Statistics Assignment

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:

a. Here our scenario has small sample of 10 drugs and we have to find the at most 3 drugs not doing satisfactory job which is equivalent to P(X)<=3. Here this scenario follows Binomial distribution and whenever a random variable takes the value less than or equal to 'X' then it follows Cumulative Probability distribution also.

Properties of Binomial Distribution:

i. Total number of trials should be always fixed.
ii. It should be having only two possible outcomes either Success or Fail, yes or no.
iii. Probability of success should be same in all the experiments/trials.

So, here total number of trials is 10 and it is fixed. It is having only two outcomes whether drug is doing satisfactory job or not and the probability of doing the satisfactory job is also same in all the experiments. So, we can implement binomial distribution to this problem.

b. Let, the drug producing not satisfactory result = x

From the question it is given that drug is 4 times more likely produce satisfactory result than not = 4x.

As we know that total probability of satisfactory and not satisfactory result should equals to 1.

P(Satisfactory result) + P(Not Satisfactory result) = 1

$4x + x = 1$

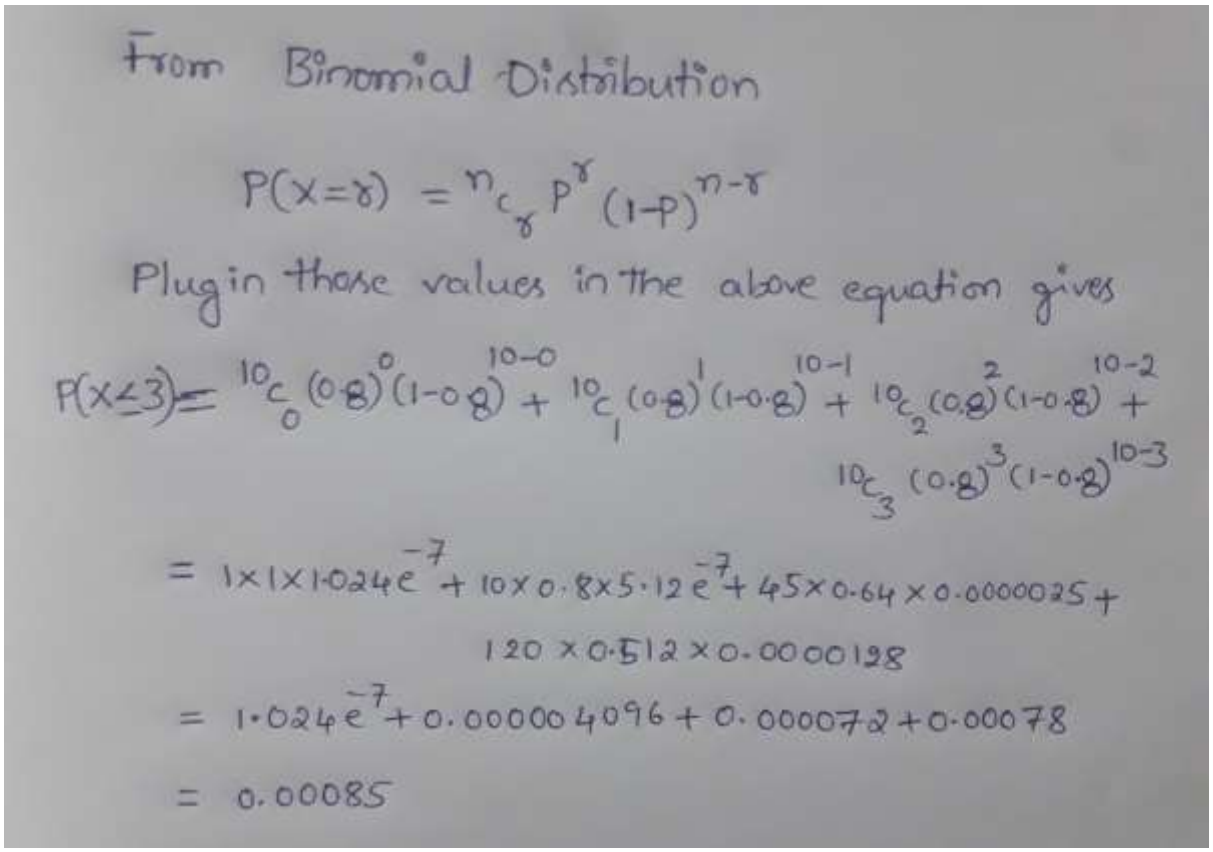
$5x = 1$

$x = 1/5 \Rightarrow 0.2$

So, P(Not Satisfactory) = 0.2 and **P(Satisfactory)=> 4x => 4*0.2 =0.8**

Given number of drugs (n)= 10, P(X<=3) = ?

