STATISTICS ASSIGNMENT

- Neelima Bhavanasi

Problem Statement

Comprehension

The pharmaceutical company Sun Pharma is manufacturing a new batch of painkiller drugs, which are due for testing. Around 80,000 new products are created and need to be tested for their time of effect (which is measured as the time taken for the drug to completely cure the pain), as well as the quality assurance (which tells you whether the drug was able to do a satisfactory job or not).

Question 1:

The quality assurance checks on the previous batches of drugs found that — it is 4 times more likely that a drug is able to produce a satisfactory result than not.Given a small sample of 10 drugs, you are required to find the theoretical probability that at most, 3 drugs are not able to do a satisfactory job.

a.) Propose the type of probability distribution that would accurately portray the above scenario, and list out the three conditions that this distribution follows.

b.) Calculate the required probability.

Answer :1A)

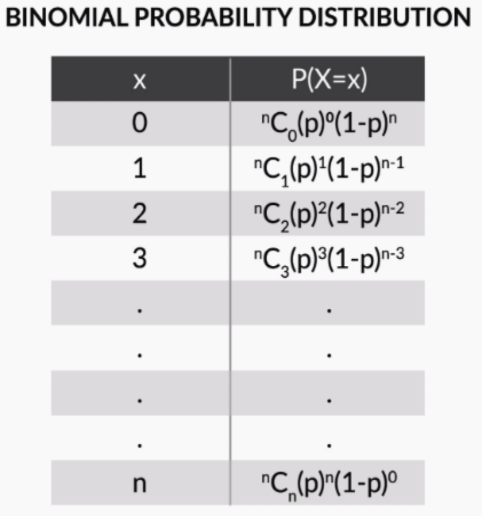
For the above scenario 'Binomial Distribution' is the best method to choose.

The Three Conditions the Binomial Distribution follows:

1. The number of trials should be fixed. (In this case the sample of drugs being fixed for 10)
2. Each trial is binary. (In this case the drug being effective and not effective.)
3. Probability of Success is the same in all trials and is denoted by P. (In this case the probability of success being the drug is effective)

Thus, satisfying all the conditions Binomial Distribution is the accurate method to follow. And then we have to calculate the cumulative probability of at most 3 drugs are not able to do a satisfactory job.

Binomial Probability Distribution:



BINOMIAL PROBABILITY DISTRIBUTION

x	P(X=x)
0	${}^{n}C_{0}(p)^{0}(1-p)^{n}$
1	${}^{n}C_{1}(p)^{1}(1-p)^{n-1}$
2	${}^{n}C_{2}(p)^{2}(1-p)^{n-2}$
3	${}^{n}C_{3}(p)^{3}(1-p)^{n-3}$
.	.
.	.
.	.
.	.
n	${}^{n}C_{n}(p)^{n}(1-p)^{0}$

Answer:1B) From the above statement: We know sample n = 10

Let us say:

If the probability of drug not likely to do a satisfactory job is denoted by = x

The probability of drug likely to do a satisfactory job is 4 times the drug not working it will be = 4x

Now we know the probability of success and failure = 1

i.e. x+4x=1

5x=1

x=0.2

Therefore P =0.2 (Probability of drug not doing a satisfactory job)

And (1-P)=0.8 (Probability of drug doing a satisfactory job.

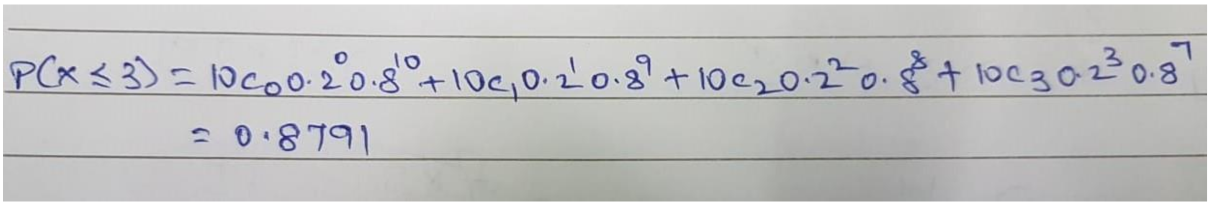
In mathematical terms, you would write cumulative probability **F(x) = P(X≤x)**.

