**STATISTICS – KRISH NAIK**

Statistics- It is the science of collecting, organising and analysing the data. Stats is used for better decision making.

Types of statistics

* Descriptive stats: It consists of organizing and summarizing data

Eg: Avg value, percentiles or percentages of some features

* Inferential stats: Technique whereas we use data to form conclusions

Eg: Does this feature of state similar to whole nation’s?

Sampling Techniques

* Simple random sampling: every member of the population (N) has equal chance of being selected for the sample (n). Eg: Exit poll, test for medicines
* Stratified sampling: Where the population (N) is split into non-overlapping groups (this is called strata means layering)

Eg: I wanna do a survey. Samples are divided based on gender. Survey will be conducted differently for male and female people. Another example is doing the survey by making layers based on age instead of gender.

**Possible interview question** : Can we do stratify sampling based on profession?

**Ans**: Probably No. Say professions are dot net, python, php, electronics engineer etc. An electronics engineer may know python or dotnet as well. So we cant create a non-overlapping group

* Systematic sampling: From the population N, we pick up every nth individual without any particular reason.

Eg: Doing a survey outside a mall. Choose the 8th person appears.

* Convenient sampling: We are conducting a survey related to a specific topic (eg electric vehicles). So the samples should be chosen in such a way that people who should have domain knowledge on this topic.

Eg: RBI conducts household survey to calculate cost estimation in kitchen. In that scenario we can apply convenient sampling (asking questions to only women). Or we can try stratified sampling.

**Histogram**

Eg: Say we have a set of age from 10 to 65. Create bins like 10-20, 20-30, etc. and plot the histogram. PDF is the kernel density estimation of histogram (ie smoothened version of histogram is pdf).

Bar chart vs histogram: Bar representation is for discrete and histogram for continuous.

Bar chart example: Say we have a list of flowers where the list contains lily, sunflower and rose. Bar chart shows the no of flowers in the list. Ie discrete values.

**Central tendency**

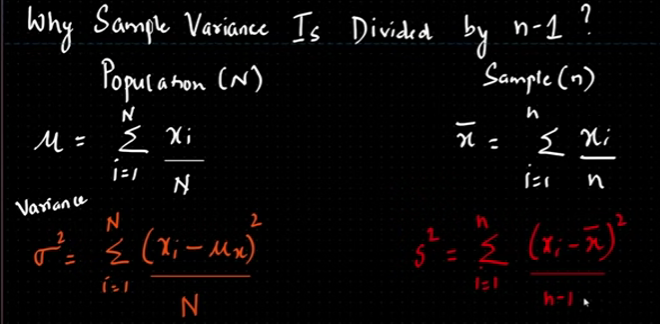
* Methods for measuring central tendency – mean, median, mode.

**Possible interview Que :** Benefits of median and situations where mode is used.

A: Median is least prone to outliers. Mode is nothing but the most frequent occurring element. If any particular outlier entry are more then mode value will be centred to that outlier value. But this is useful in case of a dataset with missing values. Ie missing values can be replaced with most frequent occurring element.

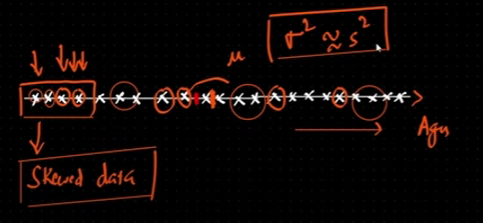
**Variance and std dev** – variance gives the spread or dispersion in the distribution of a feature. Std dev is a unit to specify how far a specific entry lies from mean. Ie, say mean is 2.8. 1 std dev is 4.17. An entry 5 will be 1.5 std dev away from the mean.

**Imp Interview Que** : Denominator of population variance is n. Then why sample variance is divided by n-1?



**A**:

When we calculate sample mean/variance, it is supposed to be very similar or near to population mean/var. But in case of skewed data, samples may pick from the tail portion. In that scenario, the sample mean/var will be not near to pop mean/var. So researchers tested to divide using n-1, n-2, n-3 and so on and found out n-1 gives more similar answer to pop mean/var. This is also called unbiased estimation.