P(X<=3) = P(X=0)+P(X=1)+P(X=2)+P(X=3)



So, Probability that at most, 3 drugs are not able to do a satisfactory job is **0.00085.**

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:

a. Here we will use Confidence Interval concept from Central Limit Theorem to solve this problem.

Properties of Central Limit Theorem:

i. Sampliing distribution mean(Xbar) = Population mean(μ)
ii. Sampling distribution's standard deviation(Standard Error) = σ/sqrt(n)
iii. For sample size of n> 30, the sampling distribution follows normal distribution.

Population mean is always lie between ± Margin of error in sample. To get that margin we have to find with the help of Confidence Intervals.

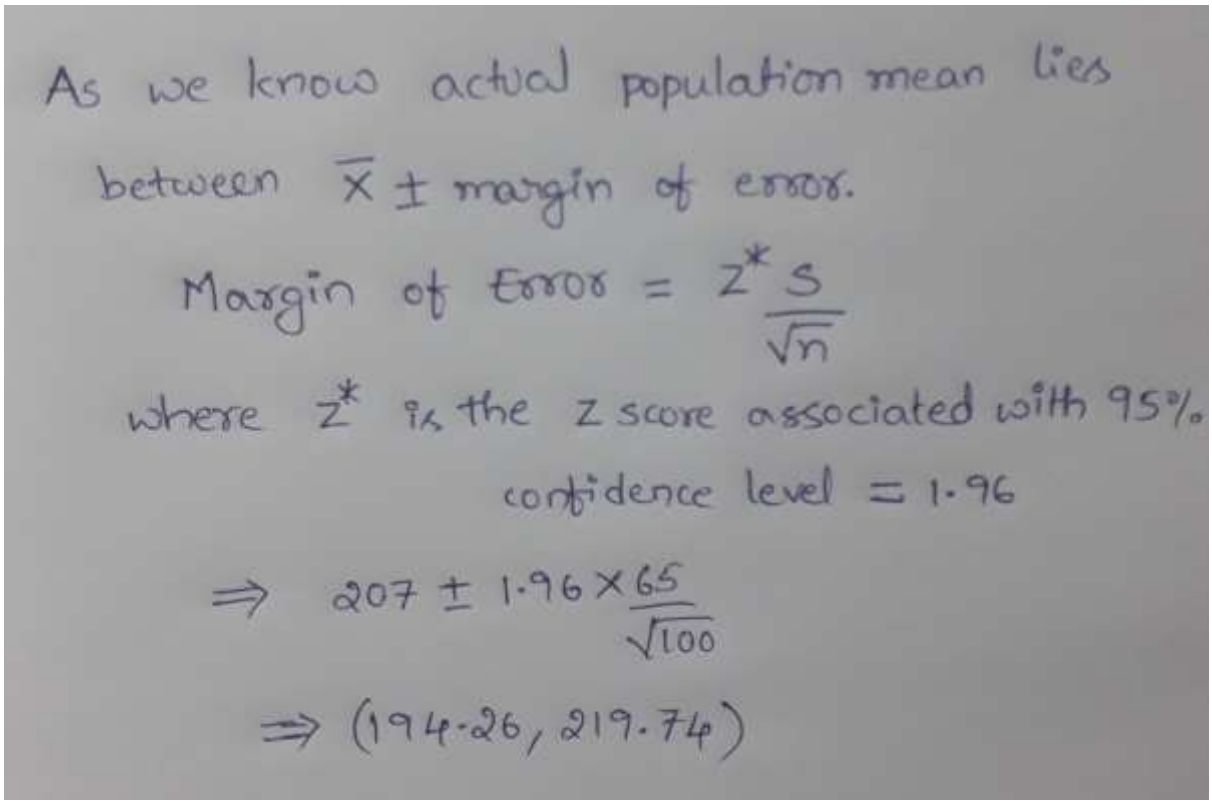
Like 1-2-3 rule of Normal Distribution there are standard values for Standard Normal Variable (Z-score).

