No.

Date . .

Central Limit Theorem

The central limit theorem states that if you have a population with mean 'μ' and standard deviation 'σ' and take sufficiently large random samples from the population with replacement, then the distribution of the sample means, it will be approximately normally distributed

Probability:

Probability is a measure of the likelihood of an event

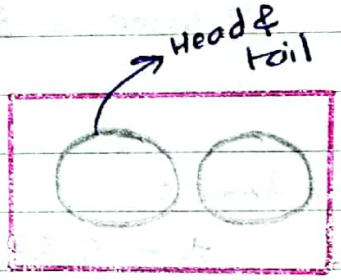
Eg: Tossing a fair coin

$P(H) = 0.5$, $P(T) = 0.5$

Mutual Exclusive Event

Two events are mutually exclusive if they cannot occur at the same time

a) Tossing a coin
b) Rolling a dice



Non Mutual Exclusive Events

Two events can occur at the same time.

• Mutual Exclusive Event Problem Statement.

1) What is the probability of coin landing on heads or tails?

↓

Addition Rule for mutual exclusive events

No.

Date . .

$P(A \text{ or } B) = P(A) + P(B)$

$= 1/2 + 1/2$

$= 1$

* What is the probability of getting 1 or 6 or 3 while rolling a dice?

$P(1 \text{ or } 6 \text{ or } 3) = P(1) + P(6) + P(3)$

$= 1/6 + 1/6 + 1/6 = 1/2.$

Non Mutual Exclusive Event Problem Statement:

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

Addition rule for non mutual exclusive event

Bag of Marbles: 10 Red, 6 Green, 3 (R & G).

* When picking randomly from a bag of marbles, what is the probability of choosing a marble that is red or green?

$P(A \text{ or } B) = P(A) + P(B) - P(A \cap B)$

$= \frac{10}{19} + \frac{9}{19} - \frac{3}{19}$

$= \frac{13}{19} + \frac{6}{19} = \frac{19}{19} = 1$

No.

Date . .

Dependent Event

Two events are dependent if they affed one another

* What is the probability of rolling a 6 sided dice to get a "5" and then a "3".

↓

Since it is independent event

$P(3) = 1/6$ & $P(5) = 1/6$

Multiplicative rule:

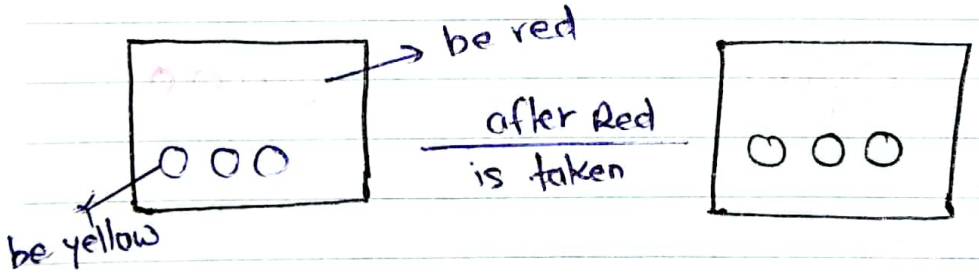
$P(A \text{ and } B) = P(A) * P(B)$

$= 1/6 * 1/6$

$= \underline{1/36}$

---

* In bags of marble, 4 Red marble, 3 Yellow marble. What is the probability of drawing a "~~orange~~ Red" and then drawing a "yellow" marble from the bag.



Soln $P(\text{Red}) = \frac{4}{7}$

$\therefore P(\text{Yello} / \text{Red}) = \frac{3}{6} = 1/2$

↖ Conditional probability

No.

Date . .

Now, P(Red and Yellow) = P(Red) * p(Yellow / Red)

$= \frac{4}{7} * \frac{1}{2} = = \frac{2}{7}$

Permutation

- each of several possible ways in which a number

Q. Considered the letters 'A', 'B' and 'C'. How many ways can these letters be arranged?

→ $\underline{3} * \underline{2} * \underline{1} = 6$ ways

3 ↑ First place has 3 choice

2 ↑ 2 choices for second

1 ↑ 1st choice left

→ ABC → ACB

→ BAC → BCA

→ CAB → CBA

} 6 option

n be total no. of object

r be no. of selection,

* Orders matter in permutation

$$ {}^{n}P_{r} = \frac{n!}{(n-r)!} $$

No.