For example, F (4) = P(X≤4), F (3) = P(X≤3).

The cumulative probability P(X<=3) = P (0) +P (1) +P (2) +P (3)

We know $P(X=r) = {}_{n}C_{r}(p)_{r}(1-p)_{n-r}$

Substituting the values for X=0,1,2,3, n=10



$$P(x \leq 3) = 10c_0 0.2^0 0.8^{10} + 10c_1 0.2^1 0.8^9 + 10c_2 0.2^2 0.8^8 + 10c_3 0.2^3 0.8^7$$
$$= 0.8791$$

Final Result:

Therefore, the probability of at most three drugs not doing a satisfactory job P(X<=3) = **0.8791.**

Question 2:

For the effectiveness test, a sample of 100 drugs was taken. The mean time of effect was 207 seconds, with the standard deviation coming to 65 seconds. Using this information, you are required to estimate the range in which the population mean might lie — with a 95% confidence level.

a.) Discuss the main methodology using which you will approach this problem. State all the properties of the required method. Limit your answer to 150 words.

b.) Find the required range

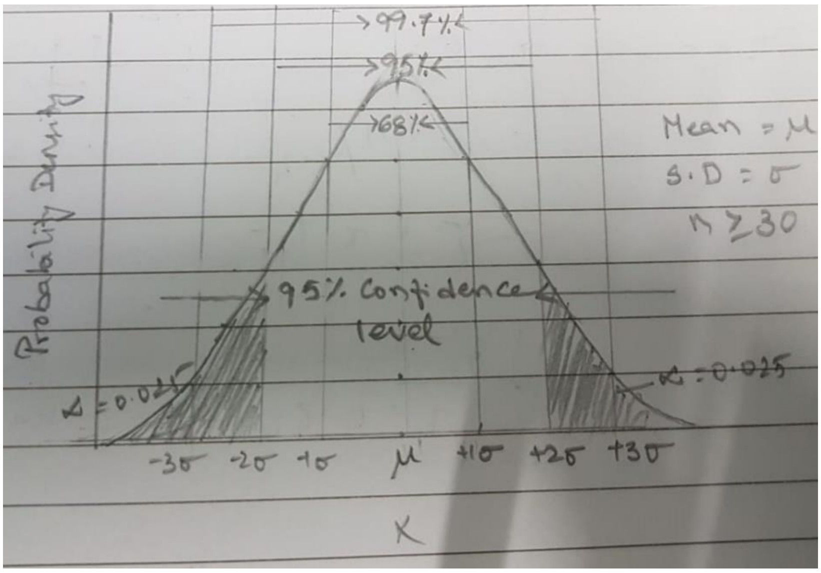
Answer: 2A)

By following **'Sampling Distribution'** by **'Central Limit Theorem',** we can approach the problem.

The properties required to follow the central limit theorem are:

1. Sampling Distribution Mean ($\mu_{\bar{x}}$) =Population Mean(μ)
2. Sampling Distribution Standard deviation (standard error) = $\sigma/\sqrt{n}$
3. For n<=30, the sampling distribution becomes a normal distribution.

