X = \begin{cases} 1, & \text{if rains} \\ 0, & \text{no rains} \end{cases}$$

→ random variable is name of the that maps us from random process.

Real line number.

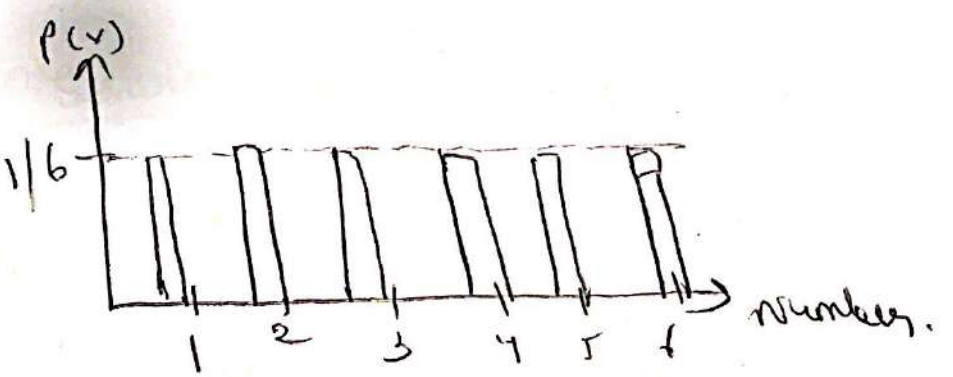
$$X = \begin{cases} 1, & \text{if heads} \\ 0, & \text{if tails} \end{cases}$$

→ Random variable, जिस variable की value कुछ भी हो सकती है।

Random variables

- Discrete
  - (i) roll a die
  - (ii) toss a coin
  - ↓ We have countable number of outcomes
- Continuous
  - ↓ infinite number of outcomes
  - ↓ X = exact amount of rain in inches

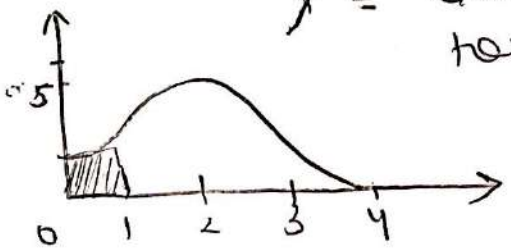
Probability distribution of X = # number facing up on the fair dice :-



$$X = \begin{cases} 1, & \text{if head} \\ 0, & \text{if tail.} \end{cases}$$

Probability density functions →

X = exact amount of rain tomorrow



$P(Y=2) = .5$

$P(|Y-2| < 1)$

$P(1.9 < Y < 2.1)$

Q what is the probability that rain will occur less than 1 inches?

Ans:

$$\int_0^1 f(x)\,dx$$ } we integrate the function from 0 to 1.

→ Area under the curve (the complete area) is equal to 1.

Binomial distribution :-

Q Coin flip 5 times.

X = No of heads after 5 flips

$P(X=0)$ ← probability of getting no head

$P(TTTTT) = \left(\frac{1}{2}\right)^5 = \frac{1}{32}$

$P(X=1)$ ← probability of exactly one head

$= \frac{5}{32}$

$P(X=2)$ ← Probability exactly two heads $= \frac{5C_2}{32} = \frac{5!}{3! \times 2!} = \frac{}{32}$

$nC_x p^x q^{n-x}$

$5C_2 \times \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^3$

$5C_2 \times \frac{1}{32}$

$= \frac{5 \times 4 \times 3!}{2 \times 1 \times 3!} = \frac{10}{32} = \frac{5}{16}$

$P(X=3)$

↗ Probability exactly three heads $= \frac{5C_3}{32} = \frac{10}{32}$

$P(X=4) = \frac{5C_4}{32} = \frac{5!}{1! \times 4!} = \frac{5}{32} = \frac{5}{32}$

↗ Probability exactly four heads.

$P(X=5) = \frac{1}{32}$

↗ Probability getting all heads. (5 heads).

Now,

$P(X=0) = \frac{1}{32}$

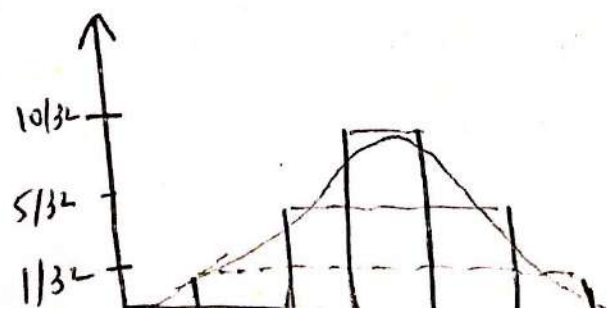
$P(X=1) = \frac{5}{32}$

$P(X=2) = \frac{10}{32} = \frac{5}{16}$

$P(X=3) = \frac{10}{32} = \frac{5}{16}$

$P(X=4) = \frac{5}{32}$

$P(X=5) = \frac{1}{32}$



Q11

$n = 6$ (basketball)

p of success = 30%

X = number of shots I made

p of failure = 70%

$P(X=0) = 0.7 \times 0.7 \times 0.7 \times 0.7 \times 0.7 \times 0.7$

$= (0.7)^6$

↓ probability that I made no shots

$P(X=1) = 0.3 \times (0.7)^5 \times 6$

↳ p that I made 1 shot

$P(X=2) = {}^6C_2 \times (0.3)^2 \times (0.7)^4$

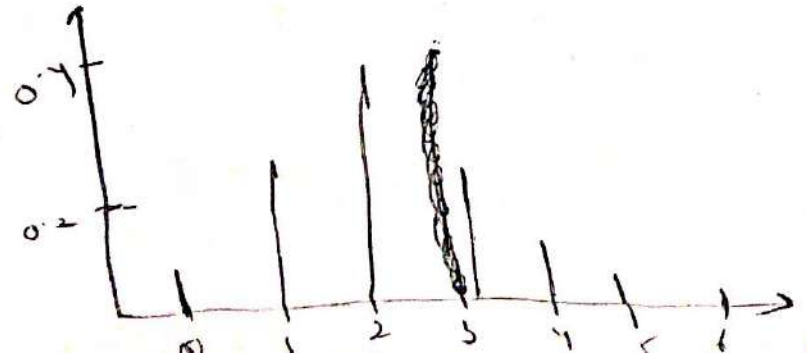
$P(X=3) = {}^6C_3 \times (0.3)^3 \times (0.7)^3$

$P(X=4) = {}^6C_4 \times (0.3)^4 \times (0.7)^2$

$P(X=5) = {}^6C_5 \times (0.3)^5 \times (0.7)$

$P(X=6) = {}^6C_6 \times (0.3)^6 \times (0.7)^0$

$= (0.3)^6$



Expected value of a Random variable :-

3 3 3 4 5

$$\frac{3+3+3+4+5}{5} = \frac{18}{5} = 3.6$$

$$\frac{3(3)+1(4)+1(5)}{5} = \frac{1}{5}\left(3(3)+1(4)+1(5)\right)$$

$$= \frac{3}{5}\cdot 3 + \frac{1}{5}\cdot 4 + 1$$

$$= 0.6 \times 3 + 0.2 \times 0.4 + \ldots$$

$$= 60\% \text{ of } 3 + 20\% \text{ of } 4 + 20\% \text{ of } 5$$

60% of numbers are 3
20% of numbers are 4
20% of numbers are 5

→ If we have the infinite population then we can find out the arithmetic mean of the infinite population by this technique.

$$E(x) = \left(60\% \text{ of } 3 + 20\% \text{ of } 4 + 20\% \text{ of } 5\right)$$

→ expected value is same as mean but दोनों अलग different है।

Probability of k success in a Binomial distribution is :-

$$P(X=k) = {}^{n}C_{k}\, p^{q}\, q^{n-q}$$

Now,

E(X) (expected value) of # of successes with probability p after n trials

$$E(X) = n \cdot p$$

For eg:-

No. of baskets I make after 10 shots. Probability of success is 40%. ?

Ans: $E(X) = 0.4 \times 10 = 4$

(0.4 ↑ probability)

Hence, I can make 4 baskets if I am throwing it 10 times.

$$E(X) = \sum_{k=0}^{n} K\, {}^{n}C_{k}\, p^{K} (1-p)^{n-K}$$

$$= np \sum_{K=1}^{n} \frac{(n-1)!}{(k-1)!\,(n-k)!} \times p^{(k-1)} \times (1-p)^{n-k}$$

$$= n p \sum_{a=0}^{b} \binom{b}{a} p^a (1-p)^{b-a}$$

↳ this will sum up to 1

$$E(X) = np$$ → expected value of binomial distribution

↳ Expected value is n times p where n is number of trials an p is probability of success.

→ This is only true for random variable X whose probability distribution is the binomial distribution

Poisson Process :-

X = No. of cars pass in an hour

$E(X) = \lambda = n \cdot p$ (success in an hour; success in an small interval)

$\lambda^{cars/hour} = 60\ min/hour \times \frac{\lambda}{60}\ cars$

we assume 1 car passes in 1 min.

$$P(X=K) = {}^{60}C_K \times \left(\frac{\lambda}{60}\right)^K \times \left(1-\frac{\lambda}{60}\right)$$

here we assume that 1 car passes in 1 sec

$$P(X=K) = {}^{3600}C_K \times \left(\frac{\lambda}{3600}\right)^K \times \left(1-\frac{\lambda}{3600}\right)$$

↳ we make it more granular.

→ We make more granular.

So,

$$\lim_{x \to \infty} \left(1 + \frac{a}{x}\right)^x = e^a$$

So, Now, we put limit to n as it is assume infinite cars passes in sec. → to make it more granular.

$$P(X = K) = \lim_{n \to \infty} \binom{n}{K} \left(\frac{\lambda}{n}\right)^K \left(1 - \frac{\lambda}{n}\right)^{n-K}$$

Solving this up we get

$$P(X = K) = \lim_{n \to \infty} \binom{n}{K} \left(\frac{\lambda}{n}\right)^K \left(1 - \frac{\lambda}{n}\right)^{n-K}$$

$P(X = K)$ — no. of cars passes in an hour.

$$= \frac{\lambda^K}{K!} e^{-\lambda}$$

λ is mean
K is required succes.
$\lambda = np$

I assume, that 9 cars passes in an hour? Find probability that exactly two cars pass in an hour?

$$P(X = 2) = \frac{(9)^2}{2!} e^{-9}$$

$$= \frac{e^{-m} m^r}{r!}$$

m = mean
r = required succes

NOTE :- (i) हम Poisson तब use करेंगे and Binomial तब use करेंगे, मगर हमारी n की value बहुत large हो जाए for eg:- $n = 100$, एक सिक्के को 100, 200 बार toss करने पर जाए then we use Poisson's in that case.

$$P(x \text{ success}) = \frac{e^{-m} \times m^{x}}{x!}$$

n = no of trials
m = mean
x = required success.

$$m = np$$

(ii) If experiment repeat upto N times

Frequency of x success = $N P(x)$

(Expected value)

$$E(X) = N \frac{e^{-m} m^{x}}{x!}$$

Q1 Find the probability that at most 5 defective fuses will be found in a box of 200 fuses if experience shows that 2% of such fuses are defective?

Ans: $p = \frac{2}{100} = 0.02$

$q = 1 - 0.02 = .98$

$n = 200$

$\text{mean} = np = 0.02 \times 200$

$= 4.0$

$x = 0, 1, 2, 3, 4, 5$

Poisons distribution $P(x) = \frac{e^{-m} m^x}{x!}$

$P(0,1,2,3,4,5) = P(0) + P(1) + P(2) + P(3) + P(4) + P(5)$

$= \frac{e^{-4}(4)^0}{\lfloor 0} + \cdots$

2) probability that an individual suffer a bad reaction from a certain injection is 0.001, determine the probability that out of 2000 individuals

(i) exactly 3
(ii) more than 2 individuals
(iii) None
(iv) more than 1 individual will suffer a bad reaction.?

Ans: $p = 0.001$, $n = 2000$

λ or $m = 0.001 \times 2000$

$= 2$

(i) $x = 3$

$P(3) = P(3) = e^{-2} \times (2)^3 / 3!$

(ii) $P(X = 3, 4, 5, \cdots 2000)$

$= P(3) + P(4) + \cdots P(2000)$

$= 1 - [P(0) + P(1) + P(2)$

(iii) $X = 0$

(iv) $P(X = 2, 3, 4 \cdots 2000) = 1 - [P(0) + P($

NOTE :- poisons में n की value बड़ी होती है और p की value छोटी होती है, binomial में vice versa होता है।

Law of large Numbers :-

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots X_n}{n}$$

$\bar{X}_n \to \mu$ for $n \to \infty$.

It says that our sample mean approaches to population mean if $n \to \infty$.

X = No. of heads after 100 tosses of fair coin

Now, $E(X) = n \times p$

$= 100 \times \frac{1}{2} = 50$

Suppose.

I shake the box one time I get 55

ready, then I shake it again

I get 65, I get 45

$$\bar{X}_n = \frac{55 + 65 + 45 + \cdots + n}{n}$$

$$\bar{X}_n \to 50 \text{ as } n \to \infty$$ } law of large numbers

Normal Distribution :-

(Gaussian distribution) → (Bell curve) → showing data near to the mean are more frequent in occurrence than data far from it.

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

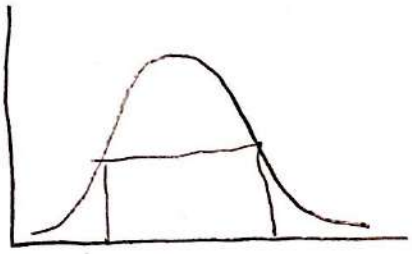
Standard z score.

σ^2 = variance

σ = standard deviation

μ = mean.

x = value.



Note :-

Binomial / poisson is discrete random variable but Normal distribution is such random variable is continuous.

$a \longrightarrow x \longrightarrow$ lies between interval.

(where $\frac{a}{a} = x$ (it is continuous)

→ Random variable here lies in a interval and it is continuous.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

$(-\infty < x < \infty)$

→ ~~दर्शाता~~ random variable infinite ~~है~~

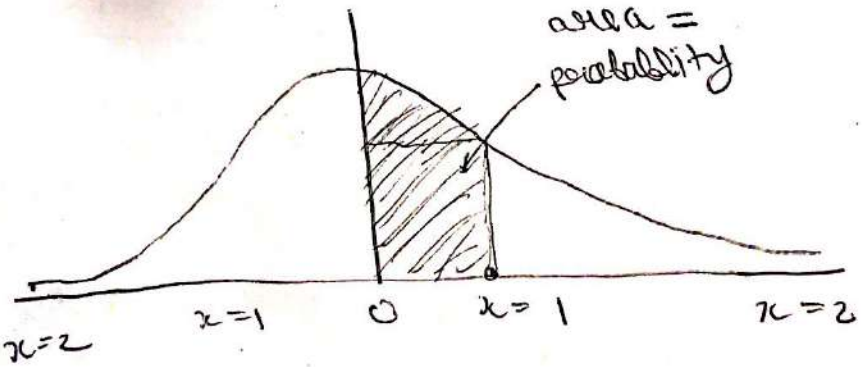
1) Standard Normal distribution :-

mean $\mu = 0$

Standard deviation $\sigma = 1$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2}$$

Standard Normal distribution



→ Standard Normal distribution is also called Z-distribution.

Step1 :- Convert x into standard Normal variate by formula : $z = \frac{x-\mu}{\sigma}$ — (1)

Step2 : Find the limits of z corresponding to the limits of x

when $x = a$ then $z = \frac{a-\mu}{\sigma}$

[put $x = a$ in (1)]

when $x = b$ then $z = \frac{b-\mu}{\sigma}$

[put $x = b$ into (1)]

Thus, the limits of z are $\frac{a-\mu}{\sigma}$ to $\frac{b-\mu}{\sigma}$

when $x = a$, $x = b$

Step3 : $P(a < x < b) = P\left(\frac{a-\mu}{\sigma} < z < \frac{b-\mu}{\sigma}\right)$

use Normal table.

Q The mean height of 500 students is 151 cm and the standard deviation is 15 cm ? Assuming that the heights are normally distributed, Find how many students have height between 120 and 155 cms

Ans:→ Height is a continuous variable

Given mean

$$\mu = 151$$

Standard deviation $= \sigma = 15$ cm

Now,

$$X = 120 \Rightarrow Z = \frac{120 - 151}{15} = \frac{-31}{15} = -2.066$$

$$= -2.07$$

when $X = 155 \Rightarrow Z = \frac{155 - 151}{15} = \frac{4}{15} = 0.266$

$$= 0.27$$

$$P(120 < X < 155) = P(-2.07 < Z < 0.27)$$

Probability that when X lies between 120 to 155 is equal to $-2.07 < Z < 0.27$

$$= P(-2.07 < Z < 0) + P(0 < Z < 0.27)$$

$$= 0.4808 + 0.1064$$

$$= 0.5872$$

} Z ka value ke liye we have standard z table

No. of students whose weights lies b/w 120 cm to 155 cm $= N \times P(120 < X < 155)$

$= 500 \times 0.5872$

$= 284$ (approx)

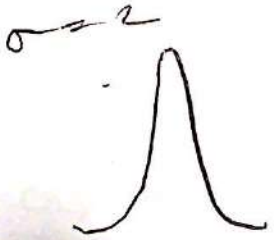
$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

↳ very imp.

Note :-

① If $\sigma = 5$ $\sigma = 10$

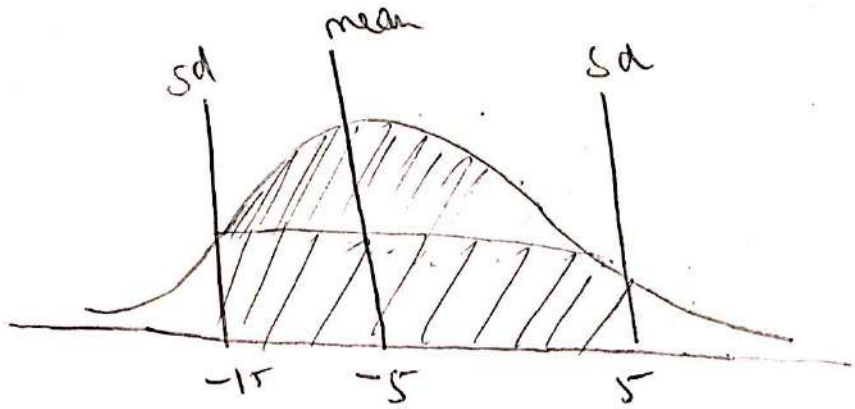
$\sigma = 2$



Cumulative density function :-

$$CDF = \int_{-\infty}^{x} P(x)\,dx$$

what is probability of getting -1 or less than $x = -1$



what is probability of standard deviation of the range of assuming normal distribution?

$$\int_{-15}^{5} p(x)\,dx \Rightarrow 68.3\%$$

Q which of the following data sets is most likely to be normally distributed? For other choices, explain why they are not normally distributed?

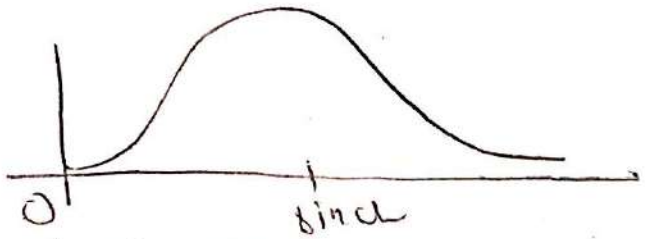
(a) hand span (measured from tip of thumb to tip of the extended 5th finger) of a random sample of high school seniors.

(b) annual salary of all employees of large Flipper's company

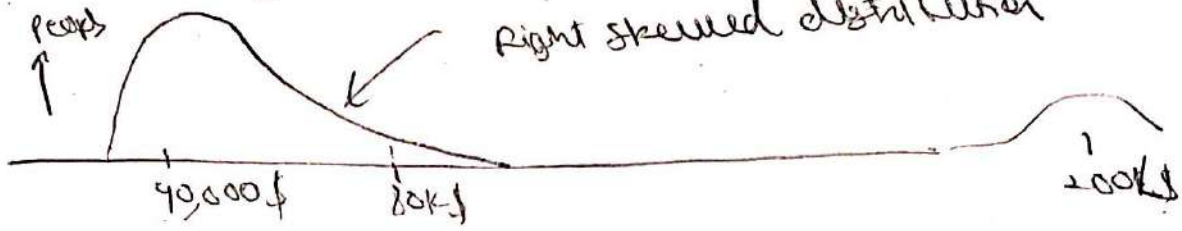
(c) annual salaries of random sample of 50 CEO's of major companies, 25 women & 25 men.

(d) dates of 100 pennies taken from a cash drawer to a convience store?

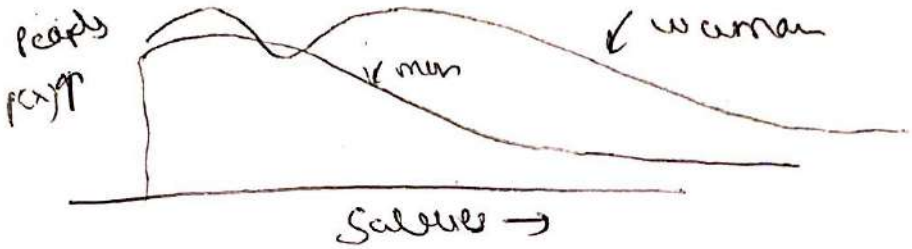
(a)



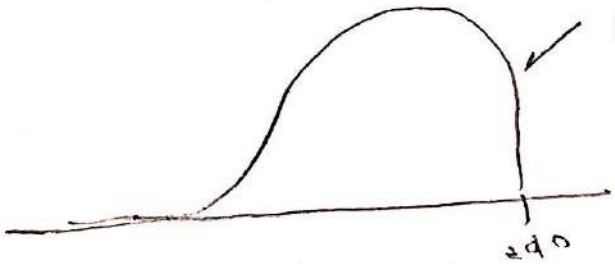
(b)



(c)



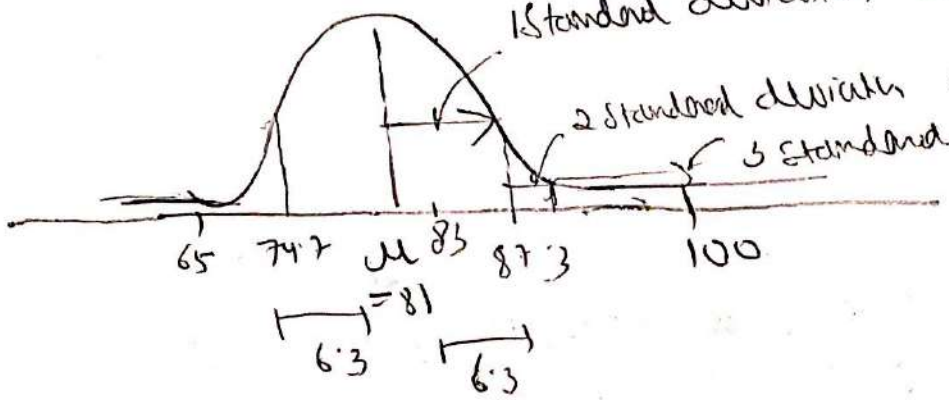
(d)



(e) Grades on a statistics mid-term for a high school are normally distributed with $\mu = 81$ and $\sigma = 6.3$. ? Calculate z scores for each of the following exam grades, Draw & label sketch for each example.

z score is how many Standard deviations away from the mean (μ).