Date . .

Combination

The number of combinations of 'n' objects taken 'r' at a time can be calculated using the combination formula.

$$C(n,r) = \frac{n!}{r!\,(n-r)!}$$

Covariance

- It is a measure of the joint variability of two ~~mean~~ random variable.

Age	weight
12	40
13	45
14	48
17	60

→ shows

age ↑ ⇒ weigh ↑

age ↓ ⇒ weight ↓

$$Cov(X,Y) = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{n-1}$$

↓

It is variance with own.

$$cov(X,X) = \sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

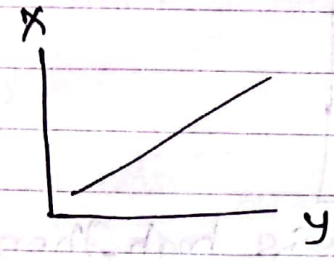
No.

Date

In +ve Covariance

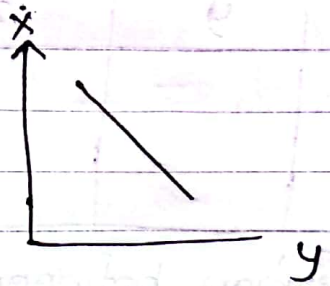
$x\uparrow \rightarrow y\uparrow$

$x\downarrow \rightarrow y\downarrow$

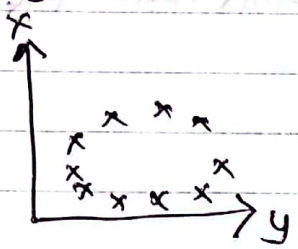


In -ve Coveriance

$x\uparrow \rightarrow y\downarrow$



In 0 coveriance



Spearman Correlation

→ It is for non linear as covariance works well in only linear datasets.

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

Rank

Rank alway in ascending

X	y	R(x)	R(y)
10	4	4	1
8	6	3	2
7	8	2	3
6	10	1	4

No.

Date

* Why do we need correlation? ~~will be~~

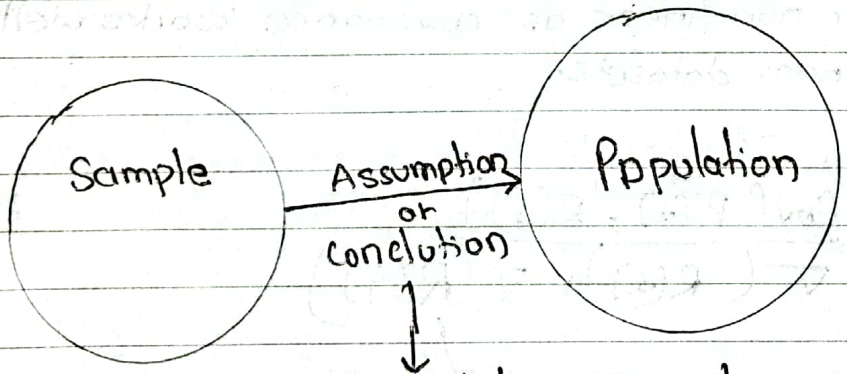
x	y	z	o/p

→ If corelation between x & o/p is high, then it is important feature

→ If there exist very +ve correlation betn x & y, then we can drop the one

Hypothesis testing:

Before comming into it, let's cover Inferential Statistics



To validating Assumption or conclusion we perform hypothesis testing

Steps of hypothesis testing:

ⓐ Null hypothesis: default value.

like: Person hasn't committed the crime

No.

Date . .

② Alternate Hypothesis: Oposite of null hypothesis
Eg: Person is criminal

③ Perform Experiment → DNA, finger Print, Weapons
Eye witness, footage.