**Percentile** : If percentile value of entry x is n% that means, n percentage of total distribution is less than x.

Eg: Dataset – {2,5,6,7,1,1,9,0,4,7,3,3,6,2,1,5,10,19,19,2}. Percentile of value 19 = # of values below 19 / total values \* 100

17/20 \*100 = 85. Meaning 85% of total distribution is less than 10.

**Interview Que** : What is the value at percentile of 25 ?

A: Index position of value = Percentile \* (n+1) / 100

Here answer is 5.25. But there is no index of 5.25. Take 5th and 6th index and take average which is 1 which is the value of 25 percentile.

**Inter Quartile Range (IQR**) – Used for removing outliers. Values less than lower fence and above upper fence are considered as outliers.

Lower fence = Q1 – 1.5(IQR) ; Q1 – 25 percentile

Upper fence = Q3+1.5(IQR) ; Q3 – 75 percentile

IQR = Q3-Q1

Eg: Data – {1,2,2,2,3,3,4,5,5,5,6,6,6,6,7,8,8,9,27}

Q1 = 25\*(19+1) / 100 = 5 ; 5th index value is 3. Similarly Q3=7

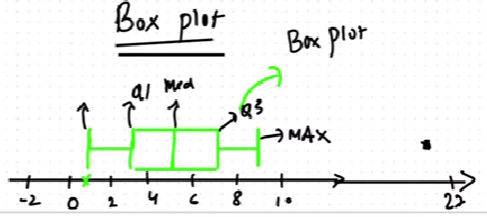
IQR = 7-3 = 4

Lower fence = 3 – 1.5(4) = -3

Upper fence = 7 + 1.5(4) = 13

**Five number summary** : min =1 ; Q1 = 3; median =5 ; Q3= 7; Max =9

Using the above data we can draw **box plot**.



**Distributions**

Empirical Formula - 68-95-99.7% ; ie 68% of data will be distributed between mean+/-sigma (ie first std dev).

Z score = x-mean / std dev. Used to spot the std dev position of a data entry. Say a distribution having mean=4 and std dev=1. At what std dev 4.75 lies ?

Z score = 4.75-4 /1 = 0.75 . So 4.75 lies at 0.75 std dev.

Consider a distribution {1,2,3,4,5,6,7} (this is a gaussian distribution) which has mean of 4 and std dev of 1. If we apply z score to every value results {-3, -2, -1, 0, 1, 2, 3} which is the std dev position of respective values. This resulted distribution has mean =0 and std dev=1 is called **standard normal distribution**.

Practical application : While doing an ML problem, the dataset may contains features having different units. So inorder to make the computation easier and uniformity we will convert the feature distribution to standard normal distribution. This is called **standardization** where internally applying z-score to every elements).

Normalization (min max scaler) – make values between 0-1 or -1 to +1. This is so common especially in CNN to make image pixel size of 0-255 to 0-1

**Problem : ODI series 2011. Series avg score – 250. Std dev- 10. Sehwag avg score - 240.**

**ODI series 2010. Series avg score – 260 , Std dev=12. Sehwag avg score – 245. Comparing both series in which year sehwag final score was better?**

**A:** 2011 : z-score = 240-250 / 10 = -1

2010 : z-score = 245-260 / 12 = -1.25

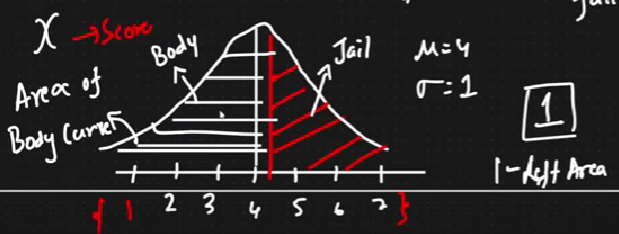
2010 z score gives more area. So this is better.