(←)



100 is 3.02 S.d above mean

(a) 65 so, $\frac{65-81}{6.3} = \frac{-16}{6.3} = -2.54$

$Z = \frac{x-\mu}{\sigma}$ → It means it 65 is (-2.54) Standard deviation away from mean.

(b) 83, $z = \frac{x-\mu}{\sigma} = \frac{83-81}{6.3} = \frac{2}{6.3} = 0.32$

→ 83 is 0.32 Standard deviation away from mean.

→ We can say 83 is two grade above the mean i.e 81 but in terms of standard deviation we have find out z score, value I can say 83 is 0.32 sd away from mean.

(c) ✓

(d) 100

$\frac{100-81}{63} = \frac{19}{6.3} = 3.02$

It means 100 is 3.02 standard deviation above the mean.

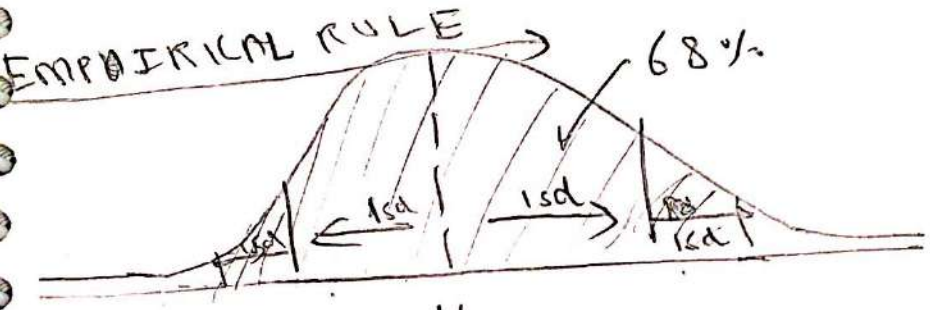
Assum that mean weight of 1 year old girls in US is normally distributed with a mean of about 9.5 grams with standard deviation of approx 1.1 grams. ? witham using calculator, estimate the percentage of 1 year old girl in the US that meets the following conditions?

(a) less than 8.4 kg

(b) between 7.3 kg & 11.7 kg

(c) more than 12.8 kg.

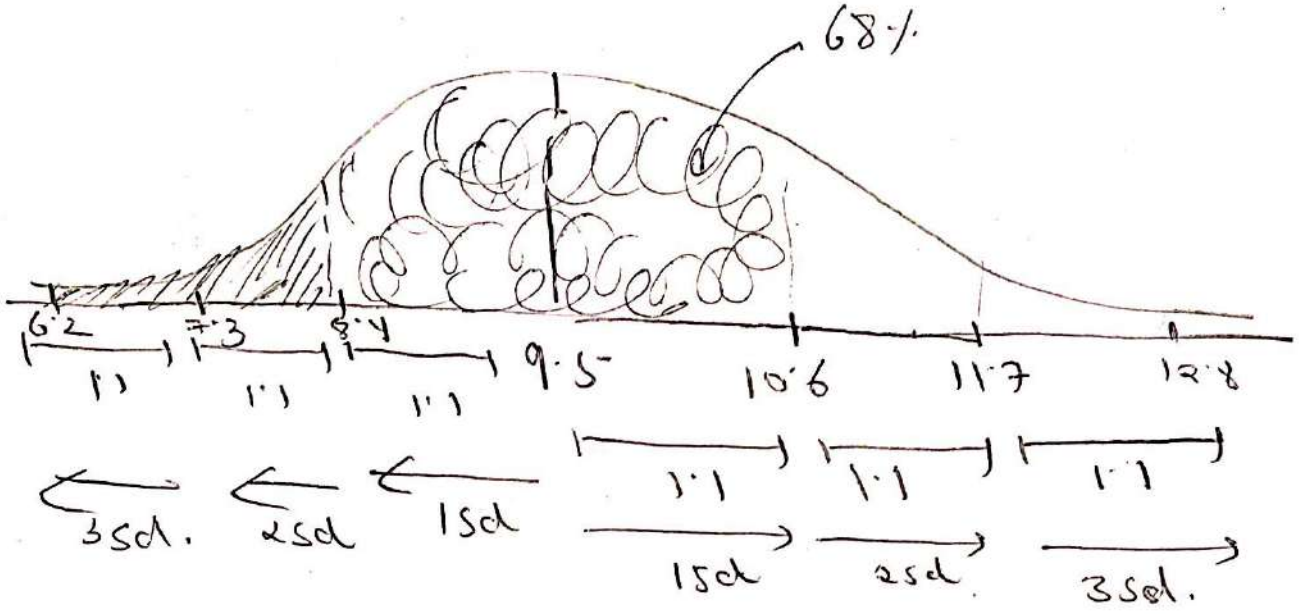
Ans: $\mu = 9.5$ gram $\quad \sigma = 1.1$ gram



→ one standard deviation below mean & one standard deviation above mean then probability is 68%.

→ 2 sd below mean & 2 sd above mean then probability is 95%.

→ 3 sd below mean & 3 sd above mean then probability is 99.7%.



(a) below or less than 8.4

hence, area below or less than 8.4 is

$$\frac{100 - 68\%}{2}$$

$$= 16\%$$

(b) between 7.3 and 11.7

It is 95%.

(c) more than 12.8 area between -3sd and +3sd = 99.7%

$$= \frac{100 - 99.7}{2}$$

$$= \frac{0.3}{2} = 0.15\%$$

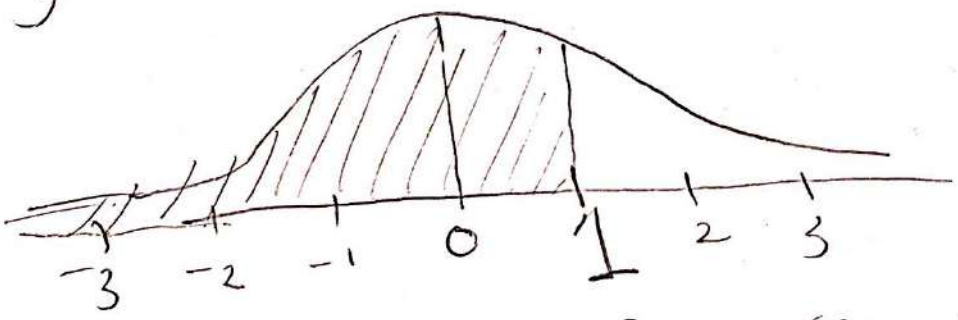
Q. For standard Normal distribution, place the following in order from smallest to largest :-

(a) percentage of data below 1
(b) percentage of data below -1
(c) Mean
(d) standard deviation.
(e) percentage of data above 2.

Therefore, standard Normal distribution

$\mu = 0$ and $\sigma = 1$

(a)



empirical rule → (68 - 95 - 99.7)

$50\% + \frac{68\%}{2}$

$50\% + 34\% = 84\% = .84$

(b) Below (-1) area

$\frac{100 - 68\%}{2} = 16\% = .16$

(c) $\mu = 0$

(d) $\sigma = 1$

(e)

Below, 2 is

$$50\% + \frac{95\%}{2}$$

Above 2 is

$$100 - \left[50\% + \frac{95}{2}\right]$$

$$= 50 - \frac{95}{2} \quad 47.5$$

$$= 2.5\% = .025$$

Now, we can arrange them,

$$c < e < b < a < d$$

Q In the 2007, AP Statistics examination scores were not normally distributed, with $\mu = 2.80$ and $\sigma = 1.34$. What is approximate z-score that corresponds to an exam score of 5 (Score range from 1-5)?

Ans:-

$$Z\text{score} = \frac{5 - 2.8}{1.34} = \frac{2.2}{1.34} = 1.64$$

(how many sd you are away from mean.

Z score can be calculated for normal distribution as well as Not normal distribution too.

Central Limit Theorem :-

→ As your sample size → ∞ then we will get perfect normal distribution curve.

→ Suppose, sampling distribution of the sample mean of population.

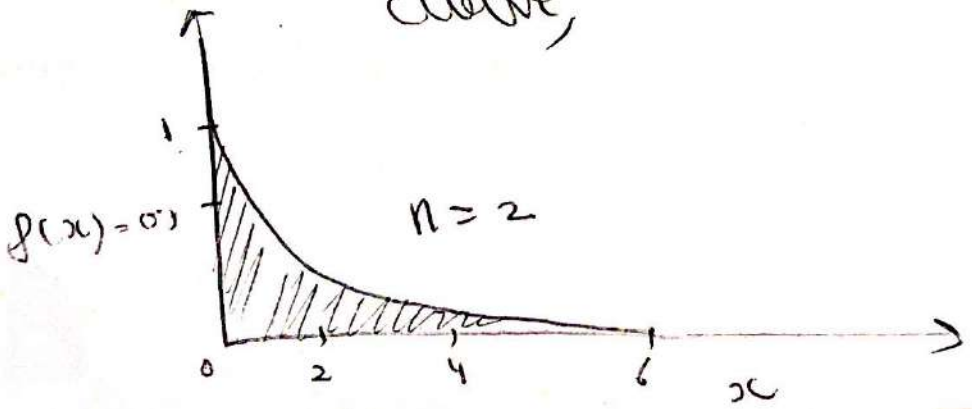
→ If population is normally distributed then our sample mean ($\bar{x}$) is also normally distributed.

→ But if population is not Normally distributed then, ?

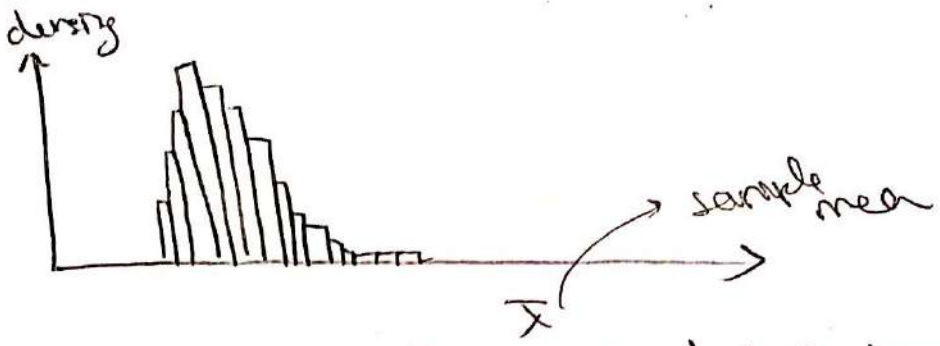
Central limit theorem

→ The distribution of the sample mean tends towards the normal distribution as sample size increases, regardless the distribution from which we are sampling

Let's suppose, we have not normally distributed curve,

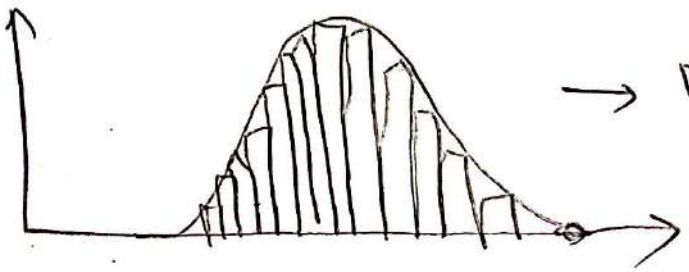


→ let take two observations $n = 2$ and take a mean out of it. Take 1000 of such observation and take mean.



If $n = 2$ then sampling distribution of the sample mean is not normal.

→ Now what if I increase sample size as $n = 4$ & plotted millions of sample means in Histogram.



→ Now we can see it's quite normal.

→ If $n \to \infty$ (n: sample size) → ~~sample~~ distribution of sample mean tends towards Normal distribution as sample size increases.

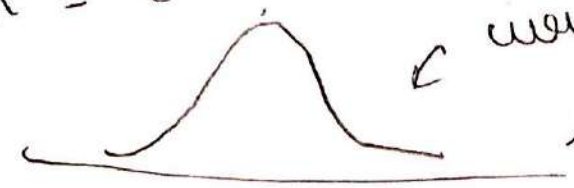
→ our sample size should atleast be 30. $n \geq 30$ for Normal distribution of sample

NOTE:-

If $n = 4$

curve for sample mean here sd is more (σ)

If $n = 20$



curve for sample mean, here sd is less (σ)

Q) why is this important?

Ans: many statistics have distributions that are approximately normal for large sample sizes, even when we are sampling from a distribution that is not normal.

→ we can often use will developed statical inference procedures that are based on normal distribution, even if we are sampling from a population that is not normal, provided we have large sample size.

$$\text{Z score} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \quad \text{as } n \to \infty$$

Q) Suppose salaries at a very large corporation have mean of $ 62000 and $\sigma = \$32,000$? If single employee is randomly selected, what is the probability salary exceeds $ 66000?

Ans: $\mu = 62000$

$\sigma = 32000$

Now $z = \frac{x - \mu}{\sigma}$

$P(X > 66000)$

$= P\left(z > \frac{66000 - 62000}{32000}\right)$

$= P(z > 0.125)$ → This question cannot be answered with the given information because salaries are not normally distributed, hence random variable X will not have the normal distribution

Now, suppose we change the question,

If 100 employees are randomly selected, what is the probability their average salary exceeds $66000?

Ans: $\mu = 62000$

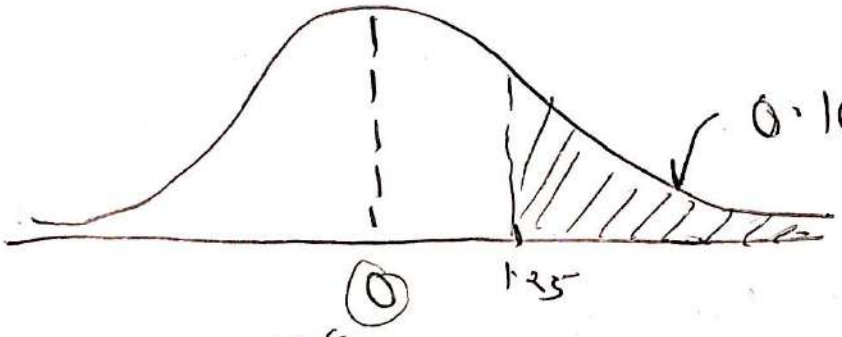
$\sigma = 32000$

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

Now, we can say that $\bar{X}$ is Normally distributed by central limit theorem.

$P\left(z > \frac{66000 - 62000}{32000/\sqrt{100}}\right) = P(z > 1.25)$

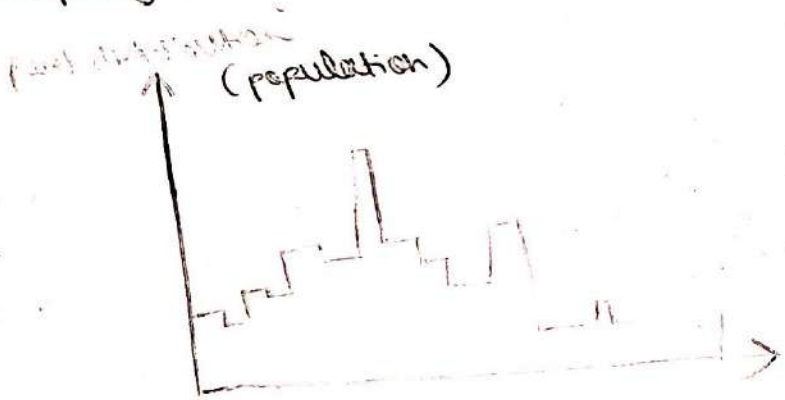
Now $z = 1.25$, Now curve is Normally distribute



standard $\mu = 0$

Measures of shape :- → Kurtosis, → Skewness

Sampling distribution of a sample mean :-



mean = 4.45

median = 11.00

sd = 9.78

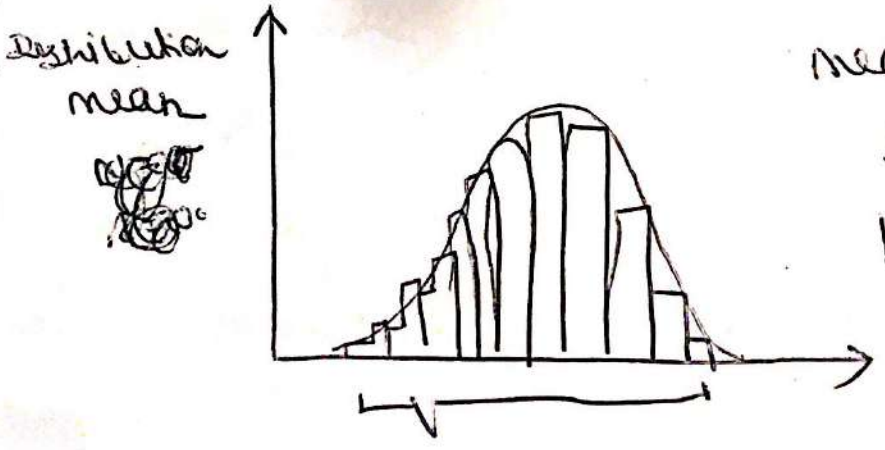
kurtosis =

$N = 4$ (4 numbers लिए then 3 बार mean निकाला)

10,000 बार किया → $n = 10000$

↳ we have taken 10,000 observations

↳ we have now (10,000 means) plot it



mean = 14.48

sd = 4.40

kurtosis = −0.26

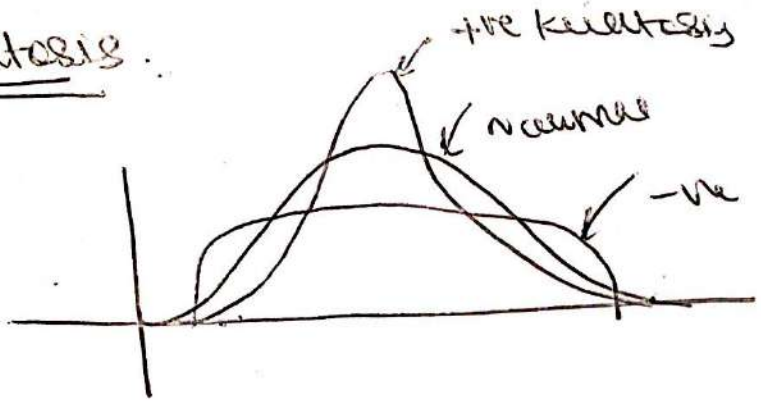
If $N = 5$

↳ complete

$= -0.64$ → $N = 25$

Note :- population std mean or sample mean when $n \to \infty$ is almost same.
↳ no. of observations

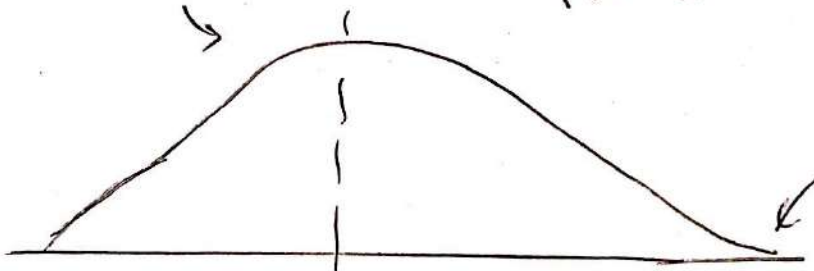
Kurtosis.



when

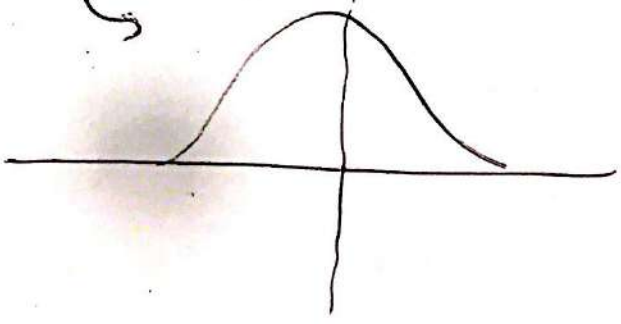
Kurtosis = -0.26
↳

Sample size / No. of observations
$N = 5$



Kurtosis = -0.04
↳

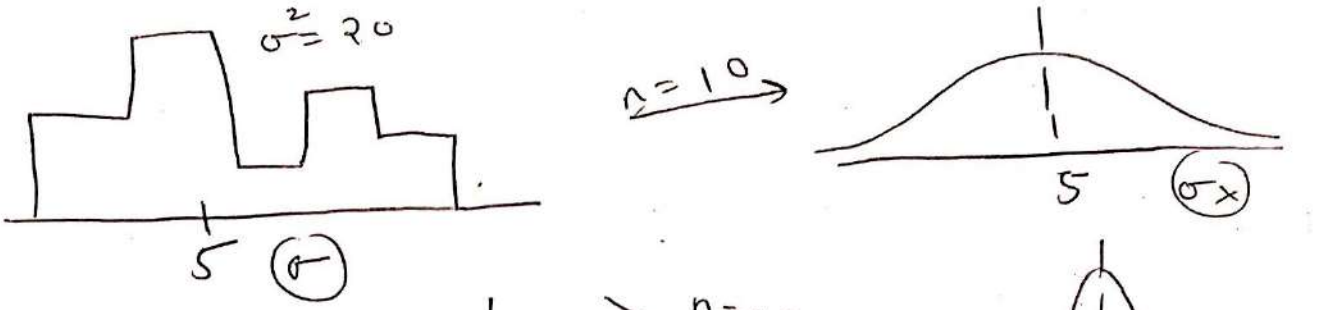
$N = 25$
↳ sample size / No. of observations

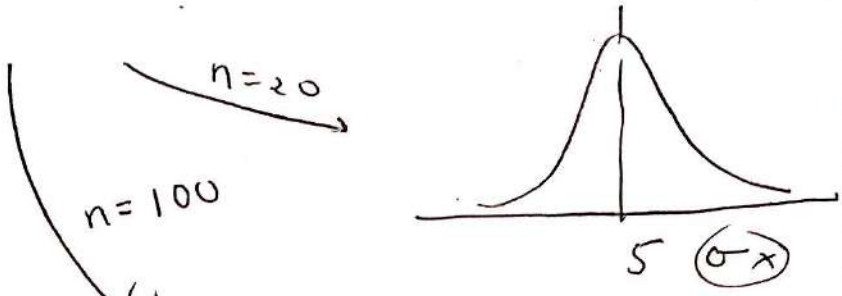


→ when N is larger (No. of sampling distribution) then mean is same in both the cases but sd is small for those whose N is larger.

$N \propto \frac{1}{sd}$ → inversely proportional.