Sample Distribution for a two tailed test:



Answer 2B)

From the problem statement we know:

n= 100 (sample size)

µ=207 sec (sample mean)

σ = 65sec (standard deviation)

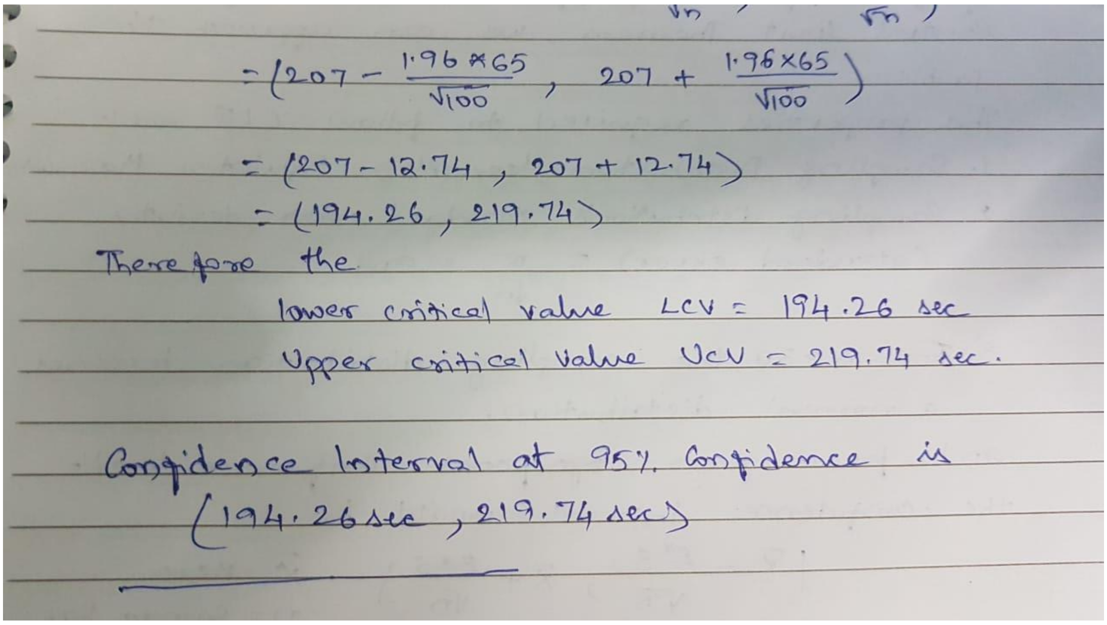
So, at 95% confidence level the Z* value is 95/100+(1-95/100)/2 = 0.975

and the Zscore associated with value 0.975 from the Z table is 1.96.

so, Z*=1.96

with a **sample size n, mean** $\bar{X}$ and **standard deviation S**. and with a **y% confidence interval** (i.e., a confidence interval corresponding to a y% confidence level) for μ will be given by the range:

Confidence interval = $(\bar{X} - \frac{Z^*S}{\sqrt{n}}, \bar{X} + \frac{Z^*S}{\sqrt{n}})$,



$\sqrt{n}$, $\sqrt{n}$)

$= (207 - \frac{1.96 * 65}{\sqrt{100}} , 207 + \frac{1.96 \times 65}{\sqrt{100}})$

$= (207 - 12.74 , 207 + 12.74)$

$= (194.26 , 219.74)$

There fore the

lower critical value LCV = 194.26 sec

Upper critical value UCV = 219.74 sec.

Confidence Interval at 95% Confidence is

(194.26 sec , 219.74 sec)

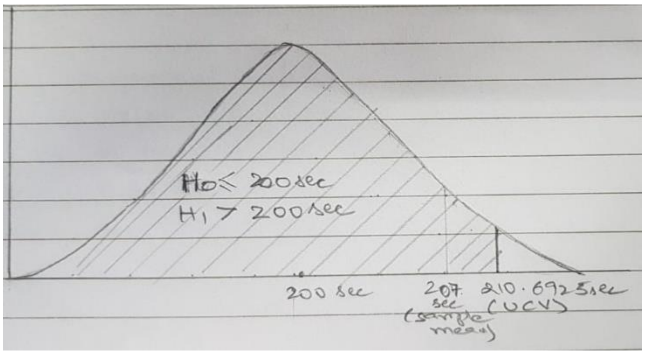
Final Result:

1. There fore the lower critical value LCV is 194.26 sec
2. Upper critical value UCV is 219.74 sec.
3. **Confidence interval at 95% is (194.26 sec, 219.74 sec)**

Question 3:

a) The painkiller drug needs to have a time of effect of at most 200 seconds to be considered as having done a satisfactory job. Given the same sample data (size, mean, and standard deviation) of the previous question, test the claim that the newer batch produces a satisfactory result and passes the quality assurance test. Utilize 2 hypothesis testing methods to make your decision. Take the significance level at 5 %. Clearly specify the hypotheses, the calculated test statistics, and the final decision that should be made for each method.

Answer 3A) Testing Method 1: By **Critical Value Method:** From the given problem statement, we understand that:



Null Hypothesis Ho <= 200 sec.

Alternate Hypothesis H1>200sec.

And therefore, it is an Upper tailed test or Right tailed test.

And we understood from the data:

Sample mean μ =207 sec

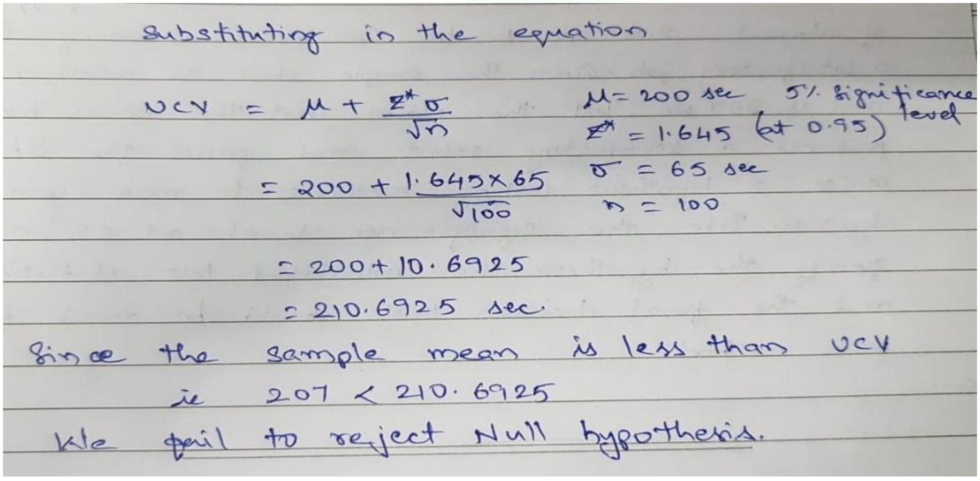
Standard Deviation σ = 65 sec

Sample size n = 100

Z* at 5% significance level: α = 0.05

Since this is a one tailed test Z score at (1-.05) = 0.95 from the Z table is 1.645.

Substituting in the equation for UCV = *μ ± Zc x (σ/√n)*



Substituting in the equation

$$UCV = \mu + \frac{Z^* \sigma}{\sqrt{n}}$$

μ = 200 sec

Z* = 1.645 (at 0.95) 5% significance level

σ = 65 sec

n = 100

$$= 200 + \frac{1.645 \times 65}{\sqrt{100}}$$

$$= 200 + 10.6925$$

$$= 210.6925 \text{ sec.}$$

Since the sample mean is less than UCV

ie 207 < 210.6925

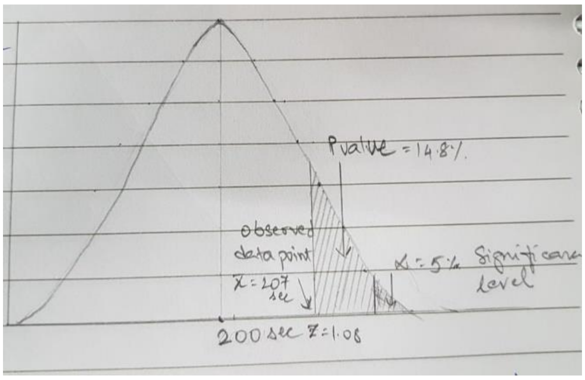
We fail to reject Null hypothesis.