Example:

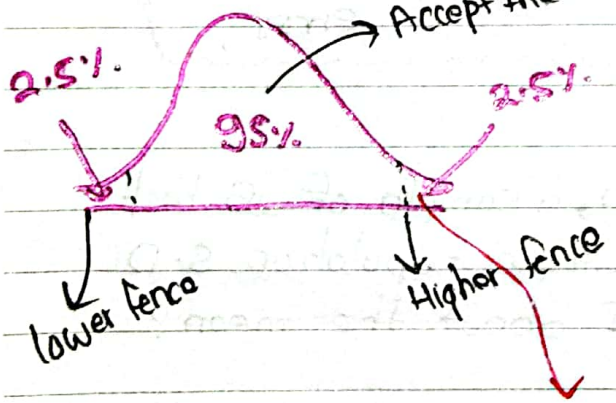
1) Null hypothesis: coin is fair
2) Alternate hypothesis: Coin is not fair
3) Experiment:

Head : 70 { 40 { 55
Tail : 30 { 60 { 45

Then, we go to confidence interval. Say, if C.I is [20 - 80], then our experiment make null hypothesis true.

Hence Null hypothesis accepted.

What is 95% C.I means? Accept the null hypothesis [we fail to reject the null hypothesis]



C.I = 95%
Significance value = 1 - 0.95 = 0.05

Reject the null hypothesis

No.

Date . .

Point Estimate: The value of any statistics that estimates the value of a parameter is called point estimate.

$\bar{x}$ —— we estimate ——→ μ

↓ ↓

Sample mean → Statistic

Population mean → Parameter

Parameter = Point Estimate ± Margin of error.

So, for confidence Interval

Lower fence = Point Estimate − Margin of error

Higher fence = Point Estimate + Margin of Error

Margin of error = $Z_{\alpha/2} \cdot \left[\frac{\sigma}{\sqrt{n}}\right]$ ← Standard error.

1 − CI = Significance value

Q. On the quant test of CAT Exam, a sample of 25 test taker has a mean of 520 with a population S.D of 80. Construct a 95% CI about the mean?

→

$n = 25$

$\bar{x} = 520$

$\sigma = 100$

C.I = 95 $\Rightarrow \alpha = 1 - 0.95$

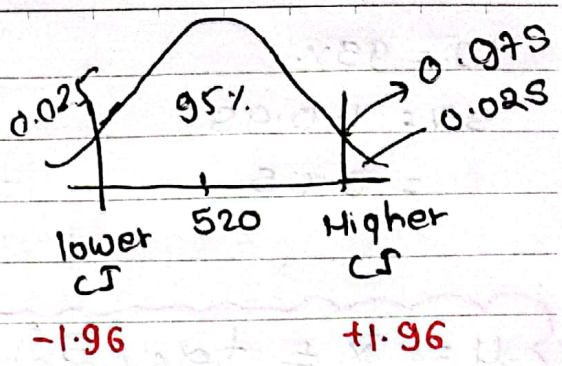
$= 0.05$

No.

Date

We know,

Point Estimate ± marginal error
||
parameter.



Also,

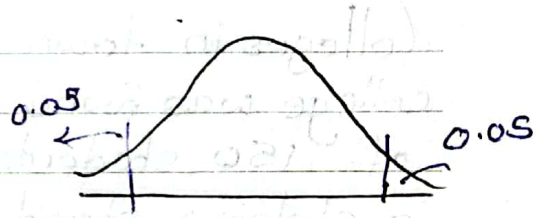
Lower CI = Point Estimate - Marginal error

$$= 520 \pm Z_{\frac{0.05}{2}} \cdot \left[\frac{80}{\sqrt{25}}\right]$$

$$= [504, 55136]$$

Q. $\bar{X} = 480, \quad \sigma = 85, \quad C.I = 90\%, \quad n = 25$

$\alpha = 1 - 0.9 = 0.1$



$$C.I = 480 \pm Z_{0.05} \cdot \frac{85}{5}$$

$$= 480 \pm 1.64 \times 17$$

$$=$$

Q. On the quant test of CAT exam, a sample of 25 test takers has a mean of 520 with a sample standard deviation of 80. Construct 85% CI about the mean?

