From here we are going to start Inferential Statistics.

Agenda

1. Central Limit Theorem ✓
2. Z-score And Z stats ✓
3. Z-test, t-test } Solve problems Assignment

In inferential statistics we draw conclusion for population data using sample data. Hypothesis testing is one of the means to do that.

} Inferential stats

} Hypothesis Testing.

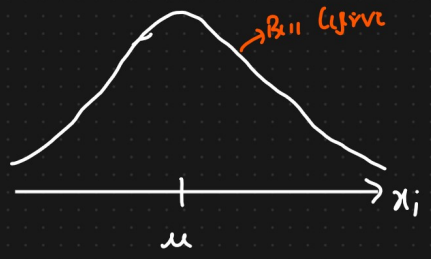
If I have a X (which is population) random variable that belongs to normal/gaussian distri^ with some mean and std dev and if we create multiple samples out of this(S1,S2,S3,...so on) in such a way that each of these sample have fixed sample size(n). Then we determine means for each of these samples and plot it then means of these samples will also follow normal/gausian distribution. This is called Central limit thereom.

1. Central Limit Theorem

1, 2, 3, 4, 5 ⇓

Note sample size can have any value h = sample size => any value

1. $X \approx N(\mu, \sigma)$



$S_1 = \{x_1\ x_2\ x_3 \cdots x_n\} = \bar{x}_1$ //

$S_2 = \{x_2, x_3, x_4, x_6 \cdots x_n\} = \bar{x}_2$ //

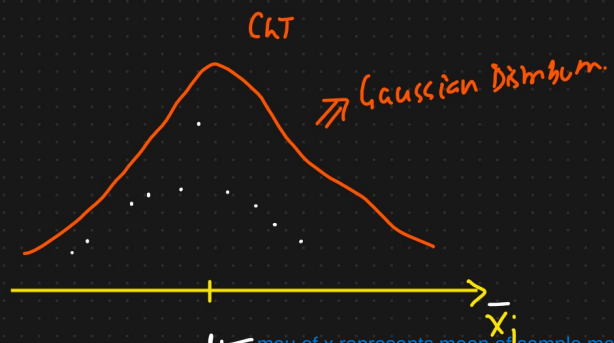
$S_3 = \{\ \} = \bar{x}_3$ //

$S_4 = \{\ \} = \bar{x}_4$ //

$S_5 = \{\ \} = \vdots$

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

means of sample



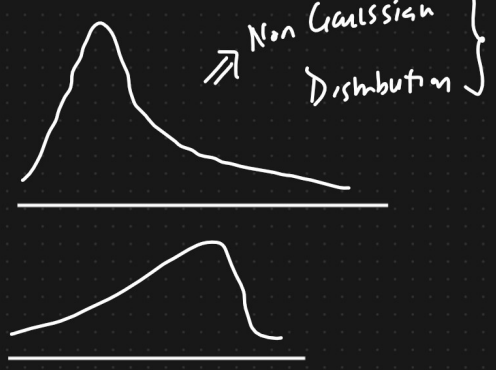
meu of x represents mean of sample means OR 'a mean' for means of different samples calculated above.

n = sample size

$n \geq 30$

In case of random variable X does not follow normal distribution then if plot sample means with sample size greater than or equal to 30 then distribution thus formed for means of sample will follow the normal distribution.

2. $X \not\approx N(\mu, \sigma)$



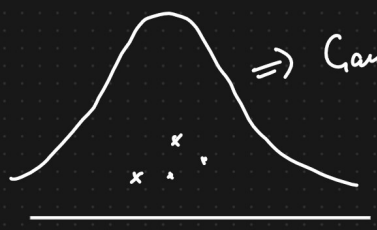
$S_1 = \{\ \} = \bar{x}_1$ //

$S_2 \quad = \bar{x}_2$ //

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

$S_4 \quad \vdots$

$\bar{x}_m$ //

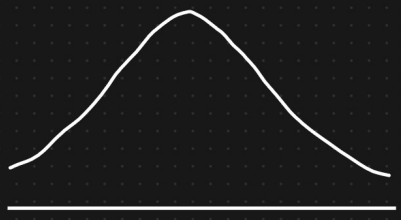


⇒ Gaussian/Normal Distribution

① Central Limit Theorem

The central limit theorem relies on the concept of a sampling distribution, which is the probability distribution of a statistic for a large number of samples taken from a population.