Standard error of mean :-





$\hookrightarrow \sigma_{\bar{x}}^2 = \frac{20}{20}$

$\sigma_{\bar{x}} = 1$

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

$\sigma_{\bar{x}}^2 = \frac{20}{100}$

$\sigma_{\bar{x}}^2 = \frac{1}{5}$

$\sigma_{\bar{x}} = \frac{1}{\sqrt{5}}$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Standard deviation of the sampling distribution of the sample mean

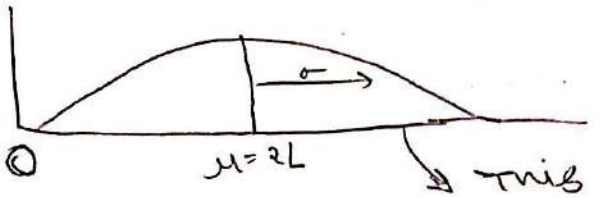
OR

Standard error of the mean.

Q Average male drinks 2L of water when active outdoor (with $\sigma = .7L$) ? you are planning a full day nature trip for 50 men and will bring 110L of water? what is probability you will run out)

Ans:

$\sigma = 0.7L$



P (run out)

= P(we more than 110L)

= P(average water used per man is greater than 2.22 /man

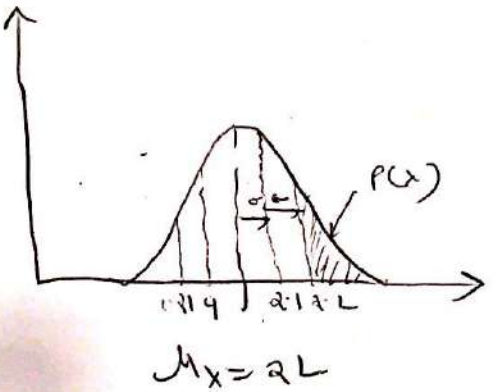
$\rightarrow \frac{110}{50} = 2.2$

Nay,

Nay,

Sampling distribution of sample mean when

$n = 50$



$\mu_x = 2L$

$\sigma^2_x = \frac{\sigma^2}{n}$

$\sigma_x = \frac{\sigma}{\sqrt{n}}$

$\sigma_x = \frac{0.7}{\sqrt{50}}$

$\sigma_x = 0.09$

$\simeq 0.1$

$P(x) = ?$

$P(x) =$ approx It is two standard deviations above the mean.

$z = \frac{2.2 - 2}{.099} = 2.02$

↳ It is 2.02 sd above the mean.

According to z table,

$$z = 0.9783$$

↳ but z table gives the area below a particular value., hence it has given the area below 2.02.

Hence,

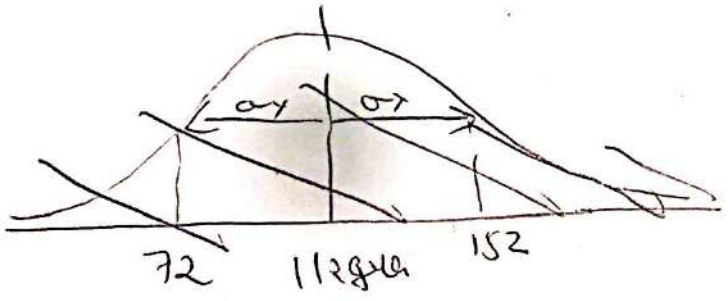
$$P(x) = 1 - 0.9783$$
$$= 0.0217$$
$$= 2.17\%$$

Example

If you sample 36 apples from your farm's harvest of over 200,000 apples? The mean weight of sample is 112 grams (with a 40 gram sample standard deviation)?

Ans: $n = 36$ $\quad \mu_x = 112$ gram

$\sigma x = 40$ gram



$$\sigma = 40$$

$$\sigma_x = \frac{40}{\sqrt{36}} = \frac{40}{6}$$

$$\sigma x = 6.67$$

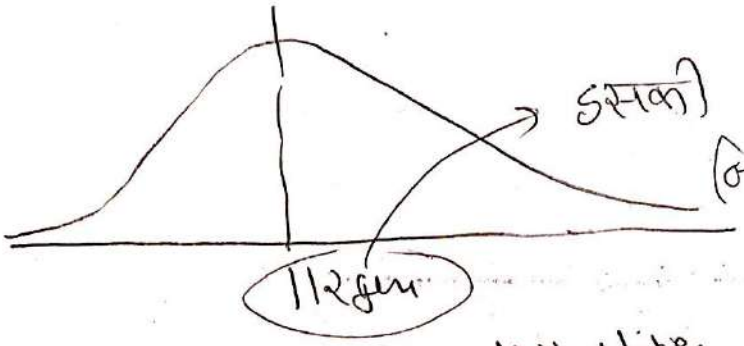
$$\sigma_x = \frac{\sigma}{\sqrt{n}}$$

$$\Rightarrow \sigma = \sigma_x \times \sqrt{n}$$

$$\Rightarrow \sigma = 40 \times \sqrt{36}$$

$$\Rightarrow \sigma = 40 \times 6 = 240$$

Now, population mean graph $\mu_{\bar{x}} = \mu = 112$ gm

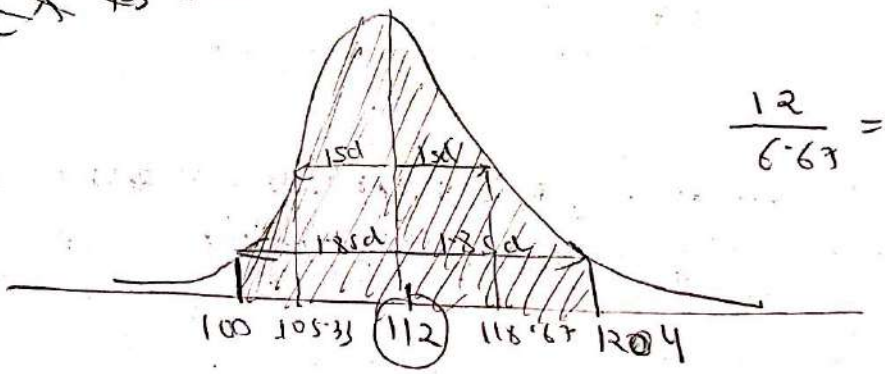


population distribution

P (population mean is within 12 of $\bar{x}$)

$= P$ ($\bar{x}$ is within 12 of μ)

$= P$ ($\bar{x}$ is within 12 of $\mu_{\bar{x}}$) [as $\mu = \mu_{\bar{x}}$]



$\frac{12}{6.67} = 1.8$

Samply distributio of $\bar{x}$ for $n = 36$

$\sigma_{\bar{x}} = 6.67$

112.00
+ 6.67
118.67

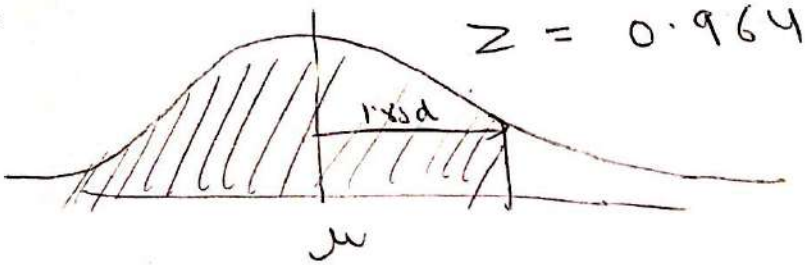
112.00
− 6.67
105.33

Go to z table,

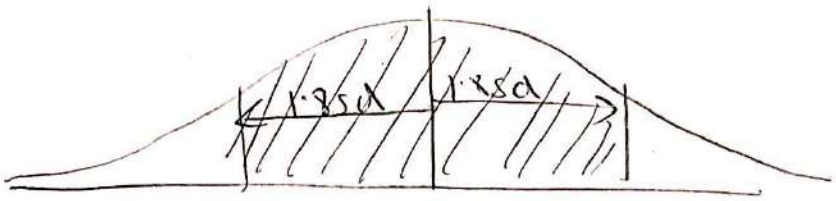
you will see z for 1.80 = 0.96

It means.

$z = 0.9641$



But you have to find out this,



Hence,

$$P(\text{required}) = (0.9\,64 - 0.5) \times 2$$

$$= \begin{array}{r} 0.9640 \\ -\ .4641 \\ \hline \end{array} \times 2$$

$$= \begin{array}{r} .96 \\ -\ .92\,82 \\ \hline \end{array}$$

$$= 0.9282$$

$$= 92.82\ \%$$

Bernoulli distribution :- only two possible outcomes when an individual trial has. It is Bernoulli random variable (discrete probability distribution)

Ex: Toss a coin once? what is the distribution of number of heads?

→ $P(\text{success}) = p$

$P(\text{failure}) = (1-p)$

Let $x = 1$ if success occurs, $X = 0$ if failure occurs

Then x has bernoulli distribution:

$$P(X = x) = p^x (1-p)^{1-x}$$

$$P(X=x) = p^x (1-p)^{1-x}$$

For $x = 0, 1$

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

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

Mean of Bernoulli $= \mu = p$

Variance of Bernoulli $= \sigma^2 = p(1-p)$

Q. Approximately 1 in 200 American adults are lawyers? One american adult is randomly selected? what is distribution of the number of lawyers?

Ans: $P(X=x) = \left(\frac{1}{200}\right)^x \left(1 - \frac{1}{200}\right)^{1-x}$

for $x = 0, 1$

$P(X=1) = \frac{1}{200}$ ↓ we get one lawyer

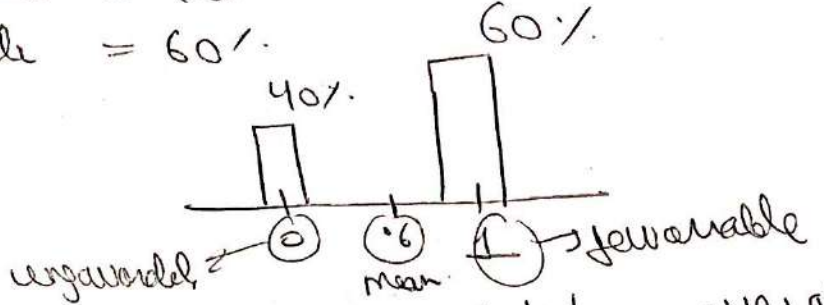
$P(X=0) = \frac{199}{200}$ ↓ we get no lawyer

Now,

Eg suppose we are taking rating

unfavourable = 40%

favourable = 60%



Mean of this distribution would be weighted value (probability weighted sums)

$$\mu = 0.4 \times 0 + 0.6 \times 1$$

$$\mu = 0.6$$

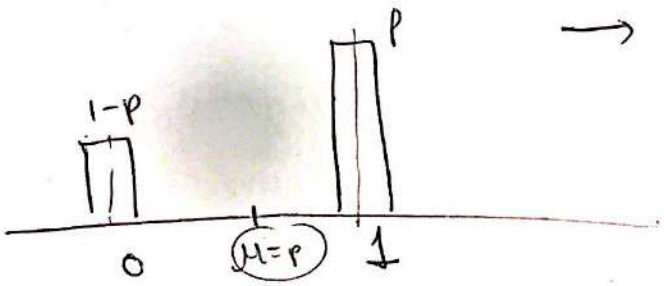
Note, **mean = probability of success = p**

$$\sigma^2 = 0.4(0 - 0.6)^2 + 0.6(1 - 0.6)^2$$

(0: x, 0.6: mean)

$$\sigma^2 = 0.24$$

$$\sigma = 0.49$$



→ Bernoulli distribution is the most simplified form of Binomial distribution.

$$\mu = (1-p) \times 0 + p \times 1 = p$$

$$\boxed{\mu = p}$$

↳ mean of Bernoulli

$$\sigma^2 = (1-p)(0-p)^2 + p(1-p)^2$$

$$\sigma^2 = p - p^2 = p(1-p)$$

$$\boxed{\sigma^2 = p(1-p)}$$

Variance of Bernoulli

Margin of error :—

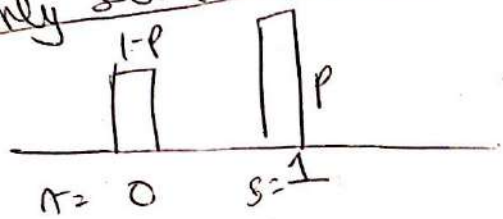
Now

Randomly surveyed 100 people from population :—

Let suppose

$p = 43\%$

$(1-p) = 57\%$



$$\overline{X} = \frac{p \times 0 + (1-p) \times 1}{n}$$

$\overline{X}$ ↓ sample mean

$$\overline{x} = \frac{57 \times 0 + 43 \times 1}{100} = 0.43$$

$$S^2 = \frac{57(0 - 0.43)^2 + 43(1 - 0.43)^2}{(100-1)}$$

S^2 ↓ variance of population

(mean, mean)

$$S^{*2} = 0.2475$$

$$S = 0.50$$

S ↓ Standard deviation

$$\sigma_x = \frac{\sigma}{\sqrt{100}} = \frac{\sigma}{10} =$$

$$\sigma_x = \frac{0.50}{10} = 0.05$$

Margin of error :-

Confidence interval :-

→ confidence intervals are based on sample data, and give range of plausible value for a parameter

Eg:-

→ we may be 95% confident that μ (parameter) lies in interval $(-0.2, 3.1)$

→ we may be 99% confident that θ (parameter) lies in the interval $(2.5, 13.4)$

Margin of error :-

→ It tells you how far off your estimate is likely to be. or how much confidence you can in your estimate

what you need for margin error :-

- Sample size = N
- Sample mean
- Standard deviation of population
- z score for level of confidence..

$(Z = 1.96)$

↳ for 95% confidence interval

How to calculate margin of error :-

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$\sigma_{\bar{x}}$ → standard error

Margin of error = standard error × z score

Eg :- $N = 44$

mean = 78

$\sigma = 5$

z score = 1.96 (for 95% confidence interval)

Ans:

$$\sigma_{\bar{x}} = \frac{5}{\sqrt{N}} = \frac{5}{\sqrt{44}} = \frac{5}{7} = 0.714$$

Standard error = 0.714

margin of error = 0.714×1.96

$= 1.40$

→ It means that 95% of the time, the population mean will be within 1.40 of the mean.

Confidence Interval = Range (mean − ME to mean + ME)

Confidence interval =

$78 - 1.40$ to $78 + 1.40$

76.6 to 79.4

To get better confidence :-

(i) Have lower sd.

(ii) Use larger sample size.

(iii) Accept broader confidence interval.

NOTE :-

→ confidence grows as the margin of error increases.

* E is symbol that we use margin of error.

* C is confidence level.

$E \uparrow \iff C \uparrow$

$E \downarrow \iff C \downarrow$

Eg :- To estimate μ, we may use $\bar{x} \mp E$,

where margin of error.

$$E = z \times \frac{\sigma}{\sqrt{n}}$$

12) A may 2000 survey found that 38% of random sample of 1012 adults said that they believe in ghost ? what is the margin of error for a 90% confidence interval for this poll?

Ans: success = 384 people who believe in ghost
failure = 627 people who don't

sample distribution :-

$$\sigma = \sqrt{p(1-p)}$$

$$\sigma = \sqrt{pq}$$

$$\sigma_x = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.38 \times 0.62}{1012}}$$

$$\sigma_x = 0.0153$$

$$E = Z \times \sigma_x \qquad \{ z = 1.65 \text{ for } 90\% \text{ confidence}$$

$$E = 1.6 \times 0.153$$

$$E = 0.252$$

$$= 2.5\%$$

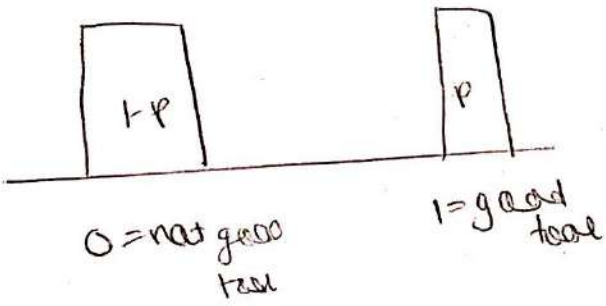
margin of error = 2.5%.

From 6250 teachers in district, 250 were randomly selected & asked if they felt that computers were an essential teaching tool for their classroom? Of those selected, 142 teachers felt computers were an essential tool.

(i) Calculate 99% confidence interval for the proportion of teacher who felt that computers are essential tool.

(ii) How could the survey be changed to narrow the confidence interval but to maintain 99% of confidence interval?

Ans:



$$\bar{X} = \frac{1\times142 + 0\times108}{250}$$

$$\bar{X} = \frac{142}{250}$$

$$p = \frac{142}{250}$$

$$\bar{X} = p = \frac{142}{250} = 0.568$$

$$\sigma = \sqrt{p(1-p)}$$

$$\sigma = \sqrt{\frac{142}{250} \times \frac{108}{250}} = 0.5$$

$$\sigma_{\bar{x}} = \frac{\text{sample sd.}}{\sqrt{n}} = \frac{0.426}{\sqrt{250}} = 0.050$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{250}} = \frac{0.50}{\sqrt{250}}$$

$$\sigma_{\bar{x}} = 0.031$$

$$E = 0.031 \times 2.58$$

$$E = 0.07$$

→ we are 99% confident, hence that $\bar{x}$ (0.568) is within 0.08 of the population (p)

$$\text{confidence interval} = 0.568 \pm 0.08$$

$$0.488 \text{ to } 0.648$$

(ii) you can take more samples

small sample size confidence interval:

⇒ patients blood pressure has been measured after having been given a new drug for 3 months? They had BP increase of 1.5, 2.9, 0.9, 3.9, 3.2, 1.1 and 1.7 Construct 95% confidence interval for true expected BP increase for all patients in population.

Ans: Sample mean

$$\bar{X} = \frac{1.5 + 2.9 + 0.9 + 3.9 + 3.2 + 2.1 + 1.9}{7}$$

$$\bar{X} = 2.34$$

$$S^2 = (1.5 - 2.3)^2 + (2.9 - 2.34)^2 + (0.9 - 2.34)^2 + (3.9 - 2.34)^2 + \cdots \text{(find it)}$$

$$S = 1.04$$

↳ standard deviation

$$\sigma \simeq S = 1.04$$

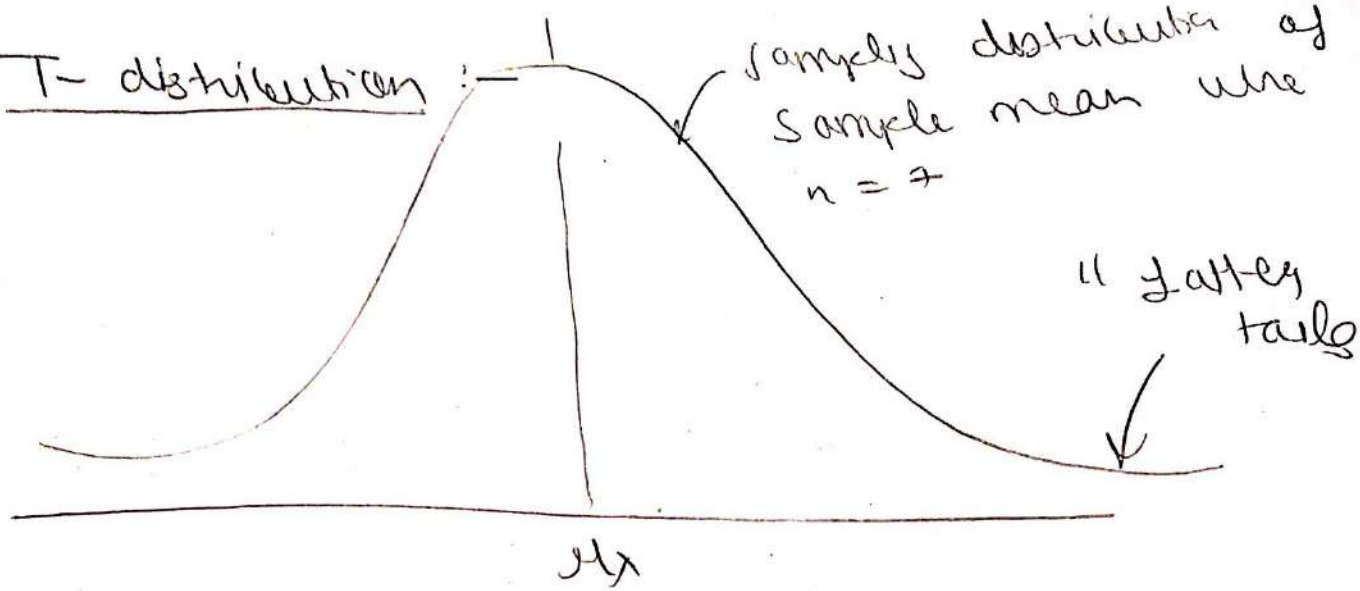
↳ not so good because n is small

$n = 7$.

→ So instead of making the normal distribution graph of the sample mean we make t-test distribution graph because our n is small $n = 7$.

→ when we have small sample size then,

T- distribution :



$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{7}} \approx \frac{s}{\sqrt{7}}$$

$$\sigma_{\bar{x}} = \frac{1.04}{\sqrt{7}} = 0.39$$

this is the bad estimation we have taken $\sigma = s$, when n is small

σ → population sd

s → sample sd

→ → T distribution is used when n is small, here → T score gives area towards the right in contrary

$$E = T \times \sigma_{\bar{x}}$$

$E = 2.447 \times 0.39$ z score gives area towards left.

$E = 0.96$

"confident"

→ 95% chance 2.34 is within 0.96 of μ

→ 95% chance μ is with 0.96 of $\bar{x}$

$(2.34 - 0.96, 2.34 + 0.96) \Rightarrow (1.38, 3.3)$ ∋ 2.34

Descriptive - Statistics

→ Descriptive statistics is the set of numbers that describe the data.

Two categories :-

1. Measures of central tendency
2. Measures of Dispersion

Measure of central tendency :-

- average ← (1) mean → = Average (c2 : c1100)
- middle observation that divides data into two parts ← (2) median → = Median (n1, n2 ...)
- (3) mode

Q when we should report median and when we should report mean?

Ans: Seven employee in firm with following salaries:-

28000 $
33000 $
33000 $
34000 $
37000 $
40000 $
400000 $

mean = 86000 $
median = 34000 $

→ hence we report median over here, since it is more appropriate.

→ mean is influenced to a greater extent by outliers.

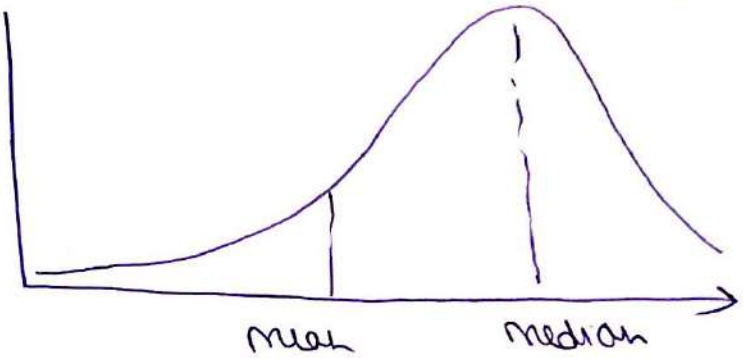
Mean vs Median :-

① Histogram

Right skewed

mean > median.

② Left skewed



median > mean.

Mode :- frequency of data.