→

$\bar{x} = 520$

$S = 80$

No.

Date . .

CI = 95%.

S.V = 1 - 0.95

= 0.05

When sample S.D is given, then t-test is performed.

$$\mu = \bar{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

Degree of freedom = n - 1

$$= 520 \pm 2.064 \times \frac{80}{\sqrt{25}}$$

$$= [486.976,\ 553.024]$$

One tail and Two Tail

Colleges in town A has 85% placement rate. A new college was found open and it found that a sample of 150 students had a placement rate of 88% with a ~~stolem~~ standard deviation of 4%. Does this college has a different placement rate with 95% CI?

→

When it say 'different placement rate', it can be greater than 85 or less than 85.. it is two tail problem.

No.

Date . .

Hypothesis Testing:

A factory has a machine that fills 80ml of baby medicine in a bottle. An employee believes the average amount of baby medicine is not 80 ml. Using 40 Samples, he measures the average amount dispersed by the machine to be 78ml. with S.D of 2.5.

(a) State null and alternate hypothesis

(b) At 95%, CI, is there enough evidence that support machine is working properly or not?

Ans:

(a) → Null hypothesis: (H_0) : $\mu = 80$

Alternate hypothesis (H_1) : $\mu \neq 80$

(b) →

$\mu = 80$

CI = 0.95 , $\alpha = 1 - 0.95 = 0.05$

$n = 40$

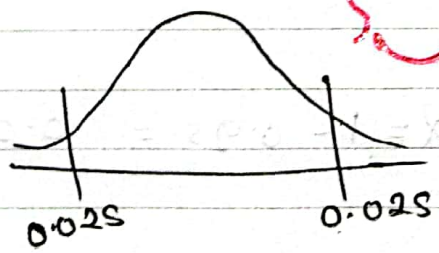
S.D = 2.5

If $n \geq 30$ or population SD → Z test

else

$n \leq 30$ & sample std → t-test

Z test



$$Z = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}} \Rightarrow \frac{78 - 80}{\frac{2.5}{\sqrt{40}}} = -5.05$$

No.

Date . .

Conclusion:

Decision rule: If $z = -5.05$ is less than -1.96 or greater than 1.96, reject null hypothesis.

Since $z = -5.05 < -1.96$

↓

Reject null hypothesis.

$\therefore \mu \neq 80$

A compaint was registered. the boys in a government school are underfed. Average weight of boys of age 10 is 32 kgs with SD: 9kg. A sample of 25 boys were selected from the government school & the average weight was found to be 29.5 kgs. with CI = 95%, check it is true or false.

Soln

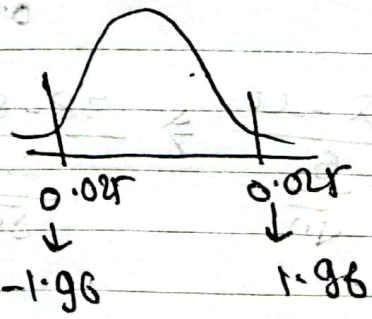
S.D = 9kg

$\mu = 32$kg

$n = 25$

$\bar{x} = 29.5$

CI = 95 $\therefore \alpha = 1 - 0.95 = 0.05$



No.

Date . .

Null hypothesis (H_0) $\Rightarrow \mu = 32$

Alternate hypothesis (H_1) $\Rightarrow \mu \neq 32$

$$Z = \frac{29.5 - 32}{\frac{9}{\sqrt{25}}} = -1.39$$

Conclusion : $-1.39 > -1.96$, So we accept the null hypothesis

"We fail to reject null hypothesis"

Q. A factory manufactures cars with a warranty of 5 years or more on the engine and transmission. An engineer believes that the engine or transmission will malfunction in less than 5 years. He tests a sample of 40 cars and finds the average time to be 4.8 years with a standard deviation of 0.50.

(a) State the null and alternate hypothesis.