The central limit theorem says that the sampling distribution of the mean will always be normally distributed, as long as the sample size is large enough. Regardless of whether the population has a normal, Poisson, binomial, or any other distribution, the sampling distribution of the mean will be normal.



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

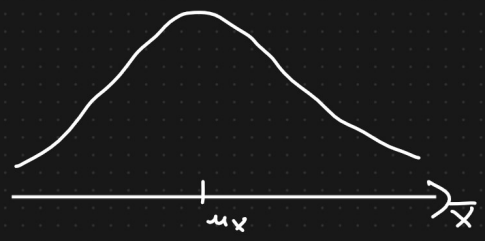
This represents population distribution

σ = population std

μ = population mean

This represents sample mean OR 'a mean' for means of samples distribution.

Sampling Distribution of the mean



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

n = Sample size

Acc to CLT sample mean will be approximately equal to population mean but sample standard deviation will be equal to population standard deviation divided by root of sample size which is denoted as standard error.

Below we will discuss about Z score and Z stats

① $X \sim N(4, 1)$

Zscore = (pt. where Zscore needs to be determined - mean)/std dev

Eg: Calculating Zscore at point 4.25 for normally distributed random variable X with mean 4 and std dev 1

$x_i = 4.25$

$$Z\text{-score} = \frac{4.25 - 4}{\sigma} = 0.25$$

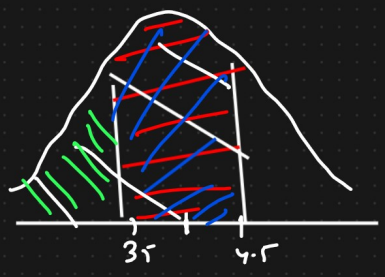
Z score is a pt. that basically tells that how much standard deviation away a pt. is from mean. This pt. is then used for calculating the area under the curve using Z table(Z table maps Zscore with area under the curve from left hand side). This curve is basically something whose internal area can be measured in terms of probability(PDF).

For Eg: Amazon manager asks to find the probability of any item to get delivered in 3 days. So here in this case we will plot histogram with kde=True where x axis represents number of days whereas y axis represents number of times that product get delivered on that particular day. After this we can simply calculate area under 4 th day using Z score and Z table and present to the Amazon manager.

(*) What percentage of score falls above 4.25?

$1 - 0.59871 = 0.4013 = 40.13\%$

(*) What percentage of score lies between 3.5 to 4.5?



$Z\text{-score} = \frac{4.5-4}{1} = +0.5 = 0.69146$

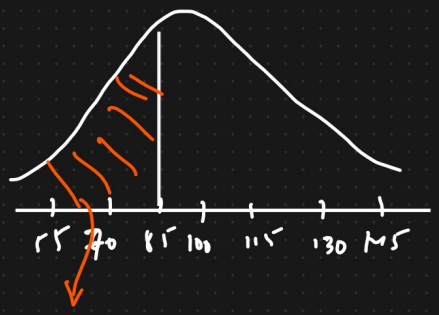
$Z\text{-score} = \frac{3.5-4}{1} = -0.5 = 0.30854$

$= 0.3829$

$= 38.29\%$

(*) In India the average IQ is 100, with a standard deviation of 15. What is the percentage of the population would you expect to have an IQ lower than 85?

Ans). $\mu = 100 \quad \sigma = 15$



$0.15866 = 15.86\%$

$Z\text{-score} = \frac{85-100}{15} = -1$

(A) $IQ \geq 85$

$1 - 0.1586 = 84.13\%$

$\boxed{Fr \geq IQ \leq 100} \Rightarrow$ Internal Assignment

① Hypothesis Testing And Statistical Analysis (Statistical analysis wrt inferential statistics.)

① Z test
② t test
③ CHI SQUARE
④ ANNOVA

Below we will try to achieve hypothesis testing using all these(Ztest,Ttest, chi square , Annova)

Note in numerical problems discussed below we are performing "1 sample Ztest and Ttest" since for these only 1 sample with some sample size is taken.

① Z_test

NOTE: Z test and T test always follows for Gaussian distribution.

①
1) The average heights of all residents (Population) in a city is 168cm. (different means > or < means 2 tail test) A doctor believes the mean to be diffuint. He measured the height of 36 individuals (Sample) and found the average height to be 169.5 cm. with a $\sigma = 3.9$

(a) State null and Alternate Hypothesis

(b) At a 95% confidence level, is there enough evidence to reject the null hypothesis.

