

Statistics Assignment

Problem Statement

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.

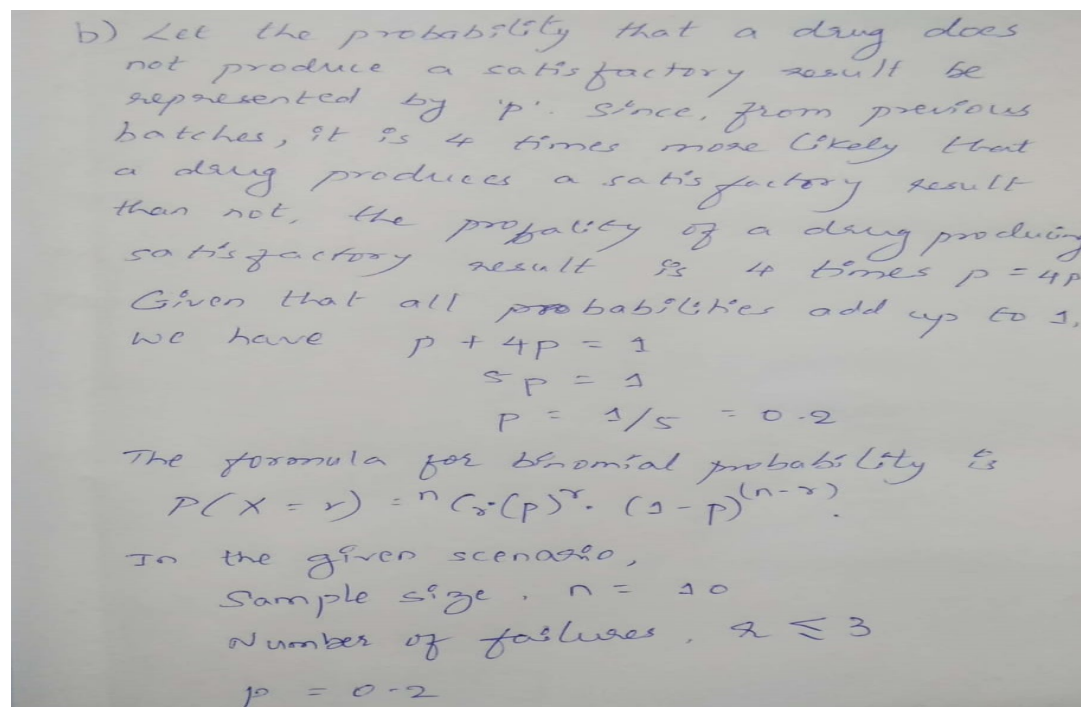
- Propose the type of probability distribution that would accurately portray the above scenario, and list out the three conditions that this distribution follows.
- Calculate the required probability.

Answer:

- The scenario mentioned above follows a **“Binomial Probability Distribution”**.

The three conditions that any Binomial Probability Distribution are:

- The total number of trials is fixed. In the given scenario, it is given that a sample of 10 drugs is taken. Hence, this condition is satisfied.
- The outcome of each trial is binary i.e., the outcome can take only 2 possible values like Success/Failure, Pass/Fail etc. In the given scenario, the possible outcomes of the quality assurance checks will be either the drug is **“Satisfactory”** (meaning drug is **“Passed”**) or **“Unsatisfactory”** (meaning drug is **“Failed”**). Since the outcome of the checks is binary (**“Passed”** or **“Failed”**). Hence this condition is satisfied.
- The probability of success is same in all the trials. It is given that the probability that the drug is **“Satisfactory”** is 4 times more than the probability that the drug is **“Unsatisfactory”**. All the trials would have the same probabilities and hence this condition is satisfied.



b) Let the probability that a drug does not produce a satisfactory result be represented by 'p'. Since, from previous batches, it is 4 times more likely that a drug produces a satisfactory result than not, the probability of a drug producing satisfactory result is 4 times $p = 4p$.
Given that all probabilities add up to 1, we have $p + 4p = 1$
 $5p = 1$
 $p = 1/5 = 0.2$
The formula for binomial probability is
 $P(X = r) = {}^n C_r (p)^r \cdot (1-p)^{(n-r)}$
In the given scenario,
Sample size, $n = 10$
Number of failures, $r \leq 3$
 $p = 0.2$

Since we need to find the cumulative probability that at most 3 drugs are not able to do a satisfactory job, we need to add probabilities of $P(X=0)$ through $P(X=3)$.