For example, Z-score of 90% confidence level is equal to 1.65 which is denoted by Z*

b. Given, Sample of drugs taken, n = 100

Mean time effect (Xbar) = 207 seconds

Standard Deviation (S) = 65 seconds



Hence, the range in which population mean might lie with 95% confidence interval would be (207= (**194.26 to 219.74**).

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:

From the above problem we have,

Pain killer drug considered as doing a satisfactory job when time of effect of at most 200 seconds.

No. of samples (n) = 100

Mean of the sample (Xbar) = 207

Sample standard deviation(S) = 65

Confidence level = 5%==> 0.05

Method 1 (Critical Value Method):

Steps to do Critical Value method:

Step1: Formulate Hypothesis

Step2: Calculate Zc value from the given significance level (α)

Step3: Find the Upper Critical Value (UCV) and Lower Critical Value(LCV)

Step4: Identify the Mean lies inside or outside Critical Region to Reject or Fail to Reject Null Hypothesis.

Implementation:

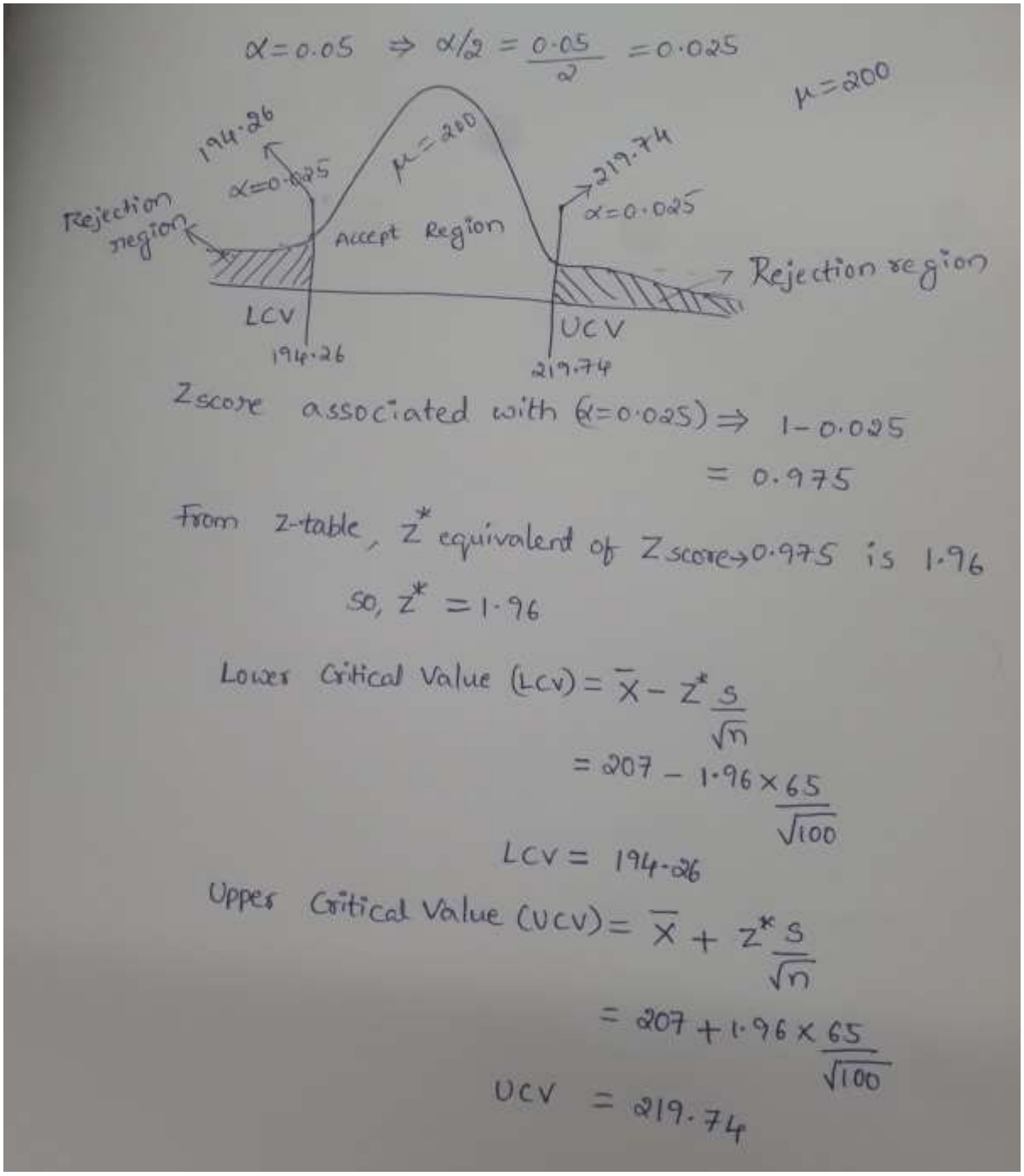
Null Hypothesis (H0) = Drug is doing Satisfactory Job (Time of Effect(μ) = 200 Seconds)