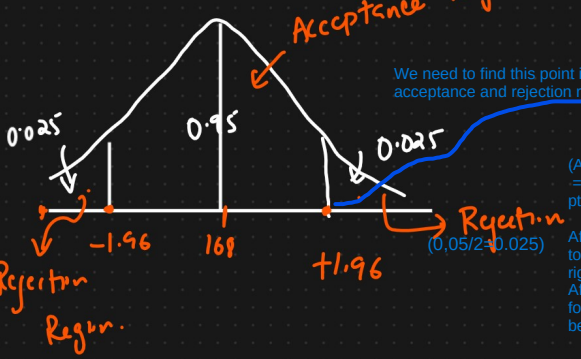
(In short in above question, using sample information we are trying to draw information for population using Z test. This is called inferential statistics).

Ans) $\mu = 168cm$ (Population mean) $\sigma = 3.9$ (Population dev) $n = 36$ (sample size) $\bar{x} = 169.5$ (sample mean) $C.I = 0.95$ (confidence interval) $\alpha = 1 - C.I = 0.05$ // (Significance value)

① Null Hypothesis H_0 : $\mu = 168cm$ (Default info given in problem)

Alternate Hypothesis H_1 : $\mu \neq 168cm$

② Decision Boundary based C.I.



We need to find this point in order to differentiate b/w acceptance and rejection regions

$1 - 0.025 = 0.9750$

(Area under curve - Area under right extreme region = area under right end rejection pt. which is Zscore pt.)

After this since we know the area we can use Ztable to determine Zscore which will be rejection pt. for right extreme end.
After knowing right rejection pt. since this test work for normal distribution so left rejection pt.(+1.96) will be symmetrical to right rejection region(-1.96).

2 Tailed Test

Since, in question it is said that population mean can be different so it can be either greater than 168 cm or less than 168 cm. That's why this lie under the category of 2 tailed test.

If, doctor would have said that he believes that population mean to be greater than 168 cm then this will cover only 1 extreme region and not both extreme regions as above(as in above case). Then this will fall under the category of 1 tailed test.

Similarly, if doctor says that he believes that population mean to be less than 168 cm then also this case will fall under 1 tailed test(since only 1 extreme region considered).

If Z test is less than −1.96 or greater than +1.96, Reject the Null Hypothesis

④ $Z\_test = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{169.5 - 168}{3.9/\sqrt{36}} = 2.31$

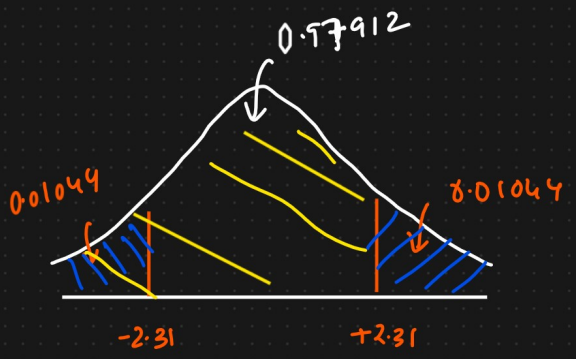
Sample mean, Population mean, Population std dev, root of sample size, (Standard error) {CLT}

NOTE: Sample size meaning is number of elements in a sample and NOT Number of samples.

Conclusion

2.31 > 1.96 Reject the Null Hypothesis.

Below is another way of determining whether accept or reject the null hypothesis using p value which is basically Probability value.



$1 - 0.98956 = 0.01044$

C.I = 99% Assignment

⑤ p Value = 0.01044 + 0.01044

= 0.02088

(Probability of distribution lying in rejection region for the points calculated based Zscore determined using Ztest)

$P < 0.05 \Rightarrow$ Reject the Null Hypothesis.

(Note here 0.05 is the default probability of rejection regions calculated based on CI and Significance level defined by domain expert)

Example ② A factory manufactures bulbs with a average warranty of 5 years with standard deviation of 0.50. A worker believes that the bulb will malfunction (malfunction) in less than 5 years. He tests a sample of 40 bulbs and find the average time to be 4.8 years.