**Stats Interview Question : Say we have a distribution {1,2,3,4,5,6,7} having mean =4 and std dev=1. What percentage of scores fall above 4.25?**

**A:** Z = xi-mu / sigma = 4.25 -4 / 1 = 0.25 ; Means 4.25 lies @ 0.25 std dev away from mean.

For finding percentage we need to find area first ie area of tail portion. Area of tail portion = 1- area of body portion. Make use of z-table to find the area of left portion, which is 0.5987 (Refer Left z table)

So area of right portion (tail portion) = 1-0.5987 = .4 = 40%.



**Q: Average IQ of Jews are 100 with std dev of 15. What % of population would you expect to have an IQ lower than 85?**

**A**: mean = 100 ; sigma = 15 ; z-score = 85-100 /15 = -1.

Area = 0.84 (from left z table). So ans = 1-0.84= 16%

**Probability**

**Addition rule**

Mutually exclusive event: If the events cannot occur at the same time. Eg: coin toss, dice roll.

Non mutual exclusive events: multiple events can occur at the same time. Eg Card. Queen and hearts can occur.

Q: In coin toss whats the prob of getting heads or tails? Whats the prob of getting 1 or 3 or 6 in dice roll?

A: Both are mutually exclusive events. For coin toss- P(A OR B) = P(A) + P(B) = 0.5 +0.5 =1

For dice roll – 1/6 + 1/6 +1/6 = 0.5

Q: Card selection. Whats the prob of getting queen or heart?

A: Non mutual exclusive. P(Q) = 4/52 ; P(H) = 13/52 ; P(Q and H) = 1/52 [This is possible cos of mutually exclusive]

Addition rule of non mutual exclusive events is P(A OR B) = P(A) + P(B) - P(A AND B)

So P(Queen OR Heart) = 4+13-1 /52 =16/52

**Multiplication rule**

Independent events: In dice roll, each and every outcome is independent. P(A AND B) = P(A) \* P(B)

Dependent events: A box contains 3 red balls and 2 green balls. Prob of getting a red ball first is 3/5. Second time its 2/4. So second probability is depended on first. P(A AND B) = P(A)\*P(B|A).

Because of this dependent event there is an algorithm called Naïve Bayes where it defines conditional probability.

Q: Whats the prob of getting 5 and then 4 on rolling a dice?

A: Independent event: So P(5 and 4) = 1/6 \* 1/6 = 1/36

Q: Prob of drawing a queen and then an ACE from a deck of card?

A: Dependent event. Once queen is drawn, total cards will be reduce

P(Q and A) = P(Q)\*P(A|Q) = 4/52 \* 4/51

Similar que : Box contains 3 green balls and 2 red balls. Prob of getting green and then red ?

P(G AND R)= P(G) \* P(R|G) = 3/5 \* 2/4

**P-value, Hypothesis testing, CI, significance value…**

**p-value** : Touchpad of laptop follows normal distribution. P value of 0.8 at a particular region means, out of 100 time we use touch pad, 80 times we touch at this particular region. P value is basically the probability.

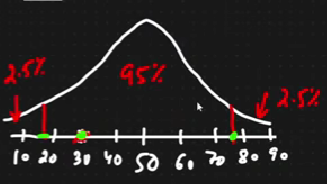
* If p-value < significance value – reject null hypothesis and accept alternative hypo and viceversa. Assuming alpha = 0.05 and if p value < alpha, which means it is less than 5% probability that the null is correct. This alpha value is the threshold for decision which is decided by domain expert.

**Hypothesis testing**

Steps: 1) Null hypothesis 2) Alternative hypo 3) Experiment 4) Reject or accept null hypothesis.

Suppose we are conducting an experiment to know whether a coin is fair or not. We will do 100 tosses. If P(H)=P(T)= 0.5. Then fair coin. In hypothesis testing first we will define null hypothesis. Here NH is the coin is fair. Alternative hypo is coin is biased.