= mode.SNGL (n1, n2, ...)

measures of spread :-

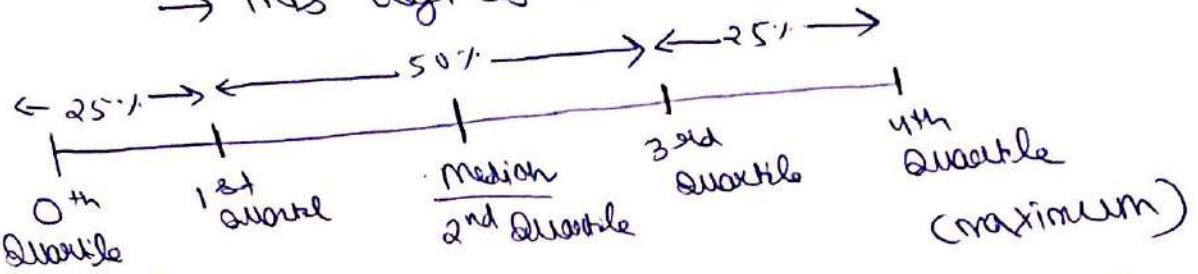
① **Range** :- max - min.

→ Higher range means greater spread in data.

Inter Quartile range :-

IQR

→ This defines 50% of data



$$IQR = (3^{rd}\ Quartile - 1^{st}\ Quartile)$$

→ we prefer Inter Quartile range over range because in statistics we prefer taking sample of data.

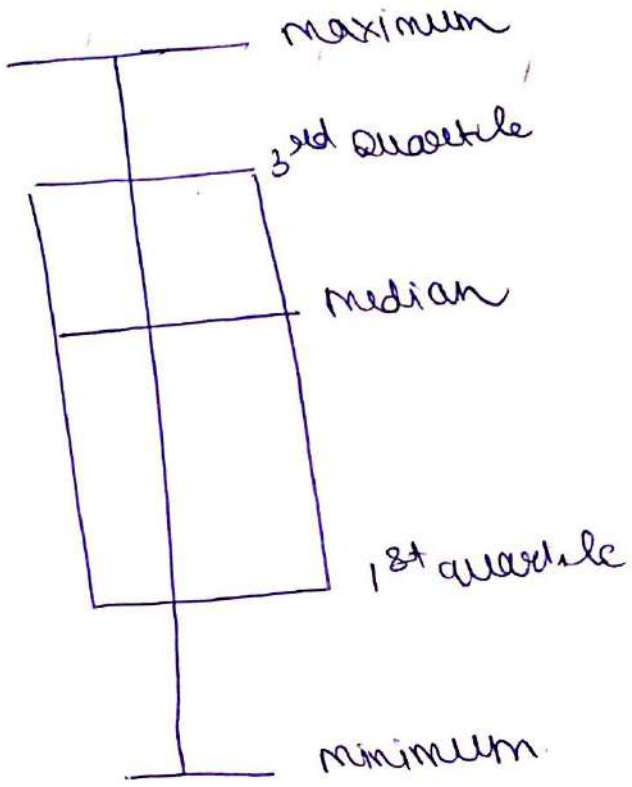
Excel :-

=Quartile.Inc (array, quart)

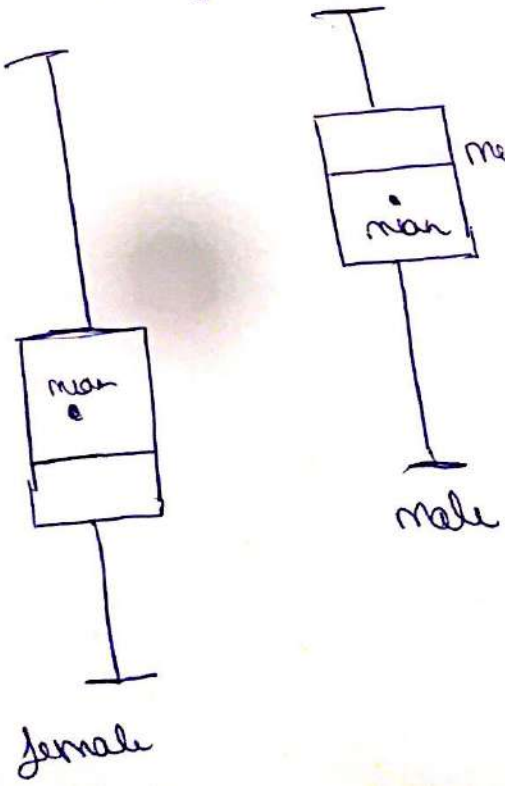
IQR = Quartile.Inc (array, 3) - Quartile.Inc (array, 1)

Box Plot :-

Visual representation of the mean and various quartiles of data.



Eg:- Hollywood stars earning



* we can infer that spread of females are more than spread of males.
* max. salary of female is equivalent to max. salary of male.
* Female is right skewed hence mean > median.
* male is left skewed hence median > mean.

③ Standard deviation

→ describes dispersion in data.

→ It calculates mean & then how far other points are deviated from mean.

= STDEV.P (number1, number2 ...)

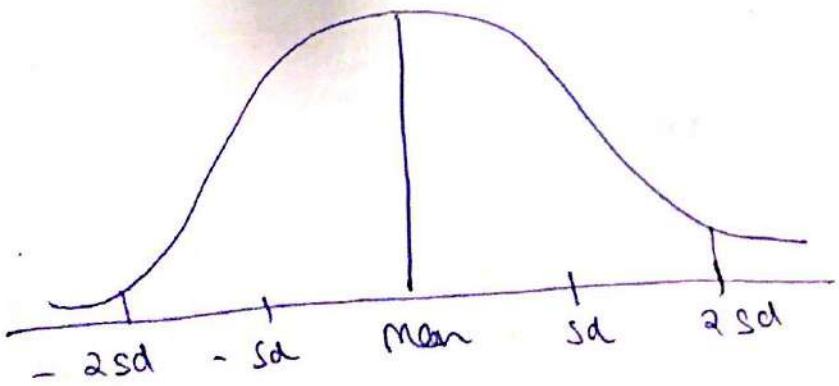
→ If we have data which was sample from some larger population of data we will use.

= STDEV.S (number1, number2 ...)

④ variance :-

Variance = $(Std)^2$

NOTE :- 68% of data lie within one standard deviation & 95% lie within 2 standard deviation from mean.



Assumption :- distribution should be Normal.

Chebyshev's theorem :—

→ At least $\left(1 - \frac{1}{k^2}\right)^{th}$ of data lie within $\pm k$ standard deviations from the mean regardless of shape of distribution

→ Specifically, It says that at least 75% of all values are within ± 2 Std from mean regardless of shape of distribution.

measures of Association :— → covariance → correlation.

Covariance :-

relationship between two variables.

$$\text{covariance of sample.} = \frac{1}{(n-1)} \sum_{i}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

= COVARIANCE.S (range 1, range 2)

→ If covariance is +ve then it indicates positive correlation.

NOTE :- covariance measure is susceptible to the unit of measurement. We can arbitrarily inflate or deflate covariance by choice of units.

Correlation :-

→ hence, covariance is not very appropriate.

→ For that correlation came,

$$\text{correlation} = \frac{\text{covariance}(X,Y)}{\text{Std}(X)\ \text{Std}(Y)}$$

$$= \text{correl}(\text{range1}, \text{range2})$$

NOTE :- **covariance**

* Range : $-\infty$ to $+\infty$
* Affected by units of measurement

correlation:

* Range : -1 to $+1$
* No effect by unit of measurement.

Causation :-

Smoking causes cancer
Cancer causes smoking

To establish Causation:

* two variable must be correlated
* causing variable must occur before caused variable.
* external variables should be ruled out.

~~Probability and Random Variables~~ :-

Statistical distributions :-

Random Experiment → Random Variable

multiple possibilities → "Salary"
of CEO salaries

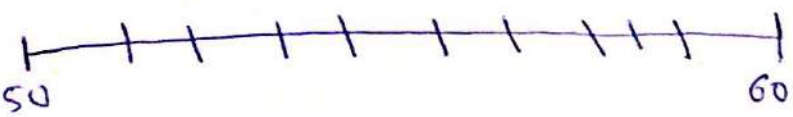
→ Statistical distribution is the tool that will help us model the salary for CEO across small firms. Benefit is that we can make predictions across different set of firms not included in the data from which the histogram is created.

Statistical distributions :-

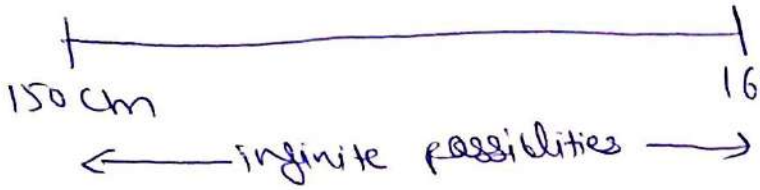
1. Beta
2. Binomial
3. Gamma
4. Poisson
5. ✓ Normal
6. T distribution

Discrete data vs continuous data :-

1. Number of students :- finite (discrete)



② Heights of Man and woman



→ continuous data.

Assumption :- many times, even though data may be strictly discrete, for eg :- annual salaries of employees, revenues earned by companies etc. we will still end up using continuous distribution to model the data.

Probability density function :-

→ It is the rule that assigns probabilities to various possible values that random variable takes.

Note :- probability of a particular outcome is always zero in continuous distribution.

Eg :-

$P(5'2) = 0$

heights
infinite

Hence this is the reason that we use range of continuous data.

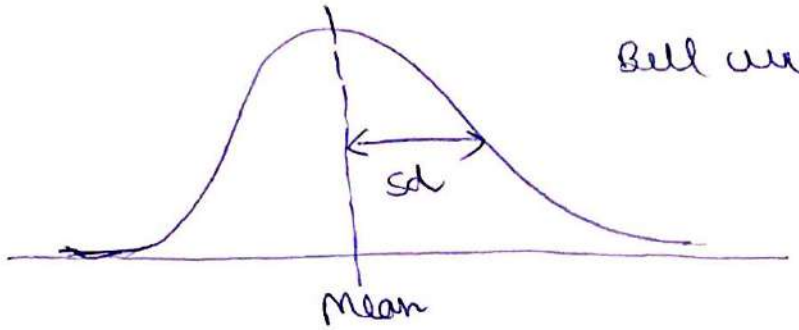
eg :- $P(\text{height} > 5'2 \text{ and less} < 5'5) = ?$

$P(\text{height} < 5\text{ feet}) = ?$

→ For discrete distribution we know how to calculate probability

→ But for continuous distribution we will use ranges.

Normal Distribution :-



Bell curve

→ Mean can have any value between $-\infty$ to $+\infty$.

$$P(x) = \int_{a}^{b} \frac{1}{\sqrt{2\pi \times Std^2}} e^{-(x-mean)^2 / 2(Std)^2} dx$$

NOTE :-

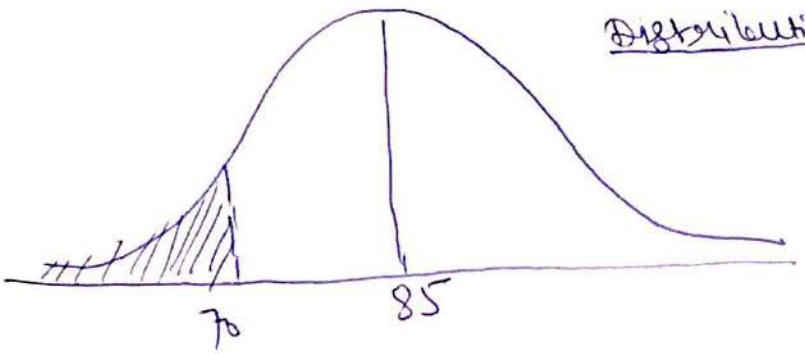
Q A bread producing company produces wheat loaves of bread in its factory. It observes that on average every day 35 loaves of bread get discarded on account of being defective. Std is 9 loaves. Predict probability that less than 20 loaves of bread will get discarded on being defective.

Ans:

Effective ~ Normal (85, 9)

Prob (Effective < 70) = ?

Distribution of effective



Syntax :-

= NORM. DIST (x, mean, std, TRUE)

x → 70
mean → 85
std → 9
TRUE → we will never use FALSE

= 0.047
= 4.7% chance

this will give area under curve to left side

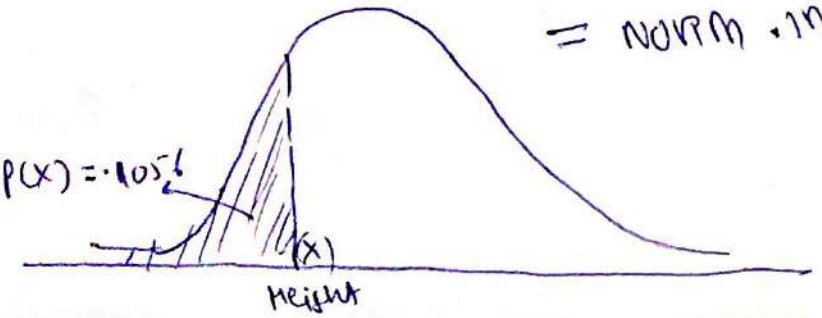
NOTE :-

Prob (Effective < 75) = Prob (Effective ≤ 75)

Now,

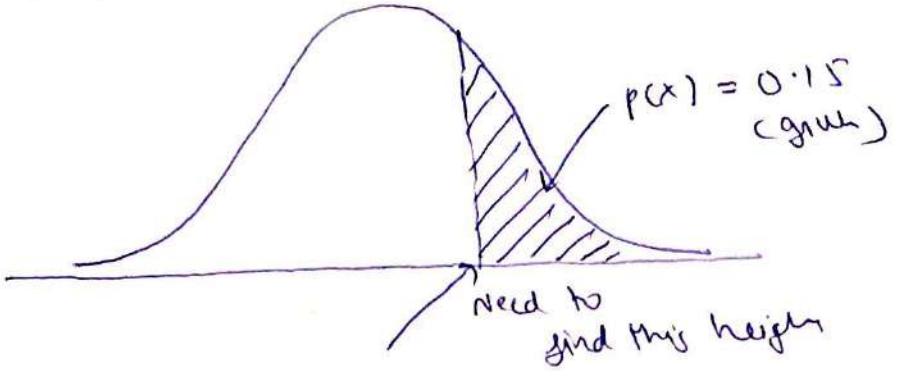
For eg :- you are given probability value and you need to find X value that is height value, you will use inverse function

= NORM.INV (probability, mean, std)



Q Find out a height such that probability Sameer's height is greater than this particular height is 0.15?

Ans:



$$P(\text{left}) = 1 - 0.15$$
$$= 0.85$$
$$= \text{NORM.INV}(0.85, \text{mean}, \text{Std})$$
$$\downarrow \quad \downarrow$$
$$5.5 \quad 0.4$$

Q John can take two roads to airport from his home.

Road A : mean = 54 minutes, Std = 3 mins

Road B : mean = 60 mins, Std = 10 mins

1) which road should he chose if on midday friday he must be at airport within 50 minutes to pick up his spouse?

Ans:

Road A

Prob (Time < 50) = NORM.DIST (50, 54, 3, TRUE)
= 0.09

Road B

Prob (Time < 50) = NORM.DIST (50, 60, 10, TRUE)
= 0.15

Binomial Distribution :-

Two popular Discreate distribution :-

1. Binomial
2. Poisson

Bernaulli process :-

Situation where random variable have only two possible outcomes.

Eg :- Pass / fail

→ NOTE :- Repeated trials to the Bernaulli process give rise to Binomial distribution.

Eg :- Probability of winning at least 4 times in 10 rolls of the dice?

win (1) loose (0)

$$P(X) = {}^{n}C_{x}\, p^{x} (1-p)^{n-x}$$

P(success)

= BINOM.DIST (x, n, p, FALSE / TRUE)

x → No. of successes

n → No. of trials

p → P of success

(success = x)

② Probality that you win 3 times in 10 rolls of dice?

$P(x=3)$ = BINOM.DIST (3, 10, 0.1667, FALSE)

= 0.1551

③ Probality that you win at most five times in 10 rolls of dice?

$P(X \leq 5)$ = BINOM.DIST (5, 10, 0.166, TRUE)

= 0.98

④ probality that you win at least three times in 10 rolls of dice?

$P(X \geq 3) = 1 - P(X \leq 2)$ = 1 − BINOM.DIST (2, 10, 0.1667, TRUE)

= 0.22

Mean of Binomial distribution = np

$Sd = \sqrt{npq}$

Poisson distribution :-

* No. of fatal traffic accidents in a city in week
* No. of customers arriving at store during given hour.

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

= POISSON.DIST(x, λ, FALSE/TRUE)

↓

(0 − ∞)

→ In Binomial distribution the random variable can take the value from 0 to the trial ends. but here in poisson Random variable can take (0 − ∞)

NOTE :- =T.INV (probability, df)

↳ we are given probality to left and we need to find the value of x

→ T distribution is symmetric about zero.

Confidence interval :-

Interval with some confidence

Eg :- Take elections.

(A) 300 vots

(B) 200 vots

P(A) = 60%, but this 60% is actually the probability from the sample, will candidate A get 60% of all vots in actual elections?

Here confidence interval comes.

95% confidence interval [55.7% ⟷ 64.3%]

60%

4.3 4.3

±4.3 is margin of error

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

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

whenever we know population std we use z statistics and when we have data about the sample we use T statistics.

Another continuous distribution

T distribution

→ Similar to normal distribution

→ However standard deviation of this distribution depends on single parameter called degree of freedom

→ As degree of freedom increases it tends towards normal distribution.

→ It is used as an tool to calculate confidence Intervals and hypothesis testing.

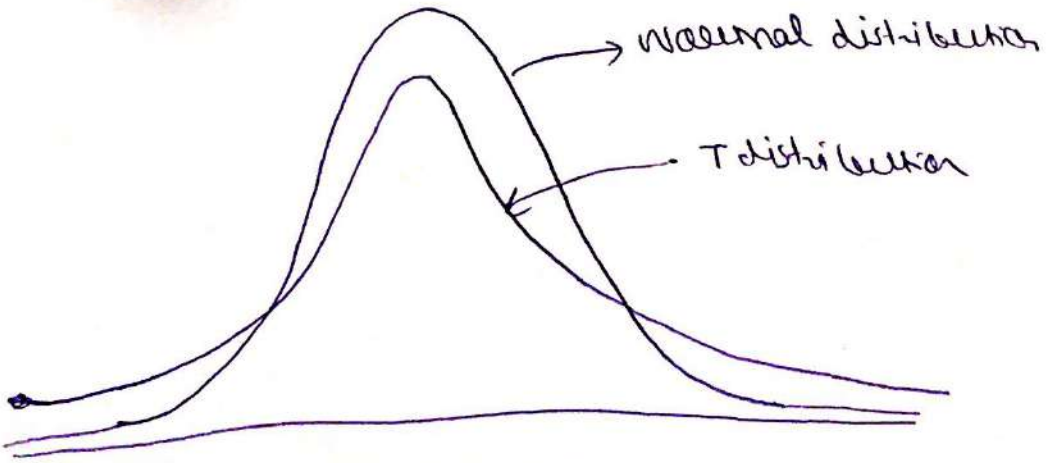
=T.DIST and =T.INV

P(X) = T.DIST (x, df, TRUE)

↳ we get area under curve left of X.

* degree of freedom is the parameter of t distribution

→ larger set of data would have greater degree of freedom



Now,

To find $z_{\alpha/2}$

$$z_{\alpha/2} = \text{NORM.INV}\left(\alpha/2, \underset{\text{mean}}{0}, \underset{\text{standard deviation}}{1}\right)$$

$$\text{margin of error} = |z_{\alpha/2}| * \frac{\sigma}{\sqrt{n}}$$

$$= \text{CONFIDENCE.NORM}(\alpha, \sigma, n)$$

→ when population standard deviation (σ) is not known

* we replace it by sample standard deviation.
* z statistics ($z_{\alpha/2}$) get replaced by t statistics ($t_{\alpha/2}$)

$$\bar{x} - |t_{\alpha/2}| \frac{s}{\sqrt{n}} < \mu < \bar{x} + |t_{\alpha/2}| * \frac{s}{\sqrt{n}}$$

here margin of error = CONFIDENCE.T(α, s, n)

(α: significance level; s: std; n: sample size.)

→ जब population standard deviation पता है तो हम z statistics and जब sample standard deviation पता है तो हम T statistics.

Confidence Interval for population proportion :-

Q A consultancy firm surveyed a randomly selected set of 210 CEO of fast growing small companies in US and Europe. only 51% of these executives had management succession plan in place, remaining did not have one.

Use this information to compute a 90% confidence interval to estimate the proportion of all 'fast growing small companies' that have management succession plan?

$$\hat{p} - |z_{\alpha/2}| \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < (p) < \hat{p} + |z_{\alpha/2}| \sqrt{\frac{p(1-p)}{n}}$$

margin of error.

actual proportion population which is unknown to us & we wish to access it with sample information we have.

Sample Size :-

Q How Big should our sample is?

A Quality control manager at battery manufacture wants to estimate the average number of defective batteries contained in a box shipped by a company? How many boxes does she needs to open to figure out average number of defective batteries contained in a box?

Given,

margin of error = $\pm$ 0.3 batteries

Confidence Interval = 95%

Population Std = 0.9

Ans:-

$$\text{margin of error} = 0.3 = Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$0.3 = \left| \text{NORM. INV } (0.05/2, 0, 1) \right| \times \frac{\sigma}{\sqrt{n}}$$

$$0.3 = 1.96 \times \frac{0.9}{\sqrt{n}}$$

$$\sqrt{n} = 5.88$$

$$n = 35$$

Sample size calculation when we have population proportion rather than mean

Q A pollster wanting make a prediction about a particular candidate.

How many voters should the pollster survey?

Pollster in a prediction wants to have margin of error +/- 3% with confidence level 95%?

Ans.

$$\text{margin of error} = |z_{\alpha/2}| * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$0.3 = |\text{NORM.INV}(0.05/2, 0, 1)| * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$\downarrow 1.96$

$$n = \frac{(1.96)^2}{(0.03)^2} * \hat{p}(1-\hat{p})$$

$\hat{p}$ is the sample proportion who would vote for that particular candidate.

Now,
we use a conservative estimate of $\hat{p}$.

$\hat{p}(1-\hat{p})$ — for this to be maximum $\hat{p} = 0.5$

Hypothesis testing and p values :—

Neurologist is testing the effect of drug on response time by injecting 100 rats with a unit dose of drug, subjecting each to neurological stimulus and recording its response time. The neurologist knows that the mean response time for rats not injected with drug is 1.2 seconds? mean of 100 injected rat response time is 1.04 seconds with sample s d = 0.5 seconds. Do you think that drug has effect on response time?