(a) State null and alternate hypothesis

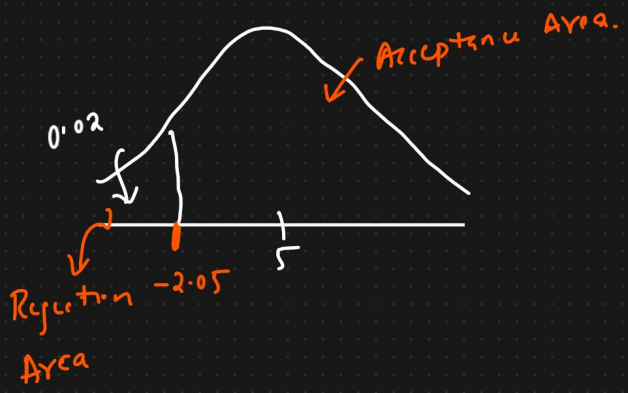
(b) At a 2% significance level, is there enough evidence to support the idea that the warranty should be revised?

Ans) $\mu = 5$ $\sigma = 0.50$ $n = 40$ $\bar{x} = 4.8$ years. C.I $= 0.98$ $\alpha = 0.02$

① H_0 $\mu = 5$

H_1 $\mu < 5$ {1 Tail Test}.

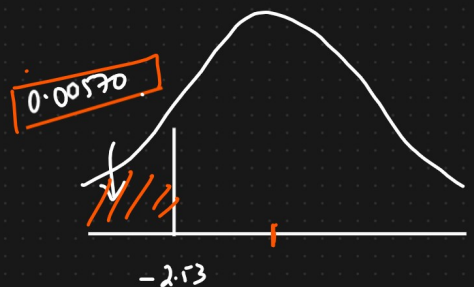
② Decision Boundary



If, $Z_{test} < -2.05$ then, Reject the Null hypothesis.

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

$-2.53 < -2.05$ => True => Reject the Null Hypothesis.



Pvalue

$0.00570 < 0.02$ (α)

95%

Reject the Null Hypothesis.

So, to conclude factory needs to revise the warranty of bulbs manufactured.
But what will be new warranty?

May be again testing of bulbs needs to done and based on outcome average warranty will be mentioned. Since, worker has already calculated warranty as 4.8 years on a sample of 40 bulbs this can be treated as new warranty. Or one can take even bigger sample of bulbs and perform testing. All this depends on domain expert means the people who are making bulbs or working in industry. Data scientist will not go and perform testing on bulbs to determine new warranty. Data scientist can only suggest whether to revise or not revise the warranty and its on domain experts to again perform testing of their product and come up with revised warranty if required. Data scientist just rely on info given by domain experts.

② T Test

Difference b/w Ztest and Ttest:
In Ztest we are provided with population standard deviation.
Whereas, in Ttest we are not provided population standard deviation. In Ttest we are provided with sample standard deviation.

// DATA ANALYST

Also, in Ttest we use Z table for determining cutoff points for rejection regions and area under the curve/p value.

Whereas, in case of T test we use T table for the same.

① In the population the average IQ is 100. A team of researchers want to test a new medication to see if it has either a positive or negative effect on intelligence, or no effect at all. A sample of 30 participants who have taken the medication has a mean of 140 with a standard deviation of 20. Did the medication affect intelligence? C.I = 95% $\alpha = 0.05$

S = sample std dev

Ans) ① $H_0 = \mu = 100$

$H_1 = \mu \neq 100$

$n = 30$ $\mu = 100$ $\bar{x} = 140$ $S = 20$ C.I = 95

$\alpha = 0.05$

② $\alpha = 0.05$

In Ttest as part of second steps after stating null and alternate hypothesis, we need to calculate Degree of Freedom. Reason why we calculate this(dof) is we are provided with sample std dev and not population std dev.

Degree of freedom

$dof = n - 1 = 30 - 1 = 29$ //.