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

Soln,

Step1: $H_0 : \mu \geq 5$

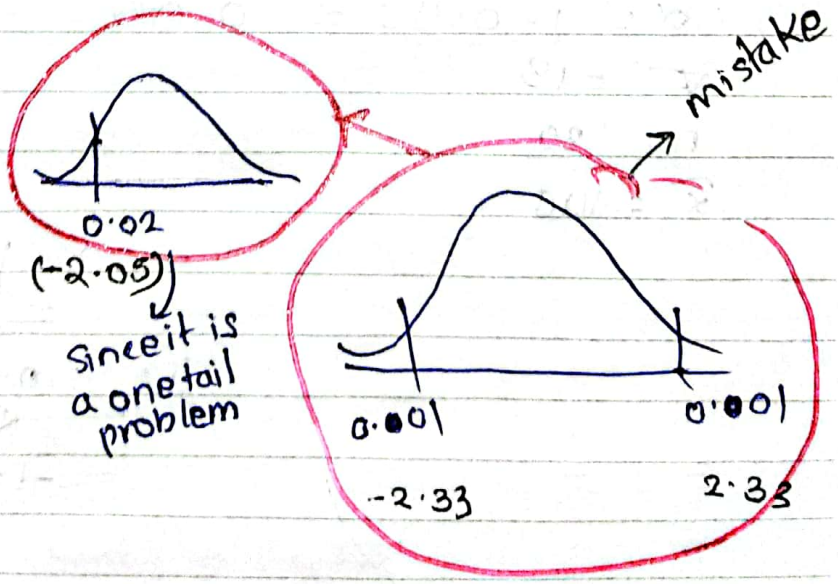
$H_1 : \mu < 5$

Step2: C.I = 98

$n = 40$

$\bar{X} = 4.8$

$S = 0.5$



No.

Date . .

$$Z = \frac{4.8 - 5}{\frac{0.5}{\sqrt{40}}} = -2.529 < -2.05$$

Hence, we reject the null hypothesis.

② In the population, the average IQ is 100 with a standard deviation of 15. A team of scientists wants to test a new medication to see if it has a +ve or -ve effect, or no effect at all. A sample of 30 participants who has taken the medicine has a mean of 140. Did the medication affect intelligence? CI = 95%.

Soln,

$H_0 \Rightarrow \mu = 100$ [No affect]

$H_1 \Rightarrow \mu \neq 100$ [effect]

↓

Two tail problem

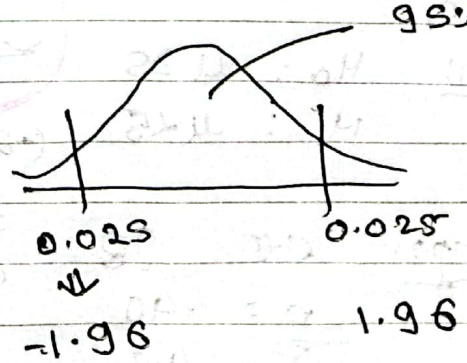
CI = 95%.

$\alpha = 1 - 0.95 \Rightarrow 0.05$

$\sigma = 15$

$n = 30$

$\bar{x} = 140$



No.

Date

$$Z = \frac{\bar{X} - u}{\frac{\sigma}{\sqrt{n}}} = \frac{140 - 100}{\frac{15}{\sqrt{80}}} = 14.60$$

Since: $14.60 > 1.96$, we reject the null hypothesis, the medication has effect.

Q. The average weight of all resident in a town XYZ is 168 pounds. A nutritionist believes the true mean to be different. She measured the weight of 36 individuals and found the mean to be 169.5 pounds with the standard deviation of 3.9?

(a) Null & Alternate hypothesis

(b) 95%., is there ~~average~~ enough evidence to discard the null hypothesis?