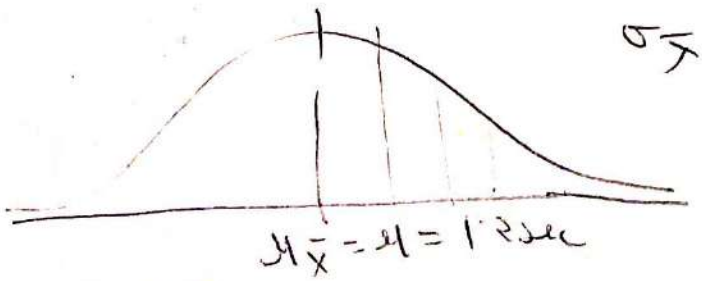
Ans:- Null Hypothesis H_0 :

↳ drug has no effect on response time

H_0 : drug has no effect $\Rightarrow \mu = 1.25$ sec when with drug

H_1 : Drug has an effect $\Rightarrow \mu \neq 1.2$ sec when drug is given

Let assume $H_0 \rightarrow$ true



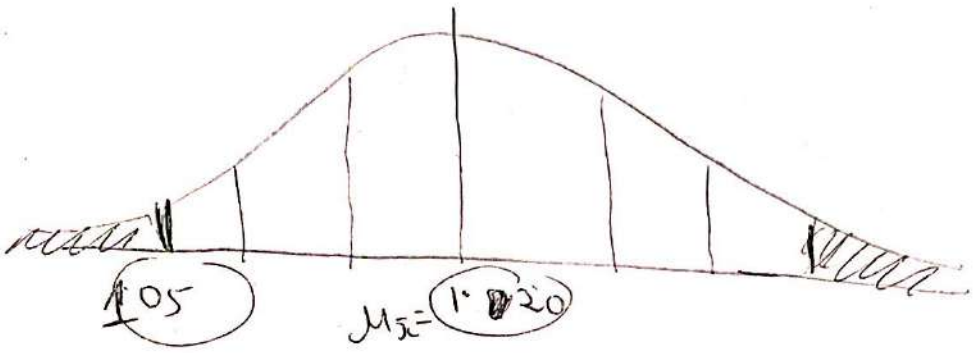
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{100}} \approx \frac{s}{\sqrt{100}} = \frac{0.5}{10}$$

$$\sigma_{\bar{x}} = 0.05$$

$$Z = \frac{1.2 - 1.05}{0.05}$$

$$\Rightarrow Z = 3$$

If mean



To find this area, $= 100 - 99.7$

$= 0.3\ \%$.

→ Hence if we assume that drug has no effect probability of getting the sample is only 0.3 %. ~~because~~ so null hypothesis is being rejected because if null hypothesis is true than their is only 0.3 % of chance of getting this result if null hypothesis is true.

$$\boxed{\text{p value} = 0.003}$$

→ Hence, drug has some effect

$n = 100 \cdot C = 0.99 \quad \alpha = 1 - 0.99$

$\alpha = 0.01$

Doctor believe that average teen sleeps on average no longer than 10 hours / day? A researcher believes that teens on average sleeps longer? write H_0 and H_1?

Ans: H_0: $\mu \leq 10$ hours / day ✓

H_1: ~~[illegible]~~

$\mu > 10$ hours / day ✓

School board claims that atleast 60% of student bring phone to school? Teacher believe this number is too high & randomly sample 25 students to test at level of significance of 0.02? write H_0 and H_a?

Ans: ~~[illegible]~~

$H_0 : p \geq .60$

$H_1 : p < .6$

$\alpha = 0.02$

$C = 0.98$

$= 98\%$

$n = 25$

Large samples

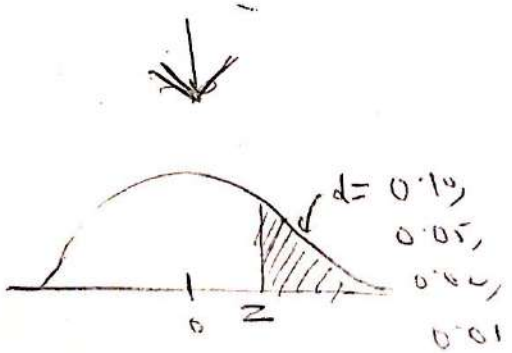
Hypothesis testing for mean & large samples

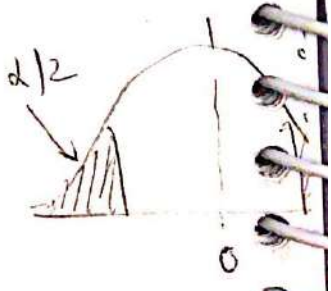
↳ Here we will take large sample as $n \geq 30$. If $n < 30$, then it is small sample & we will use T distribution there but, if it is large sample we will be using Normal distribution.

degree of freedom = n − 1

→ T distribution approaches Normal distribution when you have large number of samples.

C	α	one tail test	Two tail test
0.90	0.10	1.28 (z value)	∓ 1.645 (z value)
0.95	0.05	1.645	∓ 1.96
0.98	0.02	2.05	∓ 2.33
0.99	0.01	2.33	∓ 2.57





Hypothesis - Testing

In hypothesis testing, we turn a question of interest into hypotheses about the value of a parameter or parameters.

We create:

- A null hypothesis, denoted by H_0 (Sometimes referred to as the status quo hypothesis)
- An alternative hypothesis, denoted by H_a (Sometimes referred to as the research hypothesis)

It is the method of statistical inference. It is the an assumption that can be tested on a statement

We calculate an appropriate test statistic (based on sample data), and determine how much evidence there is against the null hypothesis.

If the evidence is strong enough (if it meets a certain *significance level*), we can reject the null hypothesis in favour of the alternative hypothesis.

Is the average weight of a certain type of candy bar different from the desired 58 grams?

$$H_0: \mu = 58$$

$$H_a: \mu \neq 58$$

- Is there significant evidence against the null hypothesis?
  (Two approaches: the rejection region approach, and the p-value approach.)
- What is an appropriate conclusion to the problem?

Z Test for one Mean :-

normal distribution

>help(pnorm)

>?pnorm

P(X <= 7)

>pnorm(q = 7, mean = 13, sd = 5, lower.tail = T(default))

0.158

Hypothesis Tests for the Population Mean μ

(When sampling from a normally distributed population)

> # P(X >= 85)

> pnorm(q = 85, mean = 75, sd = 5, lower.tail = F)

> 0.023

→ more N samples

Two scenarios:

Rare in practice (but a useful starting point)

- σ is known. Z test
- σ is not known. t test

  Much more common

→ less N samples

T-distribution :-

These can be used to find p values or critical values for constructing confidence intervals for statistics that follows t distribution.

>help(pt)

>?pt

t-statistic = 2.3, one sided.

we want to find $p(t > 2.3)$

```
> pt(q=2.3, df=25, lower.tail=F)
0.0150
```

Is there strong evidence that the population mean μ is different from some value that is of interest to us?

Is it different from some *hypothesized* value?

two sided p values

```
> pt(q=2.3, df=25, lower.tail=F) + pt(q=-2.3, d
0.03                                lower.tail=
```

Population mean — Hypothesized value

The null hypothesis is $H_0: \mu = \mu_0$ e.g. $H_0: \mu = 10$

We choose one of the possible alternatives:

- $H_a: \mu < \mu_0$ } One-sided alternatives (one-tailed tests)
- $H_a: \mu > \mu_0$ }
- $H_a: \mu \neq \mu_0$ ← Two-sided alternative (two-tailed test)

find t for 95% confidence

value of t with 2.5% in each tail.

```
> qt(p=0.025, df=25, lower.tail=T)
-2.06
```

2 sample T test and confidence interval

They are use to examine difference in mi
in the population. Also to find relationship betw
between variable and a categorical variable

The appropriate choice of alternative hypothesis depends on the problem at hand, and *should not be based on the current sample's data.*

You should not use the same data that suggests a hypothesis to test that hypothesis.

Suppose we have a simple random sample of n observations from a normally distributed population where σ is known.

> help(t.test)
> t test
> boxplot(lungcap ~ smoking)
Ho: mean lungcap of smokers = of no
Ha: ≠ (two sided)
> t.test(lungcap ~ smoke, mu=0, alt="two
conf=0.95,)

pvalue = 0.00039
confidence interval = (-1.3 -0.4)

To test H_0: $\mu = \mu_0$, we use the test statistic:

$$Z = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

where $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

If Ho is true, the Z test statistic will have the standard normal distribution

Example:
Suppose a supplier to a sushi restaurant claims their bluefin tuna contains no more than 0.40 ppm of mercury on average.

The owner fears the supplier's claim is incorrect, and the average mercury level is higher.

μ is the true mean mercury content of bluefin tuna from this supplier

We will test the null hypothesis:

H_0: $\mu = 0.40$

H_a: $\mu > 0.40$

In a random sample of 16 pieces of tuna from this supplier, there was an average of 0.74 ppm of mercury.

Does this yield strong evidence that the true mean mercury content is greater than 0.40 ppm?

(Suppose $\sigma = 0.08$ ppm)

H_0 $\mu \geq 0.40$

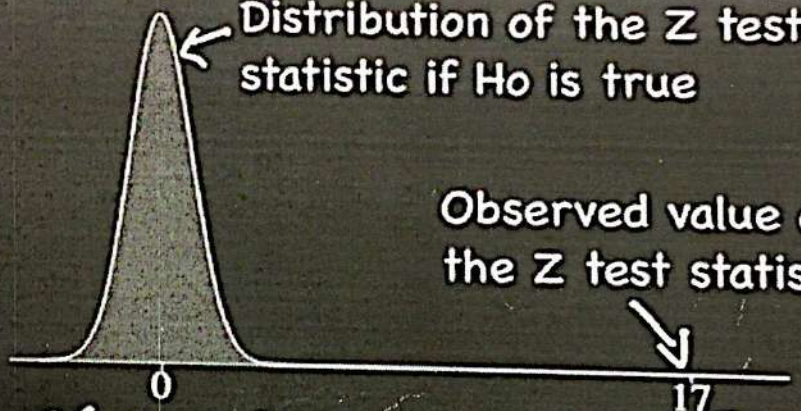
H_a < 0.40

$H_0: \mu = 0.40$, $H_a: \mu > 0.40$

$$Z = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}} = \frac{0.74 - 0.40}{0.02} = 17$$

$\bar{X} = 0.74$, $\sigma = 0.08$, $n = 16$.

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{0.08}{\sqrt{16}} = 0.02$$



$P(Z \geq 17) = 4.1 \times 10^{-65}$

There is extremely strong evidence against Ho

Here the evidence against the null hypothesis is overwhelming.

(We would almost never see what was observed if the null hypothesis were true.)

But how far out does the test statistic need to be before we can say the evidence is *significant*?

Before we can *reject* the null hypothesis in favour of the alternative?

To judge whether the evidence against the null hypothesis is significant, we use one of two approaches:

- The rejection region approach
- The p-value approach

Rejection Region Approach :-

Z Tests for the Population Mean μ

The Rejection Region Approach

Suppose we are sampling from a normally distributed population, and σ is known.

To test H_0: $\mu = \mu_0$, we use the test statistic:

$$Z = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}} \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

If Ho is true, the Z test statistic will have the standard normal distribution

The rejection region approach:

- Choose a value for α, the significance level of the test.

  (α is the probability of rejecting the null hypothesis if it is true.)

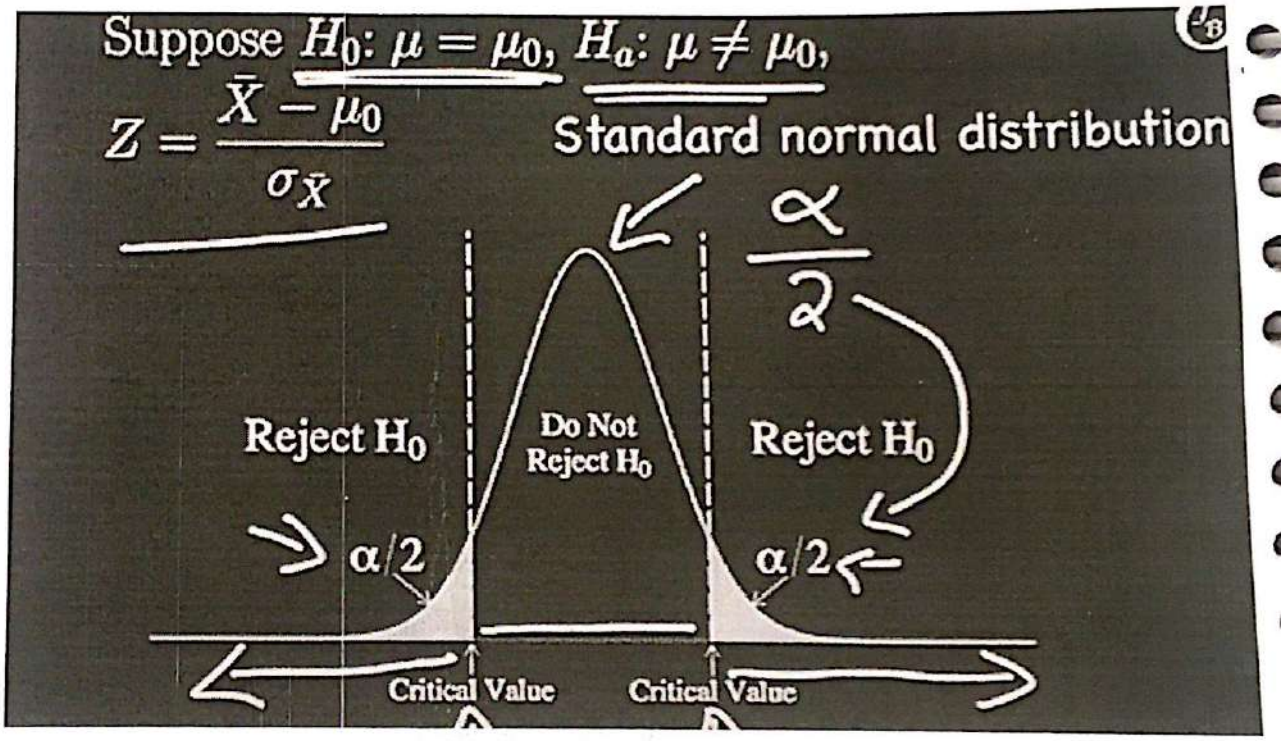
People often choose $\alpha = 0.05$

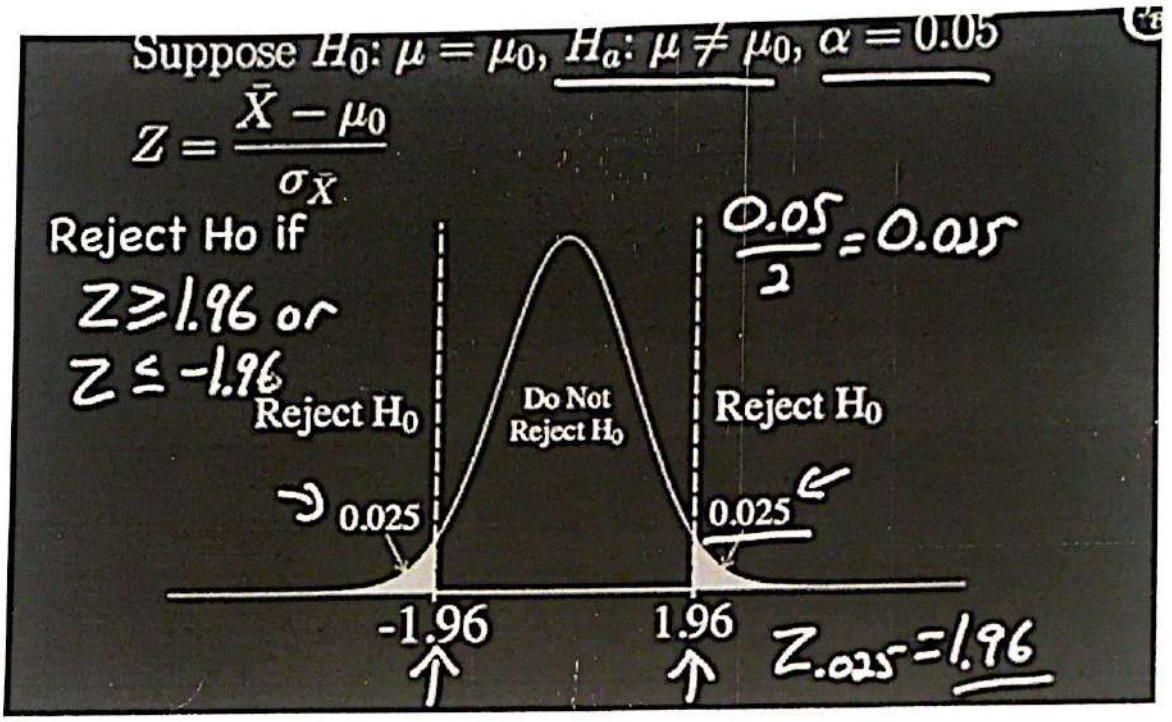
- Find the appropriate rejection region.
- Reject the null hypothesis if the test statistic falls in the rejection region.