$$\text{i.e., } P(X \leq 3) = P(X=0) + P(X=1) + P(X=2) + P(X=3)$$

$$P(X=r) = {}^{10}C_r \cdot (p)^r \cdot (1-p)^{(n-r)}$$

$$P(X=0) = {}^{10}C_0 \cdot (0.2)^0 \cdot (1-0.2)^{(10-0)}$$

$$= (1) \cdot (1) \cdot (0.8) = \underline{0.1}$$

$$P(X=1) = {}^{10}C_1 \cdot (0.2)^1 \cdot (1-0.2)^{(10-1)}$$

$$= (10) \cdot (0.2) \cdot (0.134) = \underline{0.268}$$

$$P(X=2) = {}^{10}C_2 \cdot (0.2)^2 \cdot (0.8)^8$$

$$= (45) \cdot (0.04) \cdot (0.168) = \underline{0.302}$$

$$P(X=3) = {}^{10}C_3 \cdot (0.2)^3 \cdot (0.8)^7$$

$$= (120) \cdot (0.008) \cdot (0.209) = \underline{0.20}$$

$$\therefore P(X \leq 3) = P(X=0) + P(X=1) + P(X=2) + P(X=3)$$

$$= 0.1 + 0.268 + 0.302 + 0.20$$

$$\boxed{P(X \leq 3) = \underline{0.87}}$$

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.

- 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.
- Find the required range.

Answer:

- As per “**Central Limit Theorem**”, for any data if a large number of samples are taken, the sampling distribution will follow the three properties mentioned below
 - Sampling Distribution Mean ($\mu_{\bar{X}}$) equals the Population Mean (μ)
 - Standard Error which is the Sampling Distribution’s Standard Deviation is given by
Standard Error (S.E) = σ / \sqrt{n}
 $\sigma \rightarrow$ Population Standard Deviation. If the Population Standard Deviation is not known, then the Sample Standard Deviation can be approximated to be equal to Population Standard Deviation.
 $n \rightarrow$ Sample Size
 - When the sample size $n > 30$, the sampling distribution becomes a normal distribution.

In the given case, the number of samples taken is 100 and hence the sampling distribution follows a normal distribution. We can estimate the population mean with a confidence level of 95% using the “**Central Limit Theorem**”.

- Population Mean estimated range with 95% Confidence Level

The following details are provided regarding the sample:

Sample size $n = 100$.

Sample mean time $\bar{X} = 207$ seconds of effect

Sample standard deviation $S = 65$ seconds

Confidence Level = 95%.

Confidence interval for the population mean is given by the formula

$$\text{Confidence Interval} = \left(\bar{X} - \frac{z^* S}{\sqrt{n}}, \bar{X} + \frac{z^* S}{\sqrt{n}} \right)$$

where z^* is the z-score associated with the required confidence level.

$\frac{z^* S}{\sqrt{n}} \rightarrow$ margin of error.

The z^* value associated with 95% confidence level is ± 1.96

Plugging in the values in the confidence interval formula we get

$$\text{margin of error} = \frac{z^* s}{\sqrt{n}} = \frac{1.96 \times 65}{\sqrt{100}} \\ = 12.74$$

$$\text{C.I } 95\% = \left(\bar{X} - \frac{z^* s}{\sqrt{n}}, \bar{X} + \frac{z^* s}{\sqrt{n}} \right) \\ = (207 - 12.74, 207 + 12.74)$$

$$\boxed{\text{C.I } 95\% = (194.26, 219.74)}$$

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.

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 hypothesis test conditions (sample size, mean, and standard deviation), the value of α and β come out to 0.05 and 0.45 respectively.

Now, a different sampling procedure 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.

Answer:

- a) Claim Statement: The newer batch produces a satisfactory result and passes the quality assurance test. The painkiller drug needs to have a time of effect of at most 200 seconds.

Hypotheses:

- Null Hypothesis H_0 : The time of effect of the painkiller drug is less than or equal to 200 seconds i.e., $H_0 \leq 200$
- Alternate Hypothesis H_1 : The time of effect of painkiller drug is more than 200 seconds i.e., $H_1 > 200$

Based on the sign in the Alternate Hypothesis, we can notice that this is an Upper-Tailed Hypothesis Test, in which the rejection region will be on the right side of the distribution.

Hypothesis Testing using "Critical Value Method":

Hypotheses

$$H_0: \mu \leq 200 \quad ; \quad H_1: \mu > 200$$

Details provided

Assumed Population mean time of effect, $\mu = 200$ seconds.

Sample size $n = 100$

Sample mean time of effect, $\bar{X} = 207$ seconds

Sample standard deviation, $S = 65$ seconds.

Since the population standard deviation is not known, the sample standard deviation is approximated to population standard deviation i.e., $\sigma = S = 65$ seconds.

Significance level $\alpha = 0.05$ (5%).