Alternative Hypothesis (H1) = Drug is not doing Satisfactory Job (Time of Effect(μ) ≠ 200 Seconds)

Given that Significance Level(α) is 5% which is nothing but 0.05.

As it is two-sided test significance level(α) lies on both sides of the distribution.

α/2 = (0.05)/2 = 0.025



Finally, LCV = 194.26 and UCV = 219.74.

From the problem statement it is clear that time to do satisfactory job is 200 seconds. So, μ=200 and 200 lies between LCV and UCV. So, we are **"Failed to reject Null Hypothesis"** which means **Drug is doing satisfactory job** on 95% of the times.

Method2(P-Value Method):

Steps to calculate P-value:

1. Formulate Hypothesis
2. Calculate Zscore value using sample mean and distribution mean
3. Find the P-value from the Zscore using Z-table
4. Take the decision based on the result.

Implementation:

Null Hypothesis (H0) = Drug is doing Satisfactory Job (Time of Effect(μ) = 200 Seconds)

Alternative Hypothesis (H1) = Drug is not doing Satisfactory Job (Time of Effect(μ) ≠ 200 Seconds)

Given that Significance Level(α) is 5% which is nothing but 0.05.

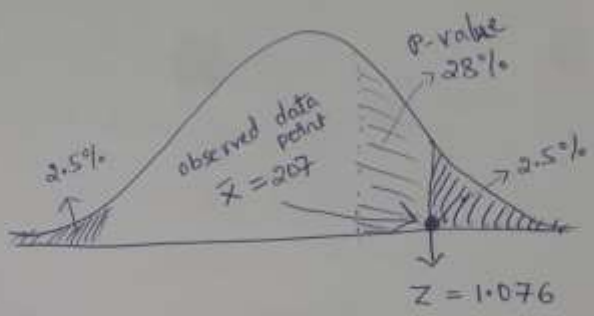
We have sample mean(Xbar) = 207 and μ=200

Standard deviation (S) = 65

$$Z_{score} = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

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

$$Z_{score} = 1.076$$



As the Z_{score} is positive

we consider positive Z-Table

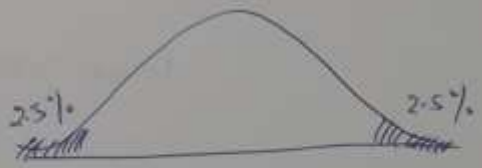
From Z-Table, $Z_{score} = 1.076 = 0.857$

$$P\text{-value} = 1 - 0.857$$

$$= 0.143$$

But from Hypothesis statement

It is a Two tailed Test



So $P\text{-value} = 0.143 \times 2$

$P\text{-value} = 0.28 \Rightarrow 28\%$

$28\% > 5\% \Rightarrow$ P-value $> \alpha$

We are failed to Reject NULL Hypothesis.