We get **UCV** as **210.6925 sec.** Since 207sec is less than 210.6925 sec (UCV), we Fail to reject the null Hypothesis.

Final Decision: Fail to Reject the Null Hypothesis.

Question 3A continued:

Testing Method 2: By P Value Method: From the statement problem we understand that:



Null Hypothesis H0<=200sec

Alternate Hypothesis H1>200sec

And this is a Right tailed test or Upper tailed test.

Sample mean µ =207 sec

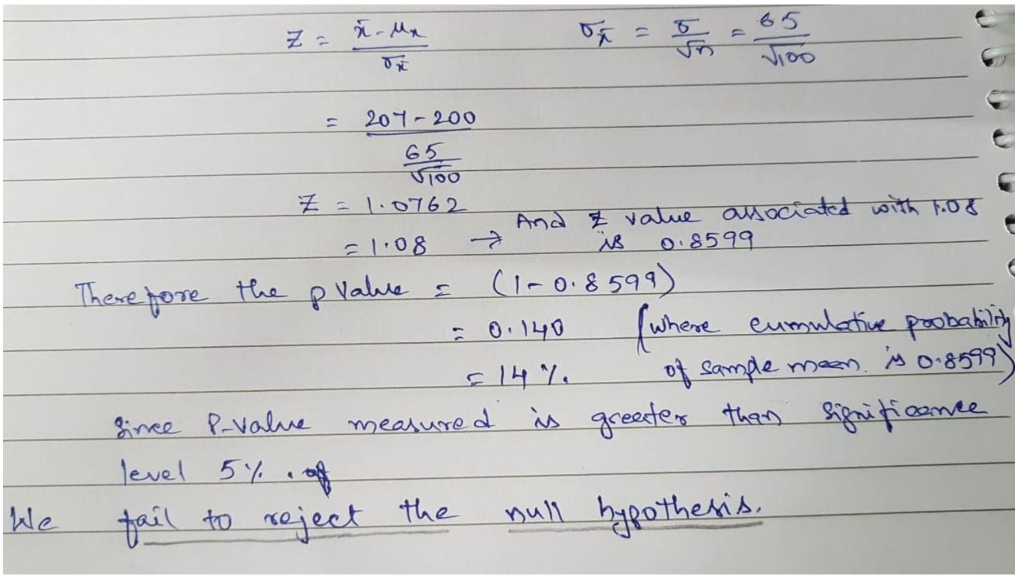
Standard Deviation σ = 65 sec

Sample size n = 100

To calculate the Z score for the observed mean is:

Z= $(\bar{x} - \mu) / (\sigma / \sqrt{\boxed{n}})$

Substituting those values in the above equation:



$$Z = \frac{\bar{x} - \mu_x}{\sigma_{\bar{x}}} \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{65}{\sqrt{100}}$$

$$= \frac{207 - 200}{\frac{65}{\sqrt{100}}}$$

$$Z = 1.0762$$

$$= 1.08$$ → And Z value associated with 1.08 is 0.8599

Therefore the p Value = $(1 - 0.8599)$

$= 0.140$ (where cumulative probability of sample mean is 0.8599)

$= 14\%$

Since P-Value measured is greater than significance level 5%. ~~of~~

We fail to reject the null hypothesis.

We got the **P value as 0.140 or 14%** and since the P value is less than significance level α = 0.05 or 5%. We conclude that we Fail to Reject the Null Hypothesis.