Soln,

$H_0 \Rightarrow u = 168$

$H_1 \Rightarrow u \neq 168$ → Two tail test

Given,

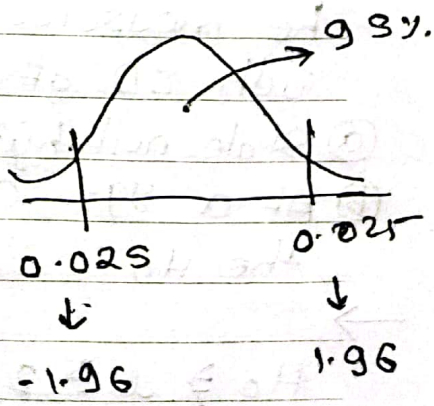
C.I = 95%

$\alpha = 1 - 0.95 = 0.05$

$\bar{X} = 169.5$

$S = 3.9$

$n = 36$



$$Z = \frac{\bar{X} - u}{\frac{S}{\sqrt{n}}} = 2.3076$$

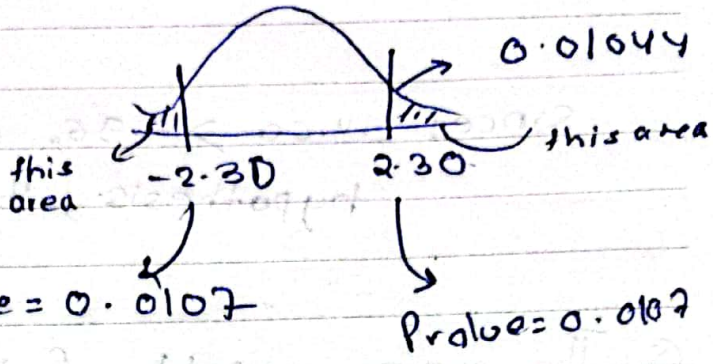
Since $Z = 2.3076 > 1.96 \Rightarrow$ Rejected.

No.

Date

Solving with P-value:

Seeing the value



Paggregate = 0.0107 + 0.0107

Paggregate = .0214

If Paggregate > Significance
accept null hypothesis

Pagg < C.I
↳ rejecting

A company manufactures bikes batteries with an average life span of 2 years or more years. An engineers believes this value to be less. Using 10 sample, he measures the average life span to be 1.8 years, with SD of 0.15.

(a) State null hypothesis

(b) At a 99% CI, is there enough evidence to discard the Ho?

→

$H_0 \Rightarrow \mu \geq 2$

$H_1 \Rightarrow \mu < 2$ → 1 tail problem

degree of freedom ⇒ max^m no. of choice from sample.

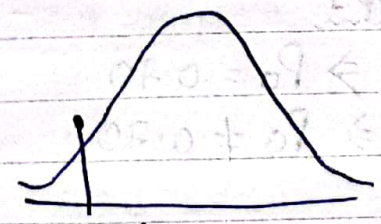
No.

Date . .

Given,

CI = 99%. ⇒ $\alpha = 1 - 0.99$

$= 0.01$



0.01

↓

2.821

↓

take negative since it is in left side

↓

-2.821

$n = 10$

$s = 0.15$ } Suggest t-test

$\bar{x} = 1.8$

degree of freedom = n - 1

= 9

Now,

$$t_{score} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{1.8 - 2}{\frac{0.15}{\sqrt{10}}} = -4.2163$$

Since: $-4.2163 < -2.821$, we reject the null hypothesis.

The average life of battery is less than 2 years.

Z test with Proportions:

A tech company believes that the ~~per~~centage of residents in town XYZ that own a cell phone is 70%. A marketing manager belives that this value to be different. He conducts the survey of 200 individuals and found that 130 responded 'yes' owning a cell phone.?

(a) state null and alternate hypothesis.

(b) At a 95%, CI, is there enough evidence to reject the null hypothesis?

No.

Date

→ Step 1

$H_0 \Rightarrow P_0 = 0.70$

$H_1 \Rightarrow P_0 \neq 0.70$

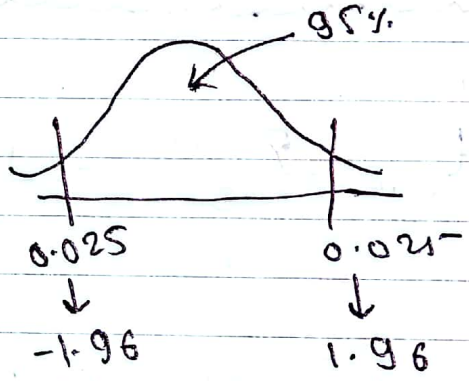
$q_0 = 1 - 0.70 = 0.30$

Step 2:

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

Step 3:

$$\hat{P} = \frac{130}{200} = 0.65$$



Step 4:

$$Z\text{ test} = \frac{\hat{P} - P_0}{\sqrt{\frac{P_0 q_0}{n}}}$$

$$= \frac{0.65 - 0.70}{\sqrt{\frac{0.70 \times 0.65}{200}}}$$

$$= -1.54$$

Since: $-1.54 > -1.96$, so we fail to reject null hypothesis.