So, finally we have P-value=0.28 and α=0.05. As P> α, We are **"Failed to Reject Null Hypothesis"** which means **drug is doing satisfactory** job on 95% of the times.

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:

Case-1: The value of **α** and **β** come out to be 0.05 and 0.45 respectively.

Consider a scenario where Customers are claiming their insurance with any Insurance company. There are situations that fraud may happen at anywhere in the claim. To accept the fraud then Insurance company must set a margin of error to handle.

Ho : Insurance claimed are not having Fraud

H1 : Insurance claimed are having Fraud.

So here 'α' should be low so that they can set low error margin to accept the fraud and company will be not under loss.

Type-I error occurs here when we are **Rejecting our Null Hypothesis** when it is **True.**

If there is a Type-I error means we believe that there is a fraud in the claimed insurances by the customer even though it is not. Due to this company need to re-check those insurance claims which is giving redundant work. Here the loss spare by the company is very less they only need to spend more time to cross check the policies having fraud or not.

Type-II error occurs when we are **Failed to Reject Null Hypothesis** when it is **False**.

If there is a Type-II error means we believe that there is no Fraud in the claimed insurances by the customer even though there is a Fraud lies in it. Here company must bear the loss from the claimed insurances by customer.

So, from the above two cases Type-II error is dangerous and need to minimize it as its cost's company's profit.

Mainly in this scenario we have to minimize the **α** value so that **β** increase as they are inversely in relation to each other, hence the error margin will be less. If both **α** and **β** values fixed at 0.15 then we are accepting 15% of error margin unlike in above statement where we accept only 5% error margin. So, fixed values of both **α** and **β** at 0.15 is not correct in this scenario.

Case-2: Both α and β values fixed at 0.15

Example 1:

Consider a scenario where banks are giving loans to the customers. Interest from different types of loans are the major source of income for banks. Now if bank has offer loan for education with interest rate of 20%-30% per year then the students really don't pay much

attention to it since they think interest rate is too high. If banks reduced that interest rate in such a way that it is economical to every middle-class family, then there are high chances of getting more applications from the students. So, in this case we must balance both α and β to attract students and at the end bank must attain profit else if α and β are different then for Type-I error banks may get profit and students will get loss. And for Type- II error banks will get huge loss and students gets good profit.

Example2:

Consider another scenario where Manufacturing company produces iphones. There is a high chance that there will be defective pieces of iphones are produced during production and company must set a margin(α) to minimize the defective pieces.

Ho : Phones are not defective

H1 : Phones are defective.

Type-I Error occurs when phones are not defective but we are saying phones are defective.

When this type of error occurs then company needs to re-check the phones functionality but it doesn't give much trouble.

Type-II Error occurs when phones are defective but we are saying phones are not defective.

Where as when Type-II error occurs then we are saying that phones are not defective even though they are defective. So, the moment when company sold those mobiles to customers then they will obviously report the complaints about malfunctioning of the phones to the company. Here it spoils the company's brand due to the malfunctioning of the phones.

Now Consumer has to purchase from manufacturer and again consumer don't want defective phones.

In this scenario α and β should be in the balanced level like in our question for example at 0.15 then both manufacturer and consumer get the profit with less error else only one will get the profit and other will get the loss.

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:

A/B testing is very useful test to identify the attractive options from the online web platform. A/B testing is also call **Two-Proportion Test.** It mainly used in the e-commerce platform where they need to identify the old website is good or not, new feature is attractive or not, website is user friendly or not, customers are spending time or not etc.

Steps to do Two-Proportion Test:

1. Formulate the Hypothesis
2. Identify the frequency and sample size from two groups
3. With the help of XLSTAT package from excel select two proportion test and enter the frequency and sample size with alternate hypothesis.
4. Give the significance level as per the business requirement.
5. Find the P-value and take decision.

This process needs to be repeated for few days on daily basis. After collecting the data and if we have enough confidence that the data, we have is sufficient to take the decision based on the A/B test then we can finalize the call about the new features are really attractive to the users or not.

A/B Testing is really helpful as users doesn't spend too much time to understand the website and they easily diverted to other sites. So, if we identify the problems and their interests as quickly as possible then it helps company to attain profits in a short span.