We have started our experiment. Tossed for 100 times and got 30 heads. Can we say that coin is fair or not? How do we decide how far we can go from the threshold value to decide? We will use **significance value** (alpha) which is defined by a domain expert. If alpha =0.05 means **confidence interval** = 1-0.05 =0.95. Means CI is 95%.



When we get heads 30 times, who are we to say that coin is not fair? Its decided by domain experts using significance value. If the experiment results fall with in the confidence interval then we can accept null hypothesis. CI says 20 and 75 are the limits. If we get 10 heads then we reject null hypo and accept alternative hyp.

If alpha= 0.2, then CI – 80%. 0.1 % ie 10 % one both tails which is out of CI. If the experiment results in 25 heads, we will reject null hyp (lower threshold is 30).

**Type 1 and Type 2 error**

Outcome 1: We reject Null hypothesis when in reality its false – good decision

Outcome 2: We reject Null hypothesis when in reality its true – bad decision – type 1 error.

Eg: Acquitting an innocent person, say he is guilty

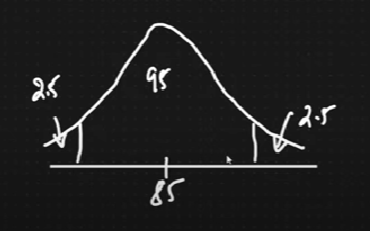
Outcome 3: We accept Null hypothesis when in reality its false – bad decision – type 2 error.

Outcome 4: We accept Null hypothesis when in reality its true – good decision

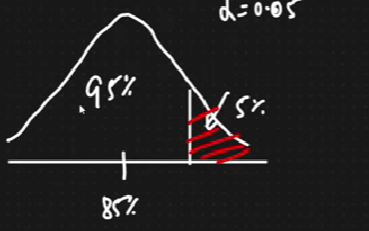
**One tailed and two tailed test**

Q: Colleges in Karnataka has a placement rate of 85%. A new college was recently opened and it was found that a sample of 150 students had a placement rate of 88% with std dev of 4%. Does this college has different placement rate? Let alpha=0.05

A: We need to know whether the college has different placement rate. So the rate can be decreased or increased so the problem is a two tailed test. 2.5% on both sides cos of 0.05 alpha and CI is 95%.



If the question was like, does the college has a placement rate more than 85%? Then the problem would be one tailed test.



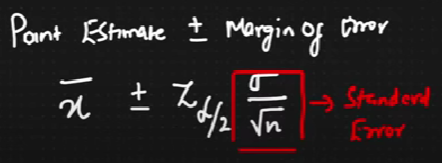
**Confidence Interval**

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

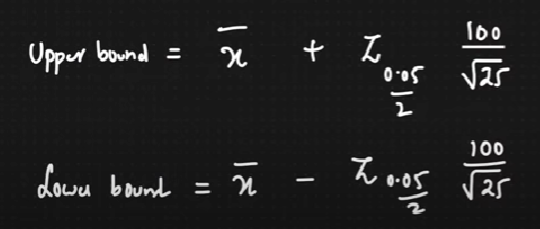
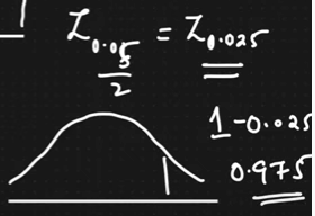
In inferential stats, using the sample data, we will estimate some statistic ink population (eg: mean, x\_bar=2.9 which is the point estimate and pop mean mu=3) . Population mean = point estimate +/- margin of error. This is how we will get confidence interval.