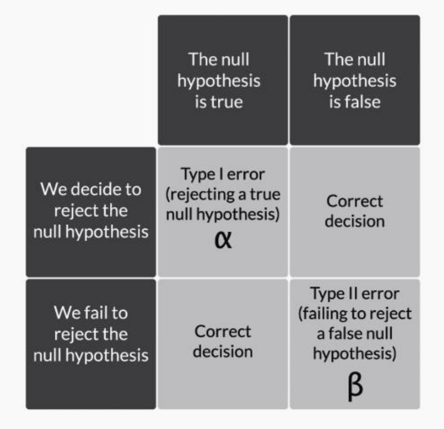
Final Decision: Fail to Reject the Null Hypothesis.

Question 3:

b) You know that two types of errors can occur during hypothesis testing — namely Type-I and Type-II errors — whose probabilities are denoted by α and β respectively. For the current sample conditions (sample size, mean, and standard deviation), the value of α and β come out to be 0.05 and 0.45 respectively.

Now, a different sampling procedure (with different sample size, mean, and standard deviation) is proposed so that when the same hypothesis test is conducted, the values of α and β are controlled at 0.15 each. Explain under what conditions would either method be more preferred than the other, i.e. give an example of a situation where conducting a hypothesis test having α and β as 0.05 and 0.45 respectively would be preferred over having them both at 0.15. Similarly, give an example for the reverse scenario - a situation where conducting the hypothesis test with both α and β values fixed at 0.15 would be preferred over having them at 0.05 and 0.45 respectively. Also, provide suitable reasons for your choice (Assume that only the values of α and β as mentioned above are provided to you and no other information is available).

Answer:



	The null hypothesis is true	The null hypothesis is false
We decide to reject the null hypothesis	Type I error (rejecting a true null hypothesis) α	Correct decision
We fail to reject the null hypothesis	Correct decision	Type II error (failing to reject a false null hypothesis) β

Type 1 Error: represented by α, occurs when you reject a true null hypothesis.

Type 2 Error: represented by β, occurs when you fail to reject a false null hypothesis.

In our Case:

Type 1 Error: Saying drug is not effective but in reality, it is effective with in 200 sec

Type 2 Error: Saying drug is effective but in reality, the drug is ineffective.

Scenario 1: α=0.05 and β=0.45 is preferred over α=0.15 and β=0.15

With Type 2 error being so high at 0.45 compared to 0.15 this scenario is **not recommended**, the company would be releasing more low quality drugs which would harm the patients as well as cause more damage to the branding and reputation of the company.

Scenario 2: α=0.15 and β=0.15 is preferred over α=0.05 and β=0.45

In case of β=0.15 compared to β=0.45, the company will be sending better quality medicines to the market all the time and consumer confidence on the company and the drug would be higher. Since **α** is higher by .1 compared to **α** being 0.05 it has higher risk of rejecting valid medicines.

Question 4: Now, once the batch has passed all the quality tests and is ready to be launched in the market, the marketing team needs to plan an effective online ad campaign to attract new customers. Two taglines were proposed for the campaign, and the team is currently divided on which option to use. Explain why and how A/B testing can be used to decide which option is more effective. Give a stepwise procedure for the test that needs to be conducted.

Answer:

Why and how A/B testing can be used:

A/B testing helps to increase the efficiency of the product by introducing a variation to the existing feature and observe which is performed best, in this scenario two Taglines were proposed and to observe for which tagline the users are attracted to.

Total No of Clicks	Group A	Group B
Tagline 1 Clicks	1000	
Tagline 2 Clicks		500

The following is an A/B testing framework to start running tests:

Set goal ⇨ Decide what to test ⇨ Create Variations ⇨ Run Test ⇨ Analyse Your Result ⇨ Repeat

1. Test Sample Size: Create two sample groups Group A and Group B of same size

2. Create Hypothesis: Generate a hypothesis with significance level (number of clicks) for the test that need to be done.

3. Create Variation: Create variation for a page with Tagline1 and Tagline2 respectively.

4. Test the Variation: Sample proportion test is done. Find the P value and see weather the p value is less than significance or greater than significance and decide weather null hypothesis should be rejected or fail to reject.

5. Observations: Tagline1 has more clicks than Tagline 2.