(dof = sample_size - 1. Dof will be used in T table to find cutoff value for rejection regions)

3 people $3 - 1 = 2$

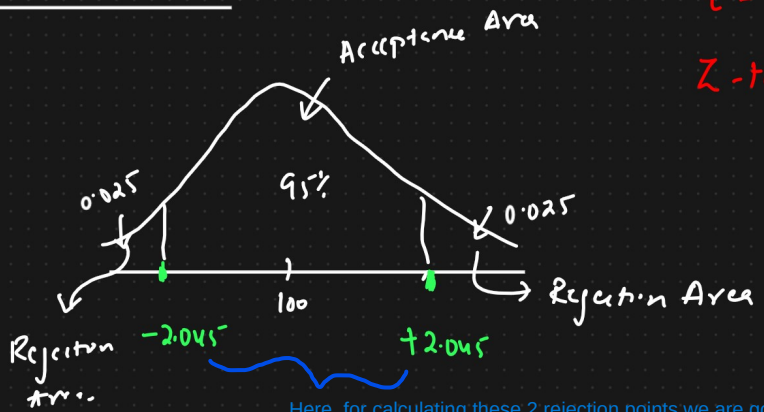
Explaining dof:
Suppose there are 3 people and 3 chairs in a room.
Choice 1: 1st person can choose b/w 3 chair to sit and will sit on one of the seat.
Choice 2: 2nd person can choose b/w 2 chairs and will sit on one of the seat.

For third person there will be no choice left and he/she has to sit on one remaining seat.
Since, there are 2 choices in total dof in this case will be 2.

③ Decision Rule

t-test ⇒ sample std.

Z-test ⇒ population std



From google/byjus:

The degrees of freedom of a system is the number of parameters of the system that may vary independently. For example, a point in the plane has two degrees of freedom for translation: its two coordinates ie; x and y coordinates.

Here, for calculating these 2 rejection points we are going to user T table and not a Z table.

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

In T table check;
- dof
- whether 1 or 2 tail
- p value/area under 1 or 2 table
- C.I

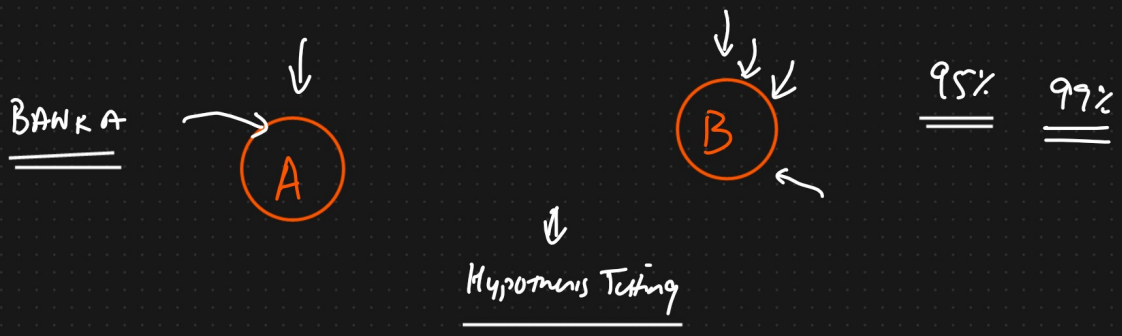
Based on above we calculate Tscore or cutoff value for rejection region in case of T test.

④ T Test statistics

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} = \frac{140 - 100}{20/\sqrt{30}} = \frac{40}{3.65} = 10.96$$

$t > 2.045$ ⇒ Reject the Null Hypothesis.

Conclusion : Medication has a +ve effect on intelligence.



2000

200 - 300 ⇒

There is "A" branch for some bank which has atm installed. As, a data analyst you are asked to perform hypothesis testing to determine whether it would be good to install another atm at a nearby bank's branch "B"?

So with the help of domain experts(tellers, cahiers, bank managers) we will collected data like mean transaction happening each day. And based of that will ask domain experts like what should be min and max transaction within which atm installation seems good and beyond which atm installation seems bad(CI and significance value). Using all these data analyst will perform hypotheses testing and tell whether to install or not to install atm at the near by location.

1. Before distributing covid vaccine one will perform hypothesis testing on its efficiency.

2. As a event manger one will perform hypothesis testing on number of plates of food to be ordered for all guests.
(Event manager tells take 50 plates extra food. +50 -50 this can be used to place C.I which is calculated from domain expert's knowledge).

3. HR wants to order T shirts for whole company. HR will not ask each employee about t shirt size. They will simply take a sample of people
and based on that will decide t shirt size of population and give order to manufacturer. But is this assumption going to be true or not? For this we need to perform hypothesis testing. For this HR needs to order first time and then based on feedback received need to judge that how much problem is there and then we need to perform hypothesis testing do determine new size of t shirts that needs to placed with manufacturer from next time.