**Q: On the CAT exam, the population std dev is known to be 100. A sample of 25 students has a mean of score 520. Construct a 95% CI about the mean.**

A: Here we are provided with pop std dev, sigma = 100, not sample value. n =25 ; sample mean x\_bar=520. We need to construct a CI about the mean of 95% which means alpha=0.05 ie 2.5% either side. We need to find the values which creates the CI. **Whenever pop std dev is given, we have to apply z test.**



For z test to be happen, there are 2 conditions, pop std dev is given and sample size n>=30. Here its 25 , that’s ok. Here we have taken 25 for easy computation.

Z score is taken from z table. Z0.025 value means what ? Total area of curve is 1. We need area of 1-0.025 = 0.975. So we need to fetch value of Z0.975. Z is 1.96 ( refer left z table)

Upper bound = 520 + 1.96 (20) = 559.2

Lower bound = 520 – 1.96 (20) = 480.8

So confidence interval – [lower, mean, upper] – [480.8, 520, 559.2]

Now after performing any experiment, when the value falls in between CI, we can accept the null hypothesis else reject.

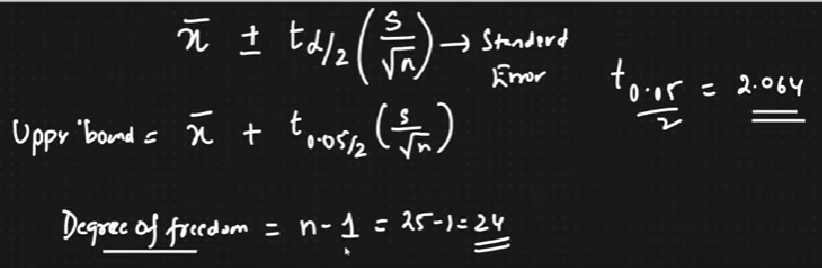
Practise question – Find the average size of the sharks throughout the world. Assume the population std dev, x\_bar and n value, alpha=0.05. Find the CI.

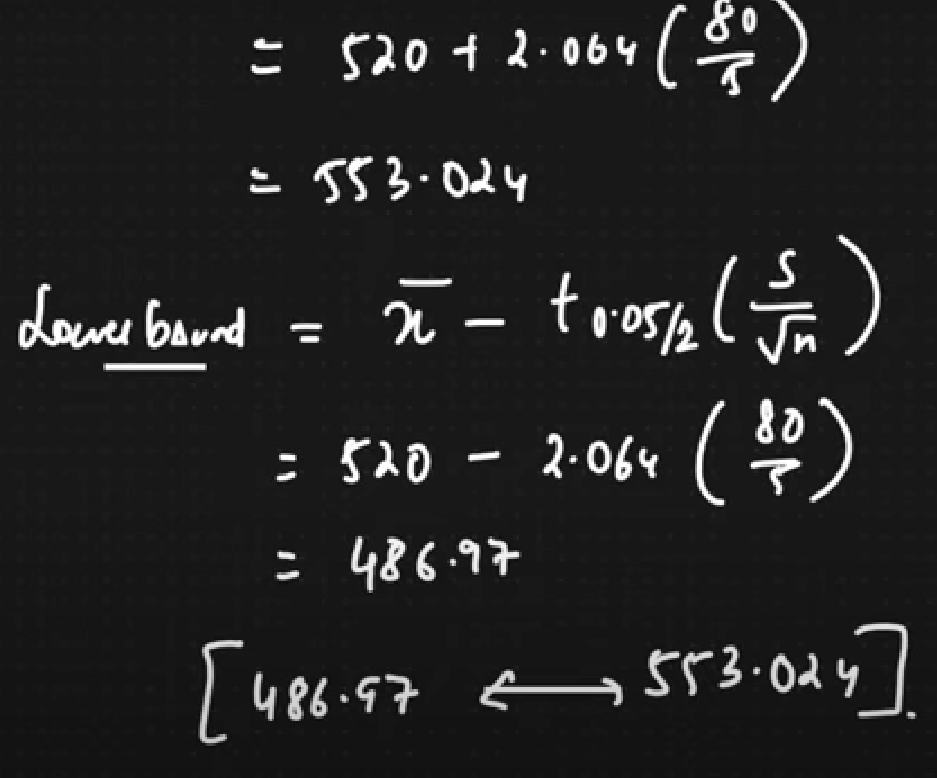
**What if population standard deviation is not given? Then we will use t-test.**