No.

Date . .

Q. A car company believes that the percentage of resident in city ABC that owns a vehicle is 60% or less. A sales manager disagree with this and conduct a hypothesis testing surveying 250 residents and found that 170 responded 'Yes' to owning a vehicle.

(a) State the null hypothesis & alternate hypothesis

(b) At 10% significance level, is there enough evidence to support the idea that vehicle ownership in city ABC is 60%.

Solo

Step 1:

$H_0 \Rightarrow P_0 \leq 0.6$

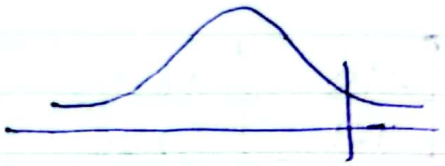
$H_1 \Rightarrow P_0 > 0.6 \longrightarrow$ 1 tail problem

Step 2:

C.I = 90%, $\alpha = 0.1$

Step 3:

$\hat{P} = \frac{170}{250} = 0.68$



Step 4:

$$z \text{ test} = \frac{0.68 - 0.6}{\sqrt{\frac{0.6 \times 0.4}{250}}}$$

$$= 2.581$$

Since: $2.581 > 1.28$, we reject the null hypothesis

No.

Date . .

Chi Square test

→ χ^2 test is a statistical test used to determine whether there is a significant association between two categorical variable.

→ It claims about population proportion.

→ It is a non-parametric test that is performed on categorical data.

In the 2000 US sensuses, the age of individuals in a small town found to be the following:

<18	18-35	≥35
20%.	30%.	50%.

In 2010, ages of n: 500, individuals were sampled. Below are the results:

<18	18-35	>35
121	286	91

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

→

	<18	18-35	>35
Expected	20%.	30%.	50%.

No.

Date

when n = 500 is sampled:

	<18	18-35	>35
Observed:	121	288	91
Expected:	100	150	250

Step 1:

Null hypothesis (H_0): The data meets the expected distribution

(H_1): The data doesn't meet the expected distribution

Step 2:

$\alpha = 0.05$, CI = 95%.

Step 3:

d.f = class - 1 $\Rightarrow$ 3 - 1 = 2

Step 4:

$\chi_\alpha = 5.991$

decision boundary:

if $\chi^2 > \chi_\alpha^2$:

Reject the null hypothesis

Step 5:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = \frac{(121-100)^2}{100} + \frac{(288-150)^2}{150} + \frac{(91-250)^2}{250}$$

Rejecting $\leftarrow$ = 232.494

No.

Date . .

Distribution

ⓐ Bernoulli Distribution

→ It models a random experiments with only two possible outcomes, 0 or 1.

→

$p = 1 - q$

&

$q = 1 - p$

→ 0 or 1
→ pass or fail

Mean ⇒ Expected Value = $1 - p$

variance ⇒ $Var(X) = p(1-p)$

or

$p \cdot q$

ⓑ Binomial Distribution

→ descrete probability distribution that describes the number of successes in fixed number of independent and identical bernoulli trials.

n → number of trials
p ⇒ probability of success at each trials

$$P_x = \binom{n}{x} p^x q^{n-x}$$

P_x ↙ number of success in n trial

No.

Date . .

Mean (expected value) $\Rightarrow E(x) = n * p$

Variance $\Rightarrow V(x) = n * p * (1-p)$.

© Power Law

→ The power law distribution is a continuous positive-only univariate distribution that describes a quantity whose probability decreases as a power of it magnitude.