Suppose $H_0: \mu = \mu_0$, $H_a: \mu \neq \mu_0$,

$$Z = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}}$$

Standard normal distribution



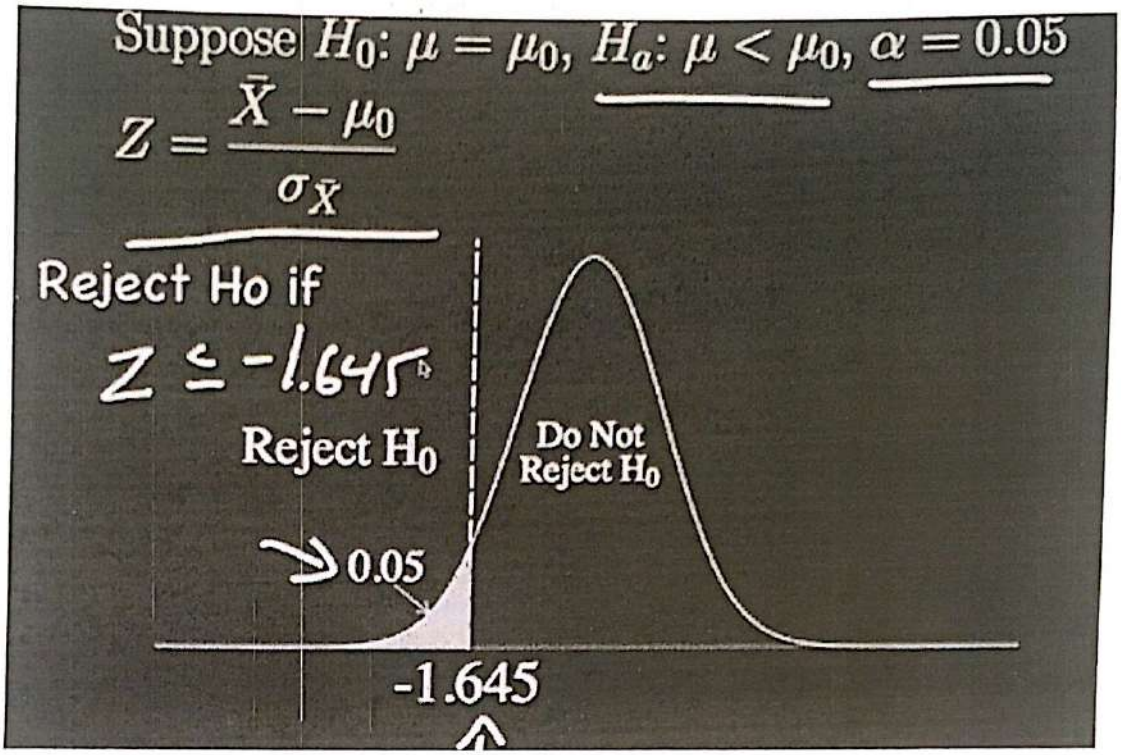


Suppose $H_0: \mu = \mu_0$, $H_a: \mu < \mu_0$, $\alpha = 0.05$

$$Z = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}}$$

Reject Ho if

$Z \leq -1.645$

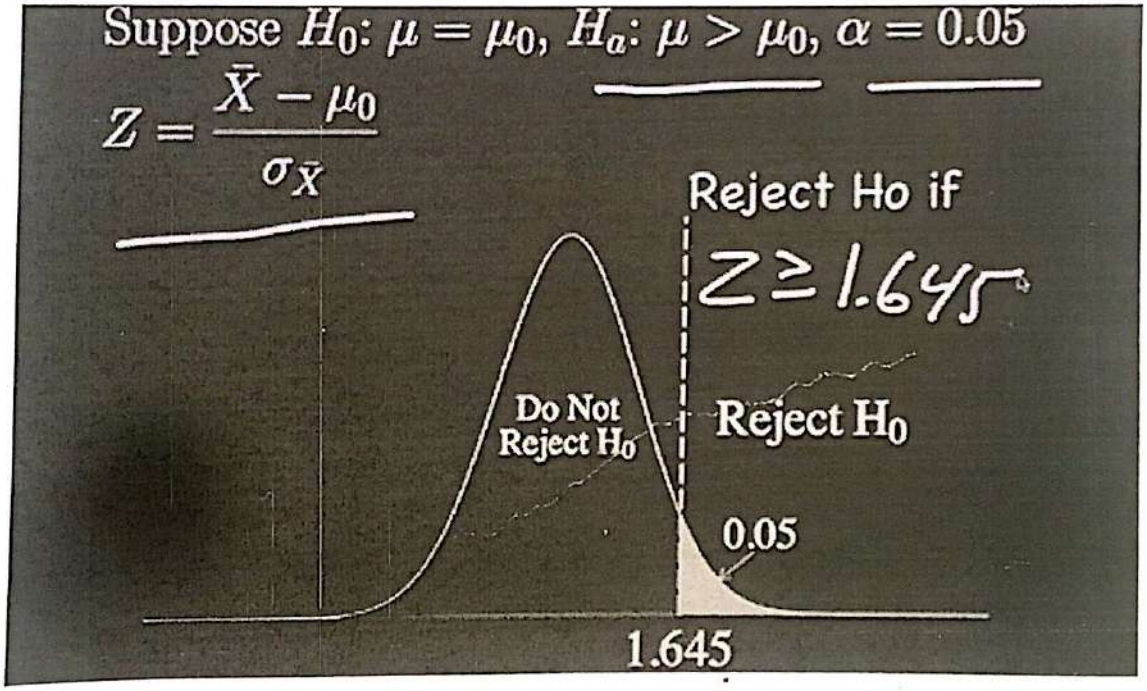


Suppose $H_0: \mu = \mu_0$, $H_a: \mu > \mu_0$, $\alpha = 0.05$

$$Z = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}}$$

Reject Ho if

$Z \geq 1.645$

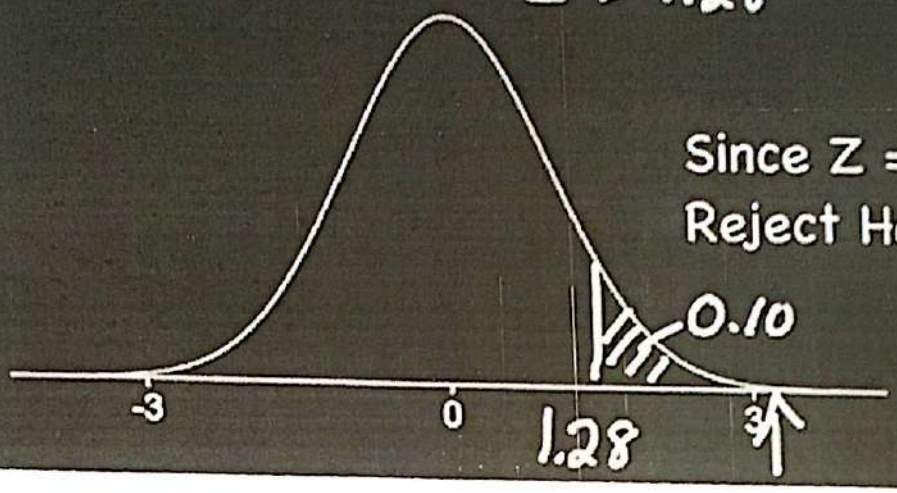


Suppose $H_0: \mu = \mu_0$, $H_a: \mu > \mu_0$, $\alpha = 0.10$

$Z = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}} = 3.1.$

Reject Ho if

$Z \geq 1.28$



Since Z = 3.1, Reject Ho

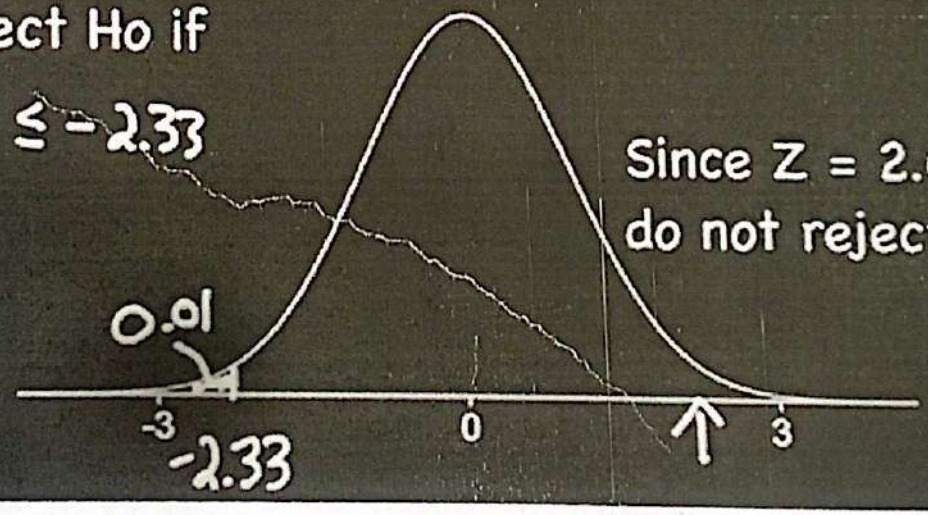
0.10

Suppose $H_0: \mu = \mu_0$, $H_a: \mu < \mu_0$, $\alpha = 0.01$

$Z = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}} = 2.6.$

Reject Ho if

$Z \leq -2.33$



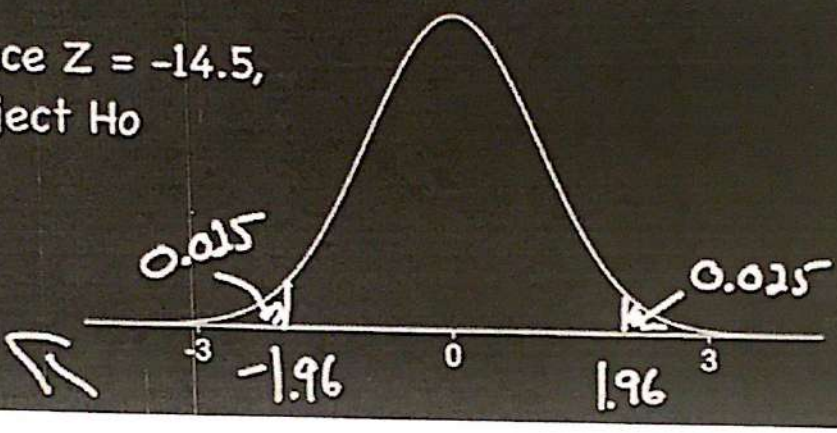
Since Z = 2.6, do not reject Ho

0.01

Suppose H_0: $\mu = \mu_0$, H_a: $\mu \neq \mu_0$, $\alpha = 0.05$

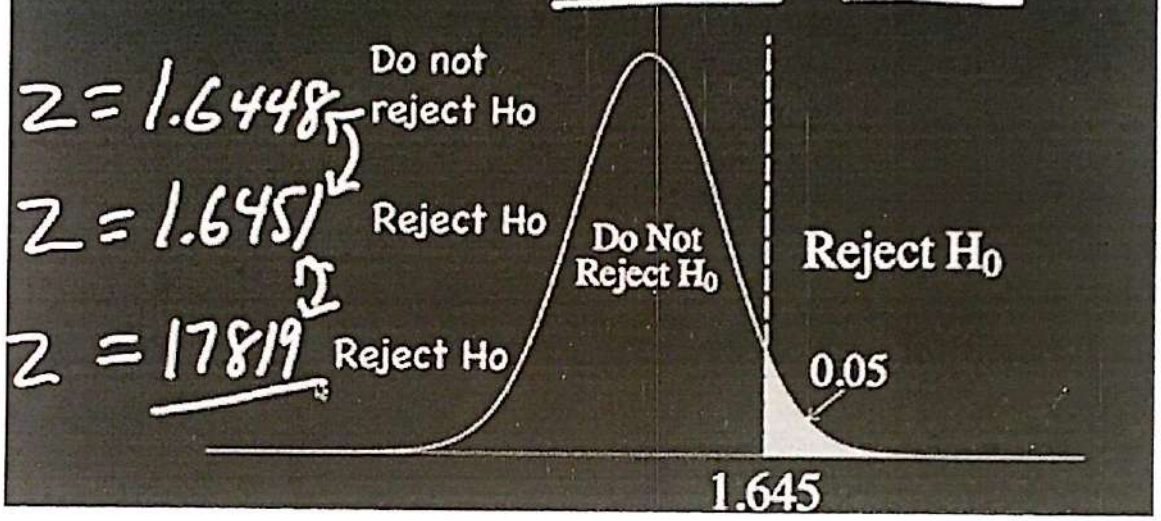
$$Z = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}} = -14.5.$$

Since Z = -14.5,
Reject Ho



But in some ways the rejection region approach is a little silly.

Suppose H_0: $\mu = \mu_0$, H_a: $\mu > \mu_0$, $\alpha = 0.05$



Many people prefer the *p-value* approach.

The *p*-value is a measure of the strength of the evidence against the null hypothesis.

P Value Approach

Z Tests for the Population Mean μ

The p-value

The definition:

The p-value is the probability of getting the observed value of the test statistic, or a value with even greater evidence against H_0, *if the null hypothesis is true.*

Example:

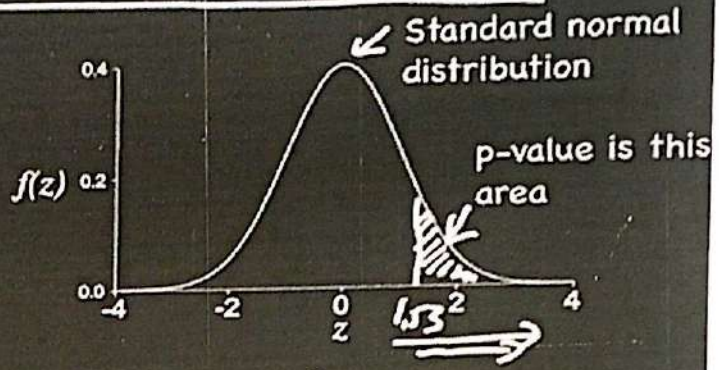
$H_0: \mu = \mu_0$

$H_a: \mu > \mu_0$

$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$

Suppose $z = 1.53$

Distribution of Z if H_0 is true:



p-value $= P(Z \geq 1.53)$

$= 0.063$

Example:

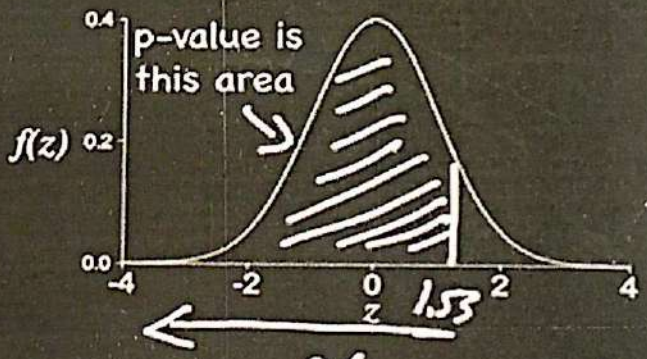
$H_0: \mu = \mu_0$

$H_a: \mu < \mu_0$

$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$

$Z = 1.53$

Distribution of Z if H_0 is true:



p-value $= P(Z \leq 1.53)$

$= 0.937$

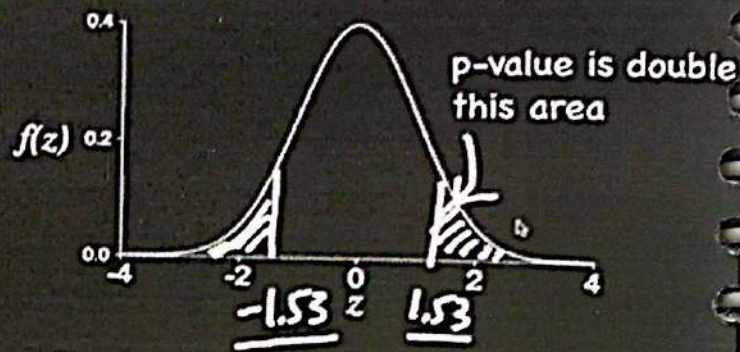
Example:

Distribution of Z if H_0 is true:

H_0: $\mu = \mu_0$

H_a: $\mu \neq \mu_0$ ←

$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$



$z = 1.53$

p-value $= P(Z \geq 1.53) + P(Z \leq -1.53)$

$= 2P(Z \geq |1.53|)$

The smaller the p-value, the greater the evidence against the null hypothesis.

If we have a given significance level α, then:

Reject H_0 if p-value $\leq \alpha$

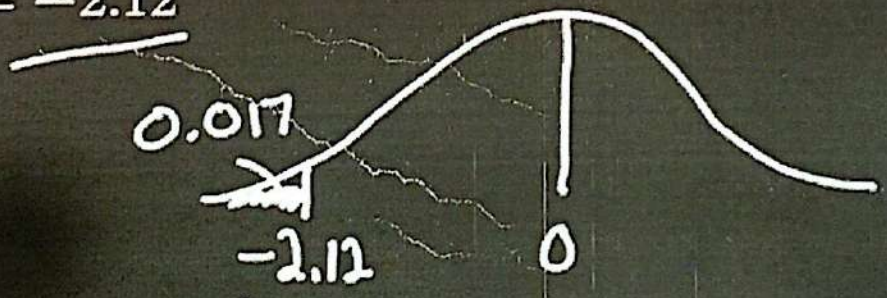
Cutoff level for significance

(If p-value $\leq \alpha$, the evidence against H_0 is significant at the α level of significance.)

Example:

Suppose H_0: $\mu = 10$, H_a: $\mu \neq 10$, $\alpha = 0.05$,

$Z = -2.12$



p-value = 2 × 0.017 = 0.034

Here we reject our null hypothesis because our p value is less than alpha (level of significance) .

If we do not have a given significance level, then it is not as cut-and-dried.

Example of Z test .

A producer of cereal is producing boxes of cereal with a stated weight of 750 grams.

It is known from a large body of past experience that the standard deviation of the weights in their filling process is approximately 16 grams.

In order to ensure that not many boxes are underfilled, the producer sets the mean fill amount at 780 grams.

As part of the quality control process, they periodically draw a random sample of 25 boxes, measure the weights, and test the null hypothesis that the mean fill amount is 780. $H_0: \mu = 780$ $H_a: \mu \neq 780$ $\alpha = 0.05$

For a sample of 25 boxes, $\bar{X} = 776$.

Does this give strong evidence the true mean weight differs from 780 grams?

$H_0: \mu = 780$, $H_a: \mu \neq 780$, $\alpha = 0.05$

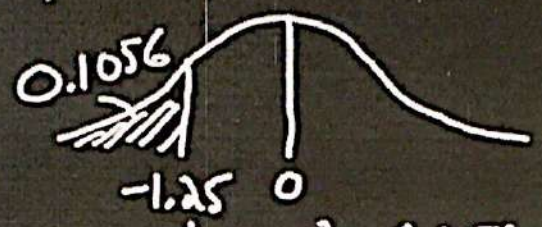
$\bar{X} = 776$, $\sigma = 16$, $n = 25$.

$$Z = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

$H_0: \mu = 780,\ H_a: \mu \neq 780,\ \alpha = 0.05$

$\bar{X} = 776,\ \sigma = 16,\ n = 25.$

$$Z = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{776 - 780}{16/\sqrt{25}} = -1.25$$



The evidence against Ho is not significant at the 0.05 level of significance

p-value $= 2 \times 0.1056 = 0.211$

There is little or no evidence (p-value = 0.211) that the true mean weight of the filling process differs from 780 grams.

Note: This does not imply there is strong evidence the mean is 780!

One sided Test vs Two sided Test

A company claims that their product contains no more than 2 grams of saturated fat on average.

You intend to test whether there is strong evidence the mean saturated fat content is greater than their claim.

$H_0: \mu = 2 \quad H_a: \mu > 2$

A production process creates silicon wafers. The desired thickness is 725 μm.

You intend to test whether there is strong evidence the mean thickness differs from this amount.

$H_0: \mu = 725 \quad H_a: \mu \neq 725$

The choice between a one-sided or two-sided alternative hypothesis can be subject to debate.

It should not be based on the current sample's data!

We might strongly suspect that a newly developed drug decreases blood pressure.

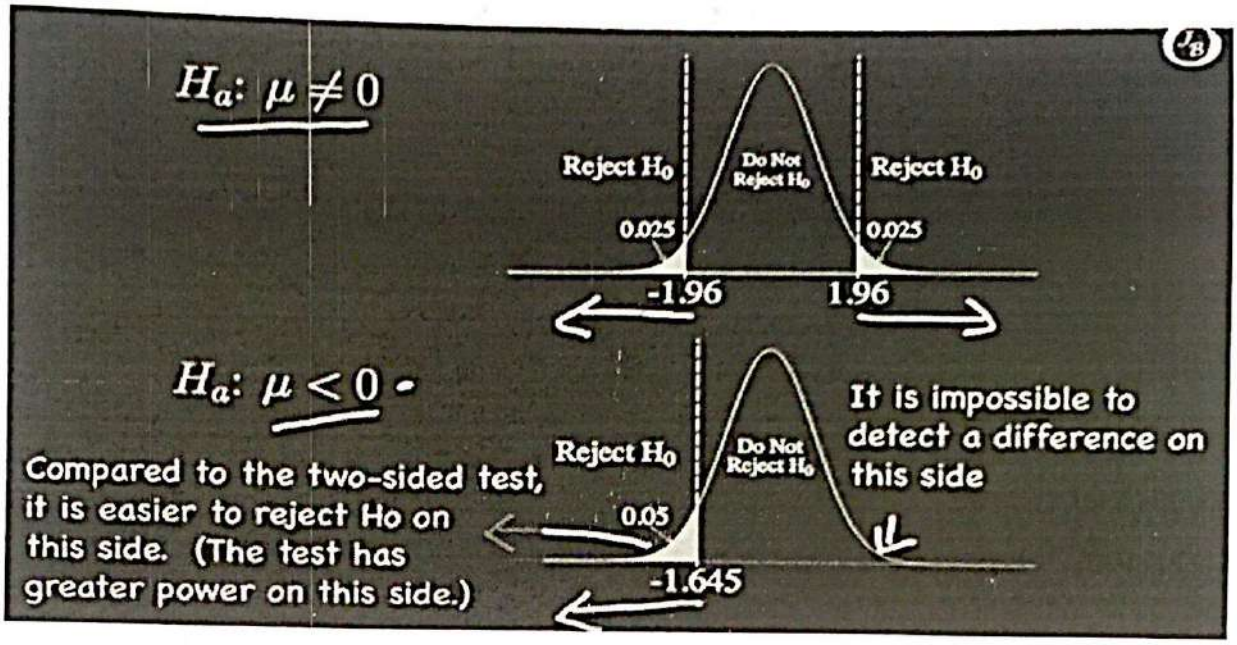
We may wish to test: H_0: $\mu = 0$ (Population mean change in blood pressure)

But which is the appropriate alternative?

H_a: $\mu \neq 0$

H_a: $\mu < 0$

Suppose we are using a Z test at $\alpha = 0.05$.



If we choose the alternative hypothesis based on the direction observed in the sample, then the reported p-value will be half of what it should be.

(We will sometimes be reporting results as significant when they are not.)

Results different from what one expects can often be the most interesting results.

In two-sided tests we are still almost always interested in the *direction* of the difference.

Example: You are investigating possible differences in five-year survival rates for two types of treatment.

$$H_0: p_1 = p_2, \quad H_a: p_1 \neq p_2$$

p_1: Five-year survival rate for Treatment 1

p_2: Five-year survival rate for Treatment 2

In tests with two-sided alternatives, the null hypothesis is almost always wrong!

If in the end we simply reject it, without giving an indication of the direction of the difference, we just wasted our time.

In the end, the choice between one or two-sided alternative hypothesis does not need to matter much:

Report the p-value (usually two-sided), and let the reader make up their own mind.

Type 1 Errors and Type 2 Errors

Hypothesis Testing

Type I errors, Type II errors, and the Power of the Test

A Type I error is rejecting H_0 when, in reality, it is true.

A Type II error is failing to reject H_0 when, in reality, it is false.

Suppose we test:

$$H_0: \mu = 10$$
$$H_a: \mu > 10$$

and we reject H_0 at $\alpha = 0.05$.

One of two things occurred:

- The null hypothesis is false and we made the correct decision.
- The null hypothesis is true, and we made a Type I error.

Suppose we test:

$$H_0: \mu = 10$$
$$H_a: \mu > 10$$

and we do not reject H_0 at $\alpha = 0.05$.

One of two things occurred:

- The null hypothesis is true and we made the correct decision.
- The null hypothesis is false, and we made a Type II error.

Possible outcomes of a hypothesis test:

Unknown

		Underlying reality	
		H_0 is false	H_0 is true
Conclusion from test	Reject H_0	Correct decision	Type I error
	Do not reject H_0	Type II error	Correct decision

Known

Consider a criminal trial. We test the hypotheses:

H_0: The defendant did not commit the crime.

H_a: The defendant committed the crime.

Type I error: Convicting a person who, in reality, did not commit the crime.

Type II error: Acquitting a person who, in reality, committed the crime.

Very Imp :-

Reducing probability of type I and type II e
→ probability of type I error is set by
choice of α / typically

→ probability of type II error can be
by taking large sample size

The probability of a Type I error, given H_0 is true, is called the *significance level* of the test (α).

$$P(\text{Type I error} | H_0 \text{ is true}) = \alpha$$

The probability of a Type II error is represented by β.

The value of β depends on a number of factors, including the choice of α, the sample size, and the true value of the parameter.

The power of a test is the probability of rejecting the null hypothesis, given it is false.

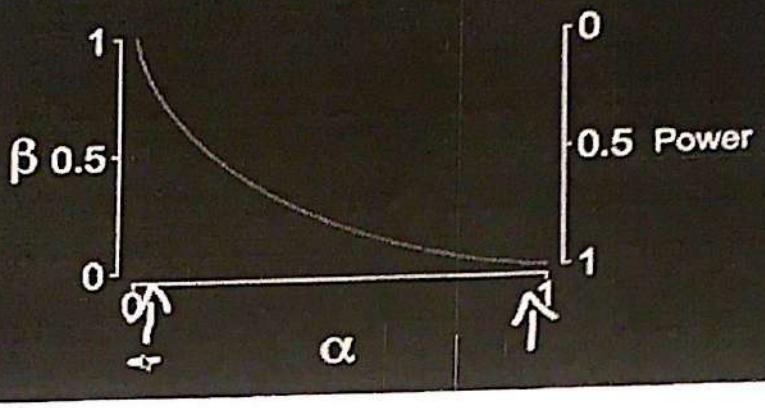
$$\text{Power} = 1 - P(\text{Type II error}) = 1 - \beta$$

The power of a test depends on several factors, including the choice of α, the sample size, and the true value of the parameter.

If we choose a very small value of α, we will be making it very difficult to reject the null hypothesis. (Type II errors will be common.)

If we choose a larger value of α, Type II errors will be less common.

The relationship between α and β for a test of H_0: $\mu = 0$:



Statistical Significance vs Practical Significance .

Hypothesis testing tests for *statistical* significance.

Statistical significance means the effect observed in the sample was unlikely to have occurred due to chance alone.

(It would be very unlikely to see what was observed in the sample if the null hypothesis is true.)

Suppose a call centre claims their average wait time is 30 seconds. We decide to test:

$$H_0: \mu = 30$$
$$H_a: \mu > 30$$

We find $\bar{X} = 30.6$, and a p-value of 0.002.

Statistical significance is strongly related to sample size:

If the sample size is large enough, even tiny differences from the hypothesized value will be found statistically significant.

Relationship b/w Hypothesis Test and Confidence Intervals .

The Relationship Between Hypothesis Tests and Confidence Intervals

Suppose we find a $(1-\alpha)100\%$ confidence interval for μ using:

(A two-sided confidence interval) $\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

And we wish to carry out a test of:

H_0: $\mu = \mu_0$ ←

H_a: $\mu \neq \mu_0$ ←

with a significance level of α.

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \leftarrow$$

The confidence interval will be made up of all values of μ_0 for which we would not reject the null hypothesis.

$\alpha = 0.10$

Suppose a 90% confidence interval for μ is found to be $(13.1, 26.4)$. The hypothesized value of 25 falls within the 90% interval

What would be the results of a test of H_0: $\mu = 25$ against H_a: $\mu \neq 25$ at $\alpha = 0.10$?

The null hypothesis would not be rejected at a significance level of 0.10

The p-value of the test would be greater than 0.10

$\alpha = 0.05$

Suppose a 95% confidence interval for μ is found to be $(48.2, 87.6)$. ← The hypothesized value of 40 falls outside the 95% interva

What would be the results of a test of H_0: $\mu = 40$ against H_a: $\mu \neq 40$ at $\alpha = 0.05$?

The null hypothesis would be rejected at a significance level of 0.05

The p-value of the test would be less than 0.05

Why is this the case?

We would reject H_0 at $\alpha = 0.05$ if: ← The Z test statistic

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \leq -1.96 \text{ or } \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \geq 1.96$$

Isolating μ_0, we would reject H_0 if:

$$\mu_0 \geq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}} \text{ or } \mu_0 \leq \bar{X} - 1.96\frac{\sigma}{\sqrt{n}}$$

The upper bound of the 95% confidence interval

The lower bound of the 95% confidence interval

If the hypothesized value falls outside the confidence interval we will reject the NULL hypothesis . If it falls within confidence interval then we will not reject NULL Hypothesis .