**Q: During the test of CAT exam, a sample of 25 students has a mean of 520 with standard dev of 80. Construct 95% CI about the mean.**

**A:** n=25 ; x\_bar = 520, s=80, alpha =0.05

Here pop std dev is not given, so we will use t test. Note that while using t test we need degree of freedom inorder to fetch value from t table. Degree of freedom = n-1 = 24





**One sample z test**

Now we will do z test. Conditions for z test is given above. We are doing Hypothesis testing.

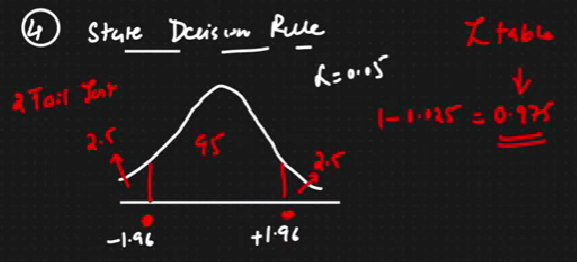
Q: In the population, avg IQ = 100 with std dev is 15. Researchers wants to test a new medicine to see if there is positive or negative effect on intelligence or no effect at all. A sample of 30 participants who have taken the medication has a mean IQ of 140. Did the medication affect the intelligence? Alpha=0.05 (means CI=95%)

A: 1) Define null hypo; H0 : mean =100

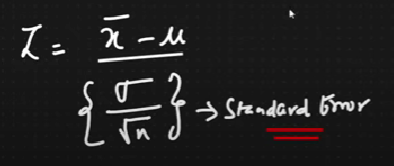
2) Alt hypo ; H1 : mean not equal to 100.

3) Define alpha

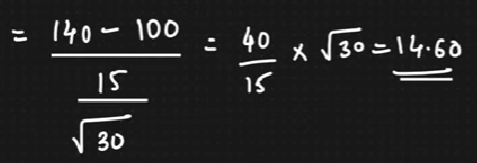
4) State decision rule . Here we need to see whether medication affect the intelligence, So it can increase or decrease. So two tail test.



5) Calculate test statistic. Here Z score.



Original formula of z score contain sigma/sqrt(n). If we are using population data we can ignore sqrt(n). Since we are processing on sample data, we need a standard error (which has obtained from experimentations.) Dividing by sqrt(n) reduces the standard error as the sample size keeps on increasing our mean values will be match with population.



6) State our decision : From decision rule we understood, if z value stands within CI we can accept H0. Here 14.6 > 1.96. So we will accept alternative hypothesis. But an important question is pending. Did the medication increase or decrease the intelligence. Answer is improved the intelligence cos z score is beyond upper CI.

**What about p-value?** Whenever we are rejecting null hypothesis, p value will be less than alpha. Here p<=0.05. Means our experiment result fall in outside CI region.

**Q: The avg weight of residents in Bangalore city is 168 pounds with a std dev of 3.9. We take a sample of 36 individuals and the mean is 169.5 pounds. Check whether the sample is saying that the weights are same. CI =95%**

**A**: we can do z test. Mu=168 ; sigma = 3.9; n=36; x\_bar = 169.5

1. Ho: mu=168
2. H1: mu not equals 168
3. Alpha=0.05
4. Decision boundary: Since it is a 2 tail test and alpha is 0.05, CI upper and lower fences are at -1.96 and 1.96 (z value at 1-0.025 = 0.975)
5. Z = 169.5-168 / 3.9/sqrt (36) = 2.307
6. Decision rule; z value is greater than 1.96; so we reject the null hypothesis

* Deriving p value from z value

Area under the curve is 1. We need to find the area of curve between z values (ie 2.307 and -2.307) from z table and we found the area is 0.9911. Now since it is a two tail test, each tail end has an area of 0.0044 (ie 1-0.9911= 0.0088 ; 0.0088/2 = 0.0044)

Now p value = 0.0044 + 0.0044 = 0.0088

We have already rejected null hypothesis. We can verify the same using p value as well. Here our p value < 0.05. So we can reject the null hypothesis.