The formula to calculate critical value i.e.,
$$C.V = \mu + (Z_c \times \sigma_{\bar{X}})$$

Since this is an upper tailed test, the critical region is equal to α i.e., 0.05 and the ~~area~~ cumulative probability of the area under the acceptance region will be $1 - \alpha$
 $= 1 - 0.05 = 0.95$.

However, the value 0.95 is not present in the z-table. Taking average of the two closest values in z-table i.e., 0.9495 and 0.9505, we get the corresponding z_c values as 1.64 and 1.65 the average $z_c = 1.645$.

Plugging in the values, we get

$$U.C.V = \mu + (Z_c \times \sigma_{\bar{X}})$$

$$\text{where } \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \cong \frac{S}{\sqrt{n}}$$

$$= 200 + \left(1.645 \times \frac{65}{\sqrt{100}} \right) = 210.69$$

Since, the sample mean of 207 seconds is less than the critical value of 210.69 seconds, we "fail to reject the Null-Hypothesis."

Hypothesis Testing using "p-Value Method":

Hypotheses

$$H_0: \mu \leq 200 \quad ; \quad H_1: \mu > 200$$

Details provided

Assumed Population mean time of effect, $\mu = 200$ seconds.

Sample size $n = 100$

Sample mean time of effect, $\bar{X} = 207$ seconds

Sample standard deviation, $S = 65$ seconds.

Since the population standard deviation is not known, the sample standard deviation is approximated to population standard deviation i.e., $\sigma = S = 65$ seconds.

The z-value for the given sample mean $\bar{X} = 207$ seconds is given by

$$z\text{-value} = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$$

$$\text{where } \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

$$z\text{-value} = \frac{207 - 200}{(65/\sqrt{100})} = \frac{7}{6.5}$$

$$= 1.0769 \approx 1.08$$

Using z-table, we look up the cumulative probability for z-value of less than or equal to 1.08 and we get the z-score as

$$z\text{-score} = 0.8599$$

Since this is a one-tailed test, the p-value is given by

$$\begin{aligned} p &= 1 - z\text{score} \\ &= 1 - 0.8599 \\ &= 0.1401 = 14\% \end{aligned}$$

We have defined the significance level $\alpha = 5\%$. Since the sample mean $\bar{X} = 207$ lies in 14% region, it does not lie in the critical region and we fail to reject the Null Hypothesis.

- b) Case 1: $\alpha = 0.05$ and $\beta = 0.45 \rightarrow$ Since the beta value is very high, this indicates that there are chances that a Type-2 error may occur, in which the mean time to effect is probably more than 200 seconds but we fail to reject the false null hypothesis, thus potentially leading to a situation where a lot of sample taking more than 200 seconds may go un-noticed. Since, the drug in consideration is a painkiller, if the nature of the pain is not very severe, then such alpha and beta values are acceptable. However, in very severe pains this might lead to more trauma to the patient and hence attract regulatory actions and loss of reputation.

Case 2: $\alpha = 0.15$ and $\beta = 0.15 \rightarrow$ Since the alpha value is relatively high, there may be chances that the criteria on time to effect becomes lenient and chances of a Type-1 error occurring become more probable. This indicates that we may end up rejecting a lot of batches of the drugs for failing to meet the regulatory guidelines, even though the batches may still adhere to it, thus causing minatory loss to the company.

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 for its existing subscribers. 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 two-sample proportion test helps in identifying if the population proportions between two groups are significant or not. In the given case, we have two campaigns that have been proposed and we need to find out which is the better campaign. A/B testing is a direct implementation of a two-sample proportion test and hence it is widely used to determine if a significant difference exists between two population proportions.

Steps to be followed in A/B testing:

1. The first step is to identify the hypotheses; Null Hypothesis and the Alternate Hypothesis. Since, currently there is no tagline which has been proved to be performing in a certain manner, the null hypothesis would be that there is no significant difference between the two taglines i.e., the difference between the two population proportions is zero. The alternate hypothesis would be that there is a significant difference between the two campaigns and that needs to be proved.

In many cases, there would be an enhancement to an already existing campaign (webpage) and in such cases the null hypothesis would be that the current campaign is equal or better than the new campaign.

2. The population is split into two groups namely the Variation Group A and the Variation Group B. In this scenario, respondents are divided between the two campaign taglines equally and are asked whether they like the campaign or not.
3. The input from each of the respondents is recorded and the population proportions are calculated i.e., of the number of people who were exposed to Variation Group A tagline, how many respondents liked the campaign. Similarly, of the number of people who were exposed to Variation Group B tagline, how many respondents liked the campaign.
4. Next, the difference of the proportions is measured by subtracting the population proportion Variation Group B from Variation Group A. In case the difference is significant and positive, then the Variation Group A is assumed to be more effective. Similarly, if the difference is significant and negative, the Variation Group B is assumed to be more effective.