Correlation Analysis :-

$x, y \rightarrow$ degree of relationship

↓

coefficient of correlation ('ρ')

Types of correlation :-

(i) +ve correlation — $x \uparrow \Rightarrow y \uparrow$ or $x \downarrow \Rightarrow y \downarrow$

(ii) -ve correlation — $x \downarrow \Rightarrow y \downarrow$ or $x \uparrow \Rightarrow y \downarrow$

(iii) no correlation — $x \uparrow \Rightarrow y \times$ or $x \downarrow \Rightarrow y \times$ $\rho = 0$

(iv) Perfect correlation — $\rho = 1$ (perfect +ve), $\rho = -1$ (perfect -ve)

methods to find correlation :-

(i) coefficient of correlation (ρ)

(ii) Rank correlation

NOTE :- $[-1 \leq \rho \leq +1]$

coefficient of correlation :- (ρ)

(1) $$\rho = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma (x-\bar{x})^2}\sqrt{\Sigma (y-\bar{y})^2}}$$

(2) $$\rho = \frac{cov(x,y)}{\sqrt{var(x)}\sqrt{var(y)}}$$

where $$cov(x,y) = \frac{1}{n}\Sigma (x - \bar{x})(y - \bar{y})$$

↙ Means how two variables vary together.

$$var(x) = \frac{\Sigma (x - \bar{x})^2}{n}$$

$$var(y) = \frac{\Sigma (y - \bar{y})^2}{n}$$

Regression Analysis :-

$$y - \bar{y} = \rho \frac{\sigma_y}{\sigma_x}(x - \bar{x}) \quad \leftarrow m_1$$

$$x - \bar{x} = \rho \frac{\sigma_x}{\sigma_y}(y - \bar{y}) \quad \leftarrow m_2$$

$$m_1 = \rho \frac{\sigma_y}{\sigma_x}, \quad m_2 = \rho \frac{\sigma_x}{\sigma_y}$$

(1) Regression lines always intersect at point $(\bar{x}, \bar{y})$

(2) If two regression lines are given and θ be the acute angle b/w them then

$$\tan\theta = \left| \frac{1-\rho^2}{\rho} \right| \left| \frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2} \right|$$

Note:-

→ If $\rho = 1 \Rightarrow \tan\theta = 0 \Rightarrow$ दोनों line coincide करती है

→ If $\rho = 0 \Rightarrow \tan\theta = \infty \Rightarrow \theta = 90$

when there is no correlation b/w the lines then they are perpendicular

(3) Coefficient of correlation is Geometric mean of Regression Coefficient.

$$\rho = \sqrt{m_{xy} \times m_{yx}}$$

Correlatio

Many times, we ha

Such as Temperatu

If Temp

Correla

To see how the

Temperat

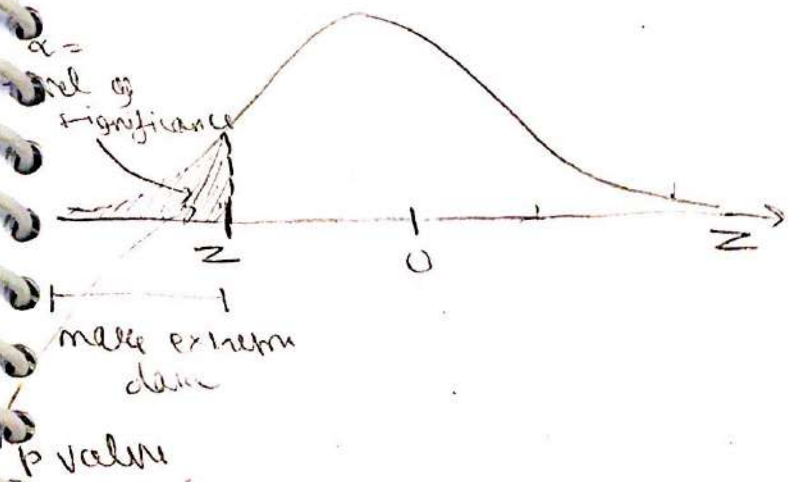
17

16

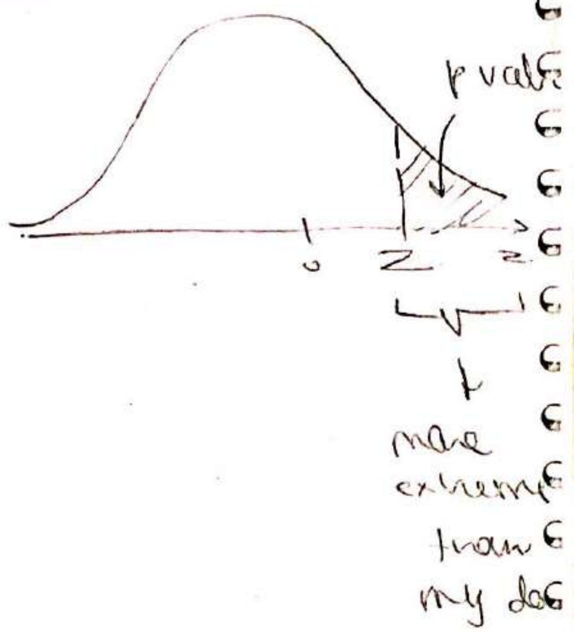
Scanned with CamScanner

p value :- probability of obtaining a sample "more extreme" than the ones observed in your data, assuming null hypothesis is true.

Left tail test :-



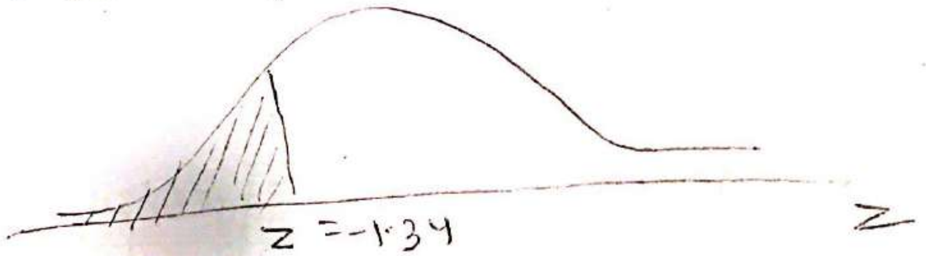
Right tail test :-



eg. - left tail test :-

$H_0 = \mu \geq 0.15$ From data : $Z = -1.34$

$H_1 = \mu < 0.15$



$P(Z < 1.34) = 0.0901$

$p = 0.0901$

→ z की value देखो। then देखो area under the curve left side पता करो। calculate ie the p value

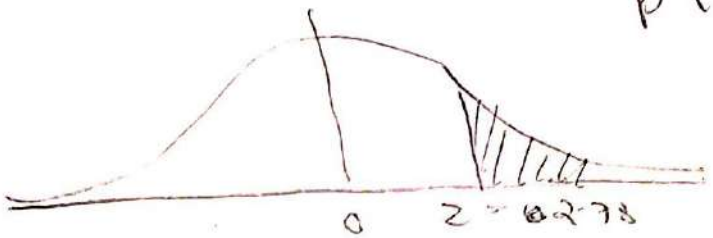
Right tail test eg :-

$H_0 : \mu \leq 0.43$

$H_1 : \mu > 0.43$

from data

$z = 2.78$



$p(z > 2.78)$

$= (1 - 0.9973)$

$= 0.0027$

p value = 0.0027

Two tail test :-

$H_0 : \mu = 0$

$H_1 : \mu \neq 0$

↳ 2 tail test

If we have two means then,

$H_0 : \mu_1 - \mu_2 = 0$

$H_1 : \mu_1 - \mu_2 \neq 0$ ↳ 2 tail test

Here test statistic :-

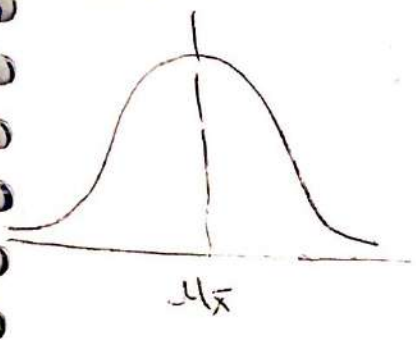
$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

↓

z score involving two means

Z statistics V/s T statistics →

Z statistic



$$Z = \frac{\bar{x} - \mu_{\bar{x}}}{\frac{\sigma}{\sqrt{n}}}$$

$$= Z = \frac{\bar{x} - \mu_{\bar{x}}}{\frac{s}{\sqrt{n}}} \quad [\text{if } n > 30]$$

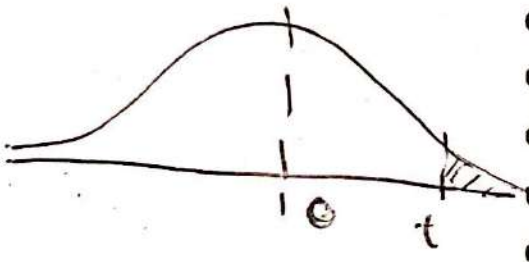
Normal distribution curve

T statistic

T distribution.

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

[if s is small]



Type - I error / Type II errors :-

Q) School district A states that its high school have 85% passage rates on high school exam. A new school was recently opened in the district, and it was found that sample of 150 students had passing rate of 88%

with standard deviation 4.1, does school this new school have passage rate than rest of school district A?

Null hypothesis:-

H_0 ; $\mu = 85\%$

Alternative hypothesis:

H_1 ; $\mu \neq 85\%$

$\alpha = 0.05 - 5\%$

→ If statistic we calculate (for example $\bar{x}$) is within 95% of range, we will conclude that what are testing (usually the mean) is right where we expect it to be, so we will retain the null hypothesis

→ If statistic we calculate is outside of that range, we will conclude that what we are testing is not where we expect it to be. So, we reject H_0 & accept H_1.

→ we haven't measured the entire. Shaal we only measured sample of students. The decision we make may or may not accurately reflect reality.

Outcome 1

We reject Null Hypothesis when in reality it is false → GOOD

Outcome 2

We reject Null hypothesis when in reality, it is true → Type I error.

Outcome 3

We retain Null hypothesis when in reality, it is false → Type II error.

Outcome 4 :-

We retain Null hypothesis when in reality, it is true → (good).

Small sample hypothesis test :—

A mean emission of all engine of a new design needs to below 20 ppm of design to meet new emission requirements. = 10 engines all manufactured for testing purpose & emission level of each is determined.

emission data is :-

15.6 16.2 22.5 20.5 16.4 19.4 16.6
17.9 12.7 13.9

does the data supply sufficient evidence to conclude that this type of engine meets new standard? Assume we are willing to risk a type - I error with $\alpha = 0.01$?

$\bar{x} = 17.17$

$s = 2.98$

$H_0 : \mu \geq 20$

$H_a : \mu < 20$

$H_0 \leq 20$
$H_a > 20$

$H_0 : \mu \geq 20$ ppm
$H_1 : \mu < 20$ ppm

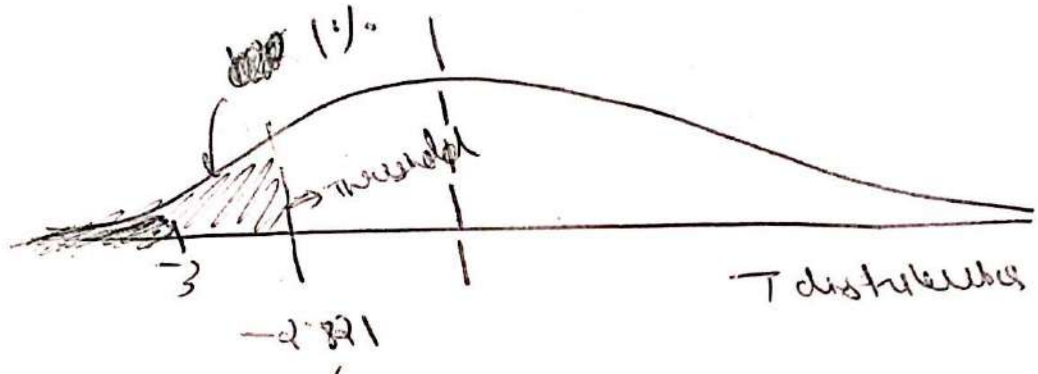
Assume H_0 is true.

We will reject H_0 if $P(\bar{x} = 17.17 \mid H_0 \text{ true}) < 1\%$

We will going to follow T statistics.

i.e. $n < 30$

$$t = \frac{17.17 - 20}{\frac{2.98}{\sqrt{10}}} = -3.00$$



ये कैसे आया → (p = 0.01)

as given

Hence p = 1% के लिए T distribution का graph देखोगे तो -2.821 आएगा |

जो calculate करा t that is = -3

t = -3 from our calculation

Its probability is less than 1% as -3 तो -2.8 से भी कम है, hence we reject Null hypothesis

$H_1 \rightarrow$ ✓

$$n = \frac{1.96^2}{0.03^2} \times 0.5(1-0.5)$$

$$n = 1068$$

Hypothesis test :-

→ All hypothesis follows a basic logic:

① you made an assumption.

③ If your data contradicts the assumption than you can conclude that assumption must be wrong.

→ It is the scientific tool for decision making. It is widely used procedure to test variety of claims.

Eg :- manufacturer claims that average bottle contains 200 ml of water. Now you came you tested 10 bottles and average comes out to be 170 ml of water. Will you reject ? Probably yes? Suppose average comes out to be 205 ml, probably you won't have rejected the ~~hypothesis~~ claim.

Hypothesis testing allows us to tell that till where we can reject ~~the~~ our claim. you can do that by your gut feeling also, but hypothesis testing is the scientific & statistical way of doing that.

NOTE :- normal

Any distribution can be turned to standard normal distribution by doing this:

$$\bar{x} \sim \text{Normal}\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \Rightarrow \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim \text{Normal}(0,1)$$

$\sigma/\sqrt{n}$ → Z Statistics

→ Generally we will be finding T statistics in our problems since we will be having sample data.

$$t_{\text{cutoff}} = \pm \left| T.INV(\alpha/2, n-1) \right|$$

↳ excel formula

NOTE :-

Null hypotheses cannot have

$$\neq \quad > \quad <$$

It can only have :-

$$= \quad \geq \quad \leq$$

↓ two tailed test

↓ one tail LHS

↓ one tail, RHS

We wish to test a claim that average age of men MBA students across various MBA programs in US is greater than 28 years. For this we collect data on average ages of men MBA students across sample of 40 MBA programs in US.

Step 1 formulate hypothesis

Null hypothesis H_0 : $\mu > 28$ ✗ as H_0 cannot have >

H_A : $\mu \leq 28$

hence

H_0 : $\mu \leq 28$ ✓

H_a : $\mu > 28$

Step 2 calculate t statistic

$$t_{stat} = \frac{\bar{x} - \mu}{s/\sqrt{n}} = 0.22$$

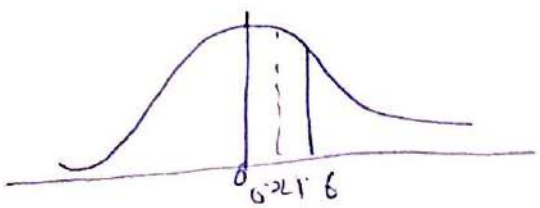
Step 3 :- cutoff value for t stat

(single tail test on RHS)

$= |T.INV(\alpha, 40-1)| = 1.6$

$\alpha = 0.05$

Step 4 :- do not reject null hypothesis.



Hypothesis test for population proportion

> Where we interest in population proportion rather than population mean.

Q University introduces a new lunch facility on campus on a trial basis. University operates lunch facility for few months and then decides to survey student body. Based on survey, university would make this facility a permanent fixture or do away with it. Specifically, if more than 70% students approve then facility would be made permanent else it would shut down.

→ University conducts a survey with 750 random selected students on campus and finds that 510 of these students (or 68% of sampled student) approves new facility & remaining 240 student or 32% do not approve it.

→ Based on criteria should university should make this permanent or not!

Ans: Step1 formulate hypothesis

$H_0 : p > 0.70$

$H_a : p_a < 0.70$

Step2 :-

Calculate Z statistics

* here in this case we always use Z statistics not t statics in population proportion case

$$Z \text{ Statistic} = \frac{\bar{p} - p_{hypothesis}}{\sqrt{\frac{p_{hy}(1-p_{hy})}{n}}}$$

($\bar{p} \rightarrow 0.68$, $p_{hypothesis} \rightarrow 0.7$, $n \rightarrow$ 750 students)

$$Z_{stat} = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = -1.1952$$

Step3 :- calculate for Z statistics.

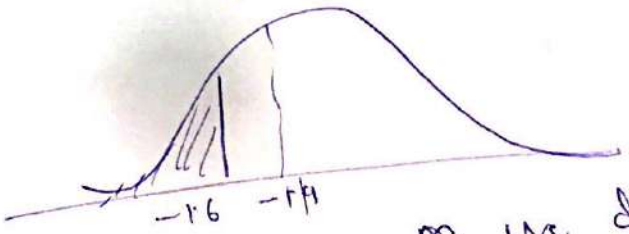
$$Z_{critical} = -\left| NORM.INV(\alpha, 0, 1) \right|$$

($\alpha \rightarrow 0.05$, 0 → mean, 1 → σ)

$$= -1.6$$

Rejection region should be on LHS.

Step4 :-



→ hence we do not reject Null hypothesis

NOTE Z test for comparing two proportions :

$$Z = \frac{(p_1 - p_2) - 0.7_{hyp}}{\sqrt{\bar{p}(1-\bar{p})\left(1/n_1 + 1/n_2\right)}}$$

Difference In means hypothesis test :-

When we are comparing population means across two different populations

→ we have athletes data, and claim was that average height of men aged 18 years to 45 years across the world was 173 cm.

Here hypothesis test can be used to test this claim for population of men athletes at olympics

→ Step 1 :-

Formulate hypothesis

$H_0 : \mu_{Height} > 173\ cm$

$H_a : \mu_{height} \leq 173\ cm$ ✗

but we cannot use that inequality here so we flip it

$H_0 : \mu_{height} \leq 173\ cm$ ✓

$H_a : \mu_{height} > 173\ cm$

here rejection region is on RHS

Step 2 :- calculate + statistic

$$t_{star} = \frac{\bar{x} - \mu_{height}}{s/\sqrt{n}}$$

Step 3 :- cut off for t statistic

$$t_{cutoff} = +\left|T.INV(\alpha, n-1)\right|$$

$$= 30.03$$

Step 4 :- Reject NULL Hypothesis.

→ final conclusion based on our data evidence we cannot claim reject the claim that male athletes at olympics are taller than 173 cm

→ same study also claimed that the difference in average heights across men & woman from various walks of life around world was 12.5 cm.

Now, to find

$(\mu_{men} - \mu_{women})$

↓

Hypothesis test in difference in means

Claim to be tested

Difference between the population mean height of men and woman athletes is 12.5 cm.

Step 1 :- Formulate hypothesis

H_0 : $\mu_{men} - \mu_{women} = 12.5$ cm

H_a : $\mu_{men} - \mu_{women} \neq 12.5$ cm

two tailed test

Step 2 :- calculate t statistic

* For calculation of t statistics we need to see the nature of variation or variance in two populations.
* Is variance more similar or more dissimilar
* need to assum either equal variance or unequal variance across two population.

→ In most applications conclusions do not change based on 'equal' or 'unequal' variance assumption.

we can do this using excel in excel we can use this when we are doing difference in mean. this formula is not applicable for single mean.

go to data, → data analysis → T test assuming equal variance = OK

Variable 1 range = men data range B1:B

Variable 2 range = C1:

Mean difference = 12.5

we will get all results

equal variance assumption

$$t_{stat} = \frac{\bar{x}_1 - \bar{x}_2 - \mu}{\sqrt{\left(\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

Unequal variance assumption

$$t_{stat} = \frac{\bar{x}_1 - \bar{x}_2 - \mu}{\sqrt{\left(\frac{s_1^2 + s_2^2}{n_1+n_2}\right)}}$$

Paired T test :-

→ used to test claims around difference in two population means

Q * what is difference in paired Test & normal difference in means hypothesis test.

we have conduct the training and we have to check that training is effective or not. we have taken 50 students and we need to check that atleast on an average after score should be 10 points greater than before score?

Claim to be tested :-

→ Training was effective. That is on an average after scores are atleast 10 point greater than before score.

Data :-

Name	Before score	After score
—	—	—
—	—	—
—	—	—
—	—	—

we called it paired because before and after data is of the similar student. they are paired with each other. last wala example men or woman ka was not paired example.

Step1 :-

$H_0: \mu_{after} - \mu_{before} \geq 10$

$H_a: \mu_{after} - \mu_{before} < 10$

$n_{before} = 50 \qquad n_{after} = 50$

Step2

* use data analysis tool box of excel. and select (Pair + test.

NOTE:- you may have your conclusions confused in using paired T test VS either equal or unequal variance T test.

→ we fill our information in data analysis of paired T test and you will get result.

$t_{stat} = -1.82$

$t_{critical} = -1.6$

conclusion :- I Reject Null hypothesis.

Applications :-

Difference in means T test :-

Examples :-

1) Testing whether customer satisfaction across Flowlie companies in US is greater than europe

2) food and drug company want to compair difference between two disease.

3) Testing whether average productivity of woman employes is greater than that for men employes

Three types of difference in means T test :-

1) Paired T test :- used when there is pairing in the data.

2) T test difference in means (assuming equal variance)

3) T test difference in means (assuming unequal variance).

To calculate pvalue,
we have formula

pvalue = 2 * T.DIST (- | t-statistic |, residual df, TRUE)

↳ formula to find pvalue.

Now,

Hypothesis Testing in regression context.

→ t-cutoff approach } already seen
→ pvalue approach.
→ confidence interval approach

confidence interval approach :-

whenever we find the regression output we will see the upper end and lower end of that coefficient.

Eg :-

Step 1 :- $H_0 : \beta = 500$
$H_a : \beta \neq 500$

Step 2 :- consider 95% confidence interval for β
$= [212.6, 1084.6]$

conclusion :-

since 500 falls in the confidence interval hence we cannot reject NULL hypothesis.

When we do regression test, we get pvalues automatically, what are they ?

→ Excel is doing hypothesis test for every coefficients and giving us pvalues

These hypothesis are of kind as shown :-

→ Sales = $\beta_0 + \beta_1$ price + β_2 Adver + β_3 Brand

Automatic hypothesis testing by excel.

$H_0 : \beta_3 = 0$
$H_a = \beta_0 \neq 0$

$H_0 : \beta_1 = 0$
$H_a : \beta_1 \neq 0$

$H_0 : \beta_2 = 0$
$H_a : \beta_2 \neq 0$

→ Automatic p values for every coefficient comes by doing the test as whether Null hypothesis is whether coefficient is zero or not.

→ If p values comes out to be low we reject Null hypothesis and that is good, that variable can help to explain variation in y variable.

low p values are good. → as we reject Null hypothesis

→ In other words that variable is important in regression.

R^2 :
(goodness of fit)

* overall variation in Y variable : Total sum of squares

Explained variation in Y variable : Regression sum of squares

unexplained variation in Y variable : Residual sum of squares

Total sum of squares = Regression SS + Residual SS

$$R\text{-square} = \frac{\text{Regression SS}}{\text{Total SS}}$$

Adjusted R square decrease if we add unnecessary explanatory variable. , by value of R^2 never decreased.