**One sample t test**

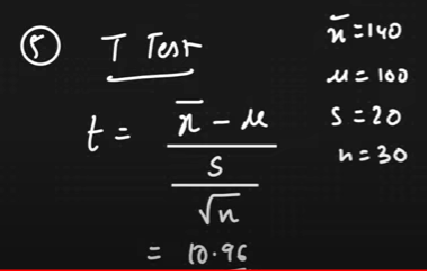
**Q: Population’s avg IQ – 100. Sample of participants, n=30 having a mean IQ =140 with a sample std dev =20. Did the medication affect intelligence? Alpha=0.05**

A: 1) Ho: Mean =100 ; H1: mean not equal to 100.

2) Calculate degree of freedom, n-1 = 29

3) Decision rule : Define the graph. Two tail test. From t table, considering degree of freedom and alpha of 0.025 (alpha will be 0.05 if its one tail test), we get CI boundaries as 2.045 and -2.045

4) Test statistic



T value > 2.045, so reject H0. So p value < significance value. So result may fall in either of two tail side. Since value is greater than upper threshold, test result is in rightside tail. So intelligence has increased.

**Real World problem** : A bank wants to open an ATM machine at a specific area. We have to formulate the problem and conduct hypothesis testing. We can assume values like eg: average money people taken from at machine.

**CHI SQUARE Test**

Chi square test claims about population proportions. It is a non parametric test that is performed on categorical (nominal or ordinal) data. Non parametric tests usually occurs wrt population proportion. Whenever you are given some proportions of data, we cant specifically use parametric kind of test, we have to go with non parametric test.

**Q:** In the 2000 census, the ages of individuals in a small town were found to be the following.

|  |  |  |
| --- | --- | --- |
| Less than 18 | 18-35 | >35 |
| 20% | 30% | 50% |

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

|  |  |  |
| --- | --- | --- |
| Less than 18 | 18-35 | >35 |
| 121 | 288 | 91 |

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

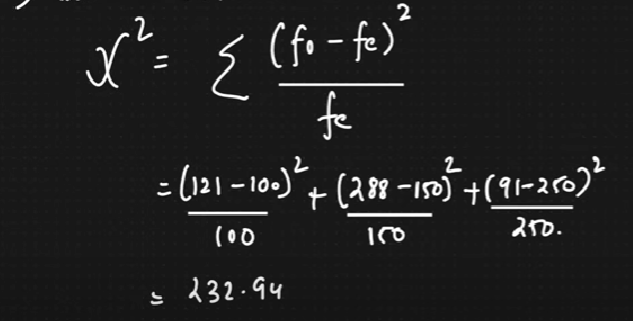
A: We have 3 categories.

|  |  |  |  |
| --- | --- | --- | --- |
| N=500 | Less than 18 | 18-35 | >35 |
| Observed | 121 | 288 | 91 |
| Expected based on 2000 sensus | 500 \* 0.2 =100 | 500\*0.3=150 | 500\*0.5 =250 |

1. Ho= Data meets the distribution of 2000 sensus.
2. H1= data doesn’t meet the distribution of 2000 sensus.
3. Alpha =0.05 (95% CI)
4. Degree of freedom; n-1= 2. [Here n is no of categories]
5. Find the decision boundary from chi square table using DOF and alpha value. Problem is a two tail test.

From the table we get value of 5.991. If chi square value is greater than 5.99, reject null hypo.

1. Calculate the test statistic. fo – observed and fe is expected.



X2 = 232.94 which is greater than 5.99, so reject null hypothesis.

**Pearson and Spearman correlation coefficient**