Dummy Variables :-

→ we change categorical variable to numeric variable ~~by applying statistics~~. when using Regression analysis.

→ We use If and else statement for that.

Ex:- If(B2 = "M", 1, 0)

NOTE:-

If focus is more on predictions then, low R^2 is problem

If focus is on understanding relation between x and y variable then low R^2 may not be problem.

Dealing Variables in Regression model

$$\text{Weight} = B_0 + B_1 \text{Male} + B_2 \text{Height}$$

(this is categorical so we make it Dummy

Male – 1
Femal – 0

Gender	Height	Weight (y variable)	Gender (male) = If(A1="m", 1, 0)
M	:	:	:
F	:	:	:
:	:	:	:

make regression model.

	coefficient	P value
Intercept	−101	6E-100
Male	5.5	2.83E-20
Height	0.966	2.83E-40

→ Now let us interpret each and every coefficients.

Height: every 1cm increase in height of olympian weight increases by 0.97kg

Male: Male olympian than female olympian have 5.5 kg more.

Intercept:- Female with zero height −101 is weight of the person, but it does not make sense since height cannot be zero. (does not have managerial interpretation)

→ We can make it meaningful by mean centering height.

Male	Height (mean) = C1 - average (C1:C100)	Height	Weight
1	:	:	:
0	:	:	:
1	:	:	:
0	:	:	:

Now, again run regression (male x_1, mean center height x_2 and weight y)

Now,

	Coeff	p-value
Intercept	69.66	0
male	5.5	2.83e^{-10}
Height mean	0.96	2.2e^{-20}

} it is same

→ Now, we can interpret intercept as

Managerial interpretation :-

→ weight of the female ~~with~~ whose height is equal to average height in the entire data is 69.66 kg.

and male olympiad weight
= 69.66 + 5.5

Interpretation
→ on an average male olympian has weight that is 5.5kg more as compared to female olympian, all other variables kept at same level.

Scanned with CamScanner

Interaction effects in Regression :-

NOTE

$$\text{weight} = \beta_0 + \beta_1 \text{Male} + \beta_2 \text{Height}$$

* Effect of height on weight is same across gender. As we interpret as increase in 1cm of height will increase the weight by 0.966 kg, irrespective of gender since it was held constant.

* No gender difference allowed based on the height.

Q How do we allow impact of 'height' based on gender?

This is where Interaction effects come into play.

$$\text{weight} = \beta_0 + \beta_1 \text{male} + \beta_2 \overbrace{\text{Height}}^{\text{mean centred}} + \beta_3 \overbrace{\text{Height} \times \text{male}}^{\text{mean centred.}}$$

interaction variable

Now,

* let us interpret the efficient of height.

→ The impact of one cm increase in height on the weight, all other variables held constant.

→ To keep the interaction variable 'at same level', we need to consider female olympian because 1cm increase in height will also increase interaction variable and we have to keep that constant.

Here, β_2 is the impact of 1cm increase in height on weight of female olympians.

Now, what about impact of height on weight of male olympians?

Ans: β_3 → it gives us additional impact on weight for male olympians

Thus, total impact of height on weight for male olympians is $(\beta_2 + \beta_3)$

Now,

Transformation of variables in Regression :-

(Improving linearity)

Q why we need transformation?

→ If relation between Y and X variable is not linear.

Various transformations

* Invert one or more variable
* Take a square of variable
* Take a square root
* log transformations

log log model

$$\ln(Y) = B_0 + \beta_1 \ln(X_1) + \beta_2 \ln(X_2) + \dots$$

Interpretation,

when natural log of X_1 increases by 1 unit, then natural log of Y increases by β_1 units, all other variables are kept at same level.

OR

we read it like this way one percentage increase in X_1 corresponds to β_1 percentage increase in Y, all other variable at same level.

Semi - log model :-

$$LN(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots$$

↓

Interpretation

one unit increase in X_1, corresponds to $100\beta_1$ percentage increase in Y, all other variables at same level.

Now,

NOTE :-

Eg :- Annual demand for cocoa in millions pounds over a period of time.

$$LN(\text{demand}) = \beta_0 + \beta_1 LN(\text{price}) + \beta_2 LN(\text{Pcinc}) + \beta_3 \text{year}$$

(Pcinc = percentage increase)

→ Here we have Log-Log as well as semilog interpretation.

→ β_1 and β_2 is log - log interpretation also known as elasticity interpretation.

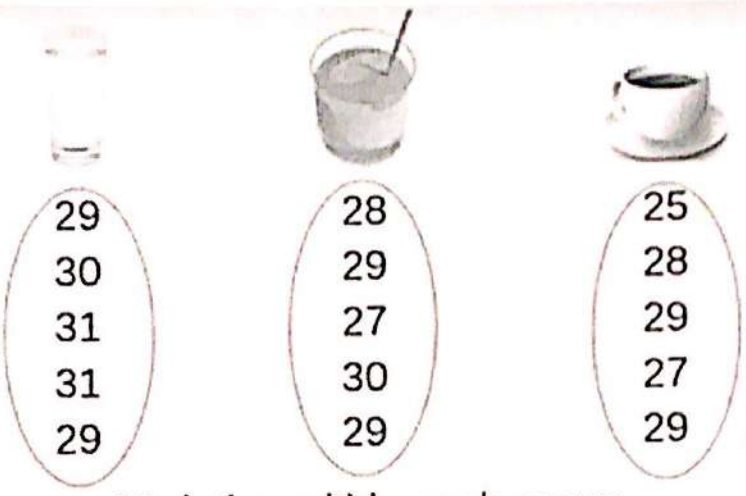
→ β_3 is semi log interpretation also known as growth rate interpretation

→ It is used for Analyzing relationship between one categorical and one numeric variable.

like total there are 60 people. divided equally into 4 diets.

ANOVA

Diet	weight loss	mean weight loss
A	96, , , ,	—
B	. . .	—
C	. . .	—
D	. . .	—



Variation within each group

The variation within each group,

H_0 = mean weight loss for all diets are same

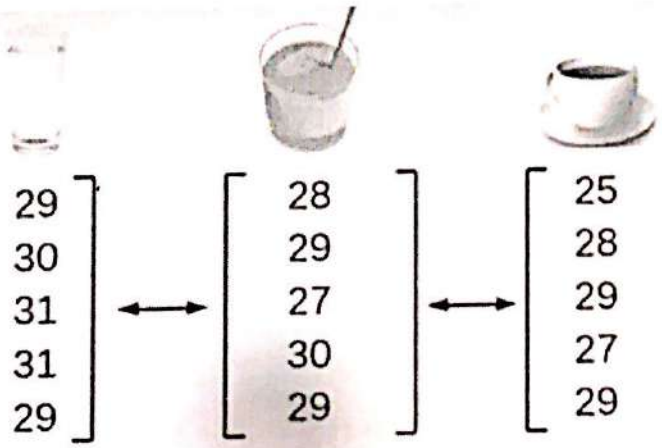
H_a : Atleast one is different

First example

Assumption :-

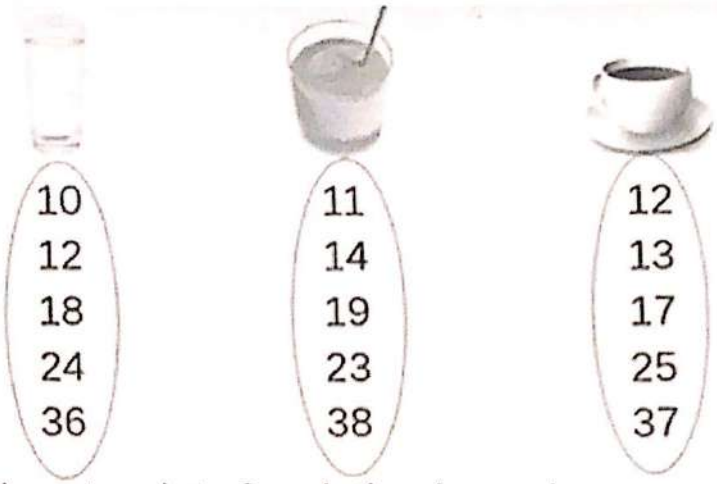
1) Independent groups, people on diet A is independent of diet B
2) Standard deviation of each group is roughly same
3) each group has large sample size. as groups are approx normal distribution.

If n is small then we can use Kruskal test or bootstrap approach.



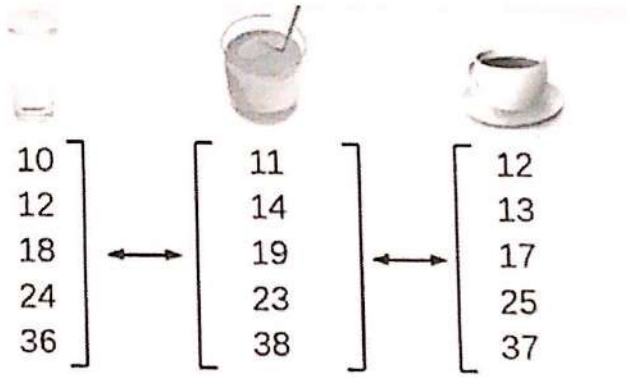
Variation between the groups

and the variation between the groups,



There's a lot of variation in each group...

to make it easier to see the patterns.



...but each group looks pretty much the same.

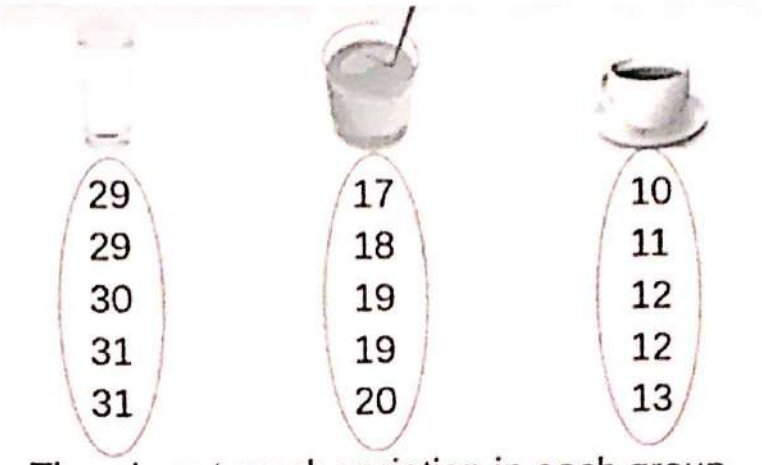
But all the groups look pretty much alike;

```
> help(aov)
> ? aov
> boxplot(weight loss ~ Diet)
> aov(weight loss ~ Diet)
> ANOVA1 <- aov(weight loss ~ Diet)
> summary(ANOVA1)
```

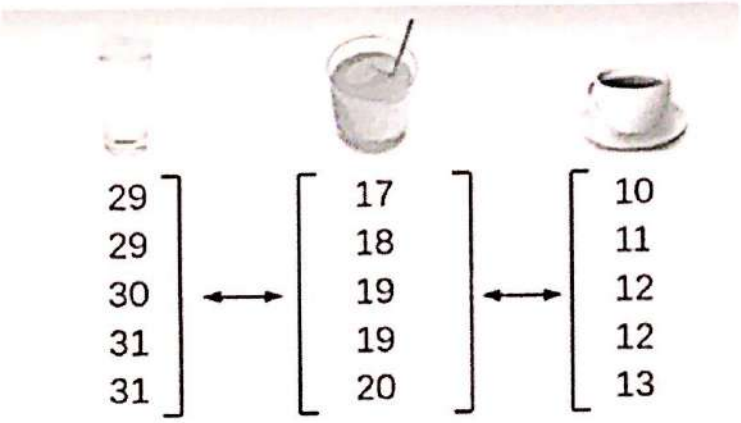
F = 6.118 p = 0.0013 hence p is small, we reject NULL hypothesis

> Tukey HSD (ANOVA) // to check की difference कौन कौन से mean में है, व कौनसे group में difference आ रहा है।
> plot (Tukey HSD))

Second example



There's not much variation in each group...



...but the groups look very different.

In this case we will reject the NULL Hypothesis .

↳ we assume our NULL hypothesis to be $H_0: \mu_0 = \mu_1 = \mu_2$

Figure out how much of the total variance comes from:

The variance *between* the groups

The variance *within* the groups

Calculate the ratio:

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

The larger the ratio, the more likely it is that the groups have different means (reject H_0).

STRESS DATA (SACHINJ, 2014)

4 PM to MIDNIGHT	MIDNIGHT TO 8 AM	8 AM TO 4 PM
7	5	1
7	6	3
6	3	4
9	5	3
8	4	1
7	6	1
6	5	2
7	4	6
8	5	5
9	5	4
	6	3
	7	4
	6	5

Within Groups Mean Sum of Squares

$$MSS_W = \frac{\sum_{g \in G} (X - \bar{X}_g)^2}{n - k}$$

Between Groups Mean Sum of Squares

$$MSS_B = \frac{\sum_{g \in G} n_g(\bar{X}_g - \bar{X}_G)^2}{k - 1}$$

Test Statistic for Tests Concerning the Differences between the Variances of Two Populations (Normally Distributed Populations)

$$F = \frac{MSS_B}{MSS_W}$$

$$df_B = k - 1$$

$$df_W = n - k$$

STRESS DATA (SACKVILLE, 2014)

4AM-6PM WEIGHT	6PM WEIGHT TO 9PM	9AM TO 4PM
7	5	1
7	6	3
6	3	4
9	5	3
3	4	1
7	6	1
6	5	2
7	4	6
3	5	5
9	5	4
	6	3
	7	4
	6	5
10	13	13
n_g	36	
$\bar{x}_G$	7.40	5.15	3.23
$\bar{x}_g$		5.08
k		3

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

$H_1: \mu_1 \neq \mu_2 \neq \mu_3$

$$MSS_W = \frac{\sum_{g \in G}(x - \bar{x}_g)^2}{n - k}$$

$$MSS_B = \frac{\sum_{g \in G} n_g(\bar{x}_g - \bar{x}_G)^2}{k - 1}$$

$$F = \frac{MSS_B}{MSS_W}$$

$df_B = k - 1$

$df_W = n - k$

	4 AM to MIDNIGHT	MIDNIGHT TO 7 AM	8 AM TO 4 PM
	7	5	1
	7	6	3
	6	3	4
	9	5	3
	7	4	1
	7	6	1
	6	5	2
	7	4	6
	8	5	5
	9	5	4
		6	3
		7	4
		6	5
n_g	10	13	13
n_G	36		
$\bar{x}_g$	7.40	5.15	3.23
$\bar{x}_G$	5.08		
k	3		

	4 AM to MIDNIGHT	MIDNIGHT TO 7 AM	8 AM TO 4 PM	$(x_1-\bar{x}_1)^2$	$(x_2-\bar{x}_2)^2$	$(x_3-\bar{x}_3)^2$
	7	5	1	0.16	0.02	4.98
	7	6	3	0.16	0.72	0.05
	6	3	4	1.96	4.64	0.59
	9	5	3	2.56	0.02	0.05
	7	4	1	0.36	1.33	4.98
	7	6	1	0.16	0.72	4.98
	6	5	2	1.96	0.02	1.51
	7	4	6	0.16	1.33	7.67
	8	5	5	0.36	0.02	3.13
	9	5	4	2.56	0.02	0.59
		6	3		0.72	0.05
		7	4		3.41	0.59
		6	5		0.72	3.13
n_g	10	13	13	$\Sigma = 10.40$	$\Sigma = 13.69$	$\Sigma = 32.31$
n_G	36					
$\bar{x}_g$	7.40	5.15	3.23			
$\bar{x}_G$	5.08					
k	3					

$$MSS_W = \frac{\sum_{g \in G} (x - \bar{x}_g)^2}{n-k}$$

$$MSS_B = \frac{\sum_{g \in G} n_g (\bar{x}_g - \bar{x}_G)^2}{k-1}$$

$$F = \frac{MSS_B}{MSS_W}$$

$$df_B = k-1$$

$$df_W = n-k$$

$(x_1 - \bar{x}_1)^2$	$(x_2 - \bar{x}_2)^2$	$(x_3 - \bar{x}_3)^2$
0.16	0.02	4.98
0.16	0.72	0.05
1.96	4.64	0.59
2.56	0.02	0.05
0.36	1.33	4.98
0.16	0.72	4.98
1.96	0.02	1.51
0.16	1.33	7.67
0.36	0.02	3.13
2.56	0.02	0.59
	0.72	0.05
	3.41	0.59
	0.72	3.13
$\Sigma = 10.40$	$\Sigma = 13.69$	$\Sigma = 32.31$

$SS_w = 56.4$

$n - k = 33$

$MSS_w = \frac{56.4}{33} = 1.71$ → This is mean squared within the group

$n_1(\bar{x}_1 - \bar{x}_G)^2 = 10(7.40 - 5.06)^2 = 53.82$

$n_2(\bar{x}_2 - \bar{x}_G)^2 = 13(5.15 - 5.03)^2 = 0.06$

$n_3(\bar{x}_3 - \bar{x}_G)^2 = 13(3.23 - 5.03)^2 = 44.49$

$\Sigma = 98.37$

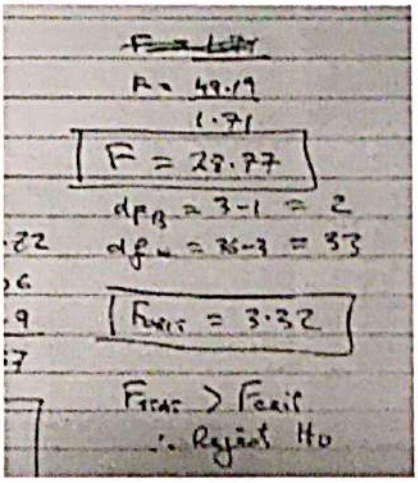
$MSS_B = \frac{98.37}{2} = 49.19$ → This is mean squared between the group

~~$F = \frac{1.71}{49.19}$~~

$F = \frac{49.19}{1.71}$

$F = 28.77$

If $F_{stat} < F_{critical}$
↳ fail to reject NULL Hypothesis


$F = \frac{49.19}{1.71}$

$F = 28.77$

$df_B = 3-1 = 2$

$df_w = 36-3 = 33$

$F_{crit} = 3.32$

$F_{stat} > F_{crit}$
∴ Reject H_0

$F_{stat} > F_{crit}$
↳ reject NULL Hypothesis

Hum F critical find करेंगे

$$F_{cri} = \frac{dof\ (between)}{dof\ (within)} = \frac{(K-1)}{(n-K)} = \frac{(3-1)}{33}$$

(2 and 33)
↓
Now इनको value को graph में देखेंगे to find F critical & found value to be

$F_{critical} = 3.32$ ✓

Note:

$F \uparrow \Rightarrow p \downarrow$
↳ If p value is less then we reject NULL Hypothesis

=

1) It works with categorical target variable "Success" or "failure"
2) It performs only Binary splits.
3) CART (classification & regression tree) uses GINI method to creat binary splits.

> tAB ← table (Gender, Smoke)
> Chisq. test (TAB, correct = T)
p value = 0.1866
> CHI ← Chisq. test (TAB, correct = T)
> Summary (CHI)
hence we fail to reject NULL hypothesis.

Chi square method :-

f_o → observed frequency
f_e → expected frequency

$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right]$$

Working Rule :-

1) consider the Null hypothesis : There is no difference between observed frequency & expected frequency. / X and Y are independent
   H_0
   H_a : X and Y are dependent.
2) calculated expected frequency f_e corresponding to each cell.
3) calculate χ^2 distribution by formula
   $$\chi^2 = \sum \left[\frac{(f_e - f_o)^2}{f_e} \right]$$ & calculate degree of freedom.
   ↓ it means एक random variable को select करेंगे कि कितने नजीक होते हैं

Assumptions :-

1) Groups are independent
2) All cell ≥ 1
3) All expected cell count ≥ 5
   ↳ if it is not met then we can use Fisher's test or Bootstrap test

(4) See and find the value of $X^2_{critical}$ from table.

(5) If $X^2 < X^2_{critical}$ then Null hypothesis is accepted otherwise Null hypothesis is rejected.

NOTE :- How to find expected frequency,

A/B	B_1	B_2	
A_1	f_{11}	f_{12}	N_1
A_2	f_{21}	f_{22}	N_2
	N_3	N_4	N

$$f_{11} = \frac{N_1 \times N_3}{N}$$

$$f_{12} = \frac{N_1 \times N_4}{N}$$

$$f_{21} = \frac{N_2 \times N_3}{N}$$

$$f_{22} = \frac{N_2 \times N_4}{N}$$

→ Chi square statistic is used for testing relationships between categorical variables. The null hypothesis of chi square is that there is no relationship between categorical variables of the population & they are independent.

→ It is also called 'goodness of fit' because it measures how well the observed distribution of data fits with distribution that is expected if variables are independent.

e of freedom :-

Binomial distribution : $(n-1)$

Poisson distribution : $(n-2)$

normal distribution : $(n-3)$

$(m \times n)$ table : $(m-1)(n-1)$

→ Pearson's chi square test can be used to find the correlation between two categorical variables, if sample size is greater than 50.

In random survey of public opinion :-

Question 1: do you drink

Question 2: Are you in favour of local option on sale of liquor

@2 :

	Yes	No	Total
Yes	56	31	87
No	18	6	24
Total	74	37	111

can you infer or not that local option on sale of liquor is dependent on individual drink?

Given that $X^2_{critical}$ at dof = 5 and level of significal = 5%.
$= 3.841$?

Sol: $f_{11} = \frac{87 \times 74}{111} = 58$, $f_{12} = \frac{87 \times 37}{111}$, $f_{21} = \frac{24 \times 74}{111} = 16$

$f_{22} = \frac{24 \times 37}{111} = 8$

$$X^2 = \frac{(58-56)^2}{58} + \frac{(29-31)^2}{31} + \frac{(16-13)^2}{16} + \frac{(8-6)^2}{6}$$

$X^2 = 0.95$

Hence null hypothesis is accepted that

⇒ Dice is tossed 120 times

No. noted up :	1	2	3	4	5	6
frequency :	30	25	18	10	22	15

Test hypothesis that dice is unbiased ($\chi^2_{0.05,5}$ = 1

Ans :-

Ho : Dice is unbiased

Calculation of expected frequency :-

$p(x) = \frac{1}{6}$

expected frequency = $120 \times \frac{1}{6} = 20$ for $x = 1, 2, 3, 4, 5$

$$\chi^2 = \frac{(30-20)^2}{20} + \frac{(25-20)^2}{20} + \frac{(18-20)^2}{20} + \cdots$$

$\chi^2 = 12.90$

Hence Null hypothesis is rejected.

→ Now we can use Chisquare technique on decision trees.

→ Since, Chisquare is algorithm to find out statistical significance between the difference between sub node and parent node.

→ Higher the chisquare value higher the statistical significance between sub-node & parent node.

→ It generate tree called (CHAID)

(Chisquare automatic Interaction Detector)

→ Ex. इस variable को target variable के साथ contigency table बनाओ, (जिस variable की chisquare value ज्यादा आएगी, वो variable root node बनेगा।)