Statistics and EDA

Inferential Statistics and Hypothesis Testing

Assignment Submitted By: Janarthanan Balasubramanian

Background: 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 (the time taken for the drug to completely cure the pain), as well as the quality assurance (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.

We apply **binomial distribution** here because:

a. The total number of trials (i.e. samples) is fixed at 10
b. Each trial has only two possible outcomes: success (drug produces satisfactory result) or failure (drug does not produce satisfactory result).
c. Probability of success is the same for all drugs is 4 times the probability of failure. These values do not change with the trials.

Number of samples = n = 10

Probability that the drug is able to produce satisfactory result = 0.8

Probability that the drug does not produce satisfactory result = 0.2

In a binomial distribution, the formula to compute theoretical probability:

$${}_nC_r\ p^r\ (1-p)^{n-r}$$

To compute theoretical probability that **at most three drugs are not able to do satisfactory job**:

n = 10   r = 10, 9, 8, 7   p = 0.2   1 – p = 0.8

$P\ (X \le 3) = {}_{10}C_0\ (0.2)^0\ (0.8)^{10} + {}_{10}C_1\ (0.2)^1\ (0.8)^9 + {}_{10}C_2\ (0.2)^2\ (0.8)^8 + {}_{10}C_3\ (0.2)^3\ (0.8)^7$

= (1 x 0.1073 x 1) + (10 x 0.1342 x 0.2) + (45 x 0.1678 x 0.04) + (120 x 0.2097 x 0.008)

= 0.1073 + 0.2684 + 0.3020 + 0.2013 = 0.8790

The probability that at most three drugs are not able to do satisfactory job is **87.9%**

Question 2: For the effectiveness test, a sample of 100 drugs were 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.

It is always not practical to find the mean μ and standard deviation σ of the entire population. Therefore, it is beneficial to find the mean and standard deviation of a representative sample. Here we have taken a sample of 100 drugs for the effectiveness test. However, we cannot draw a conclusion that the sample mean is equal to the population mean. There is a margin of error introduced due to the flaws in the sampling process.

If we take multiple samples of 100 drugs, and then plot the means of all such samples, then the distribution of such sample means is called a sampling distribution.

According to the **Central Limit Theorem (CLT)**:

a. Sampling Distribution's Mean ($\mu_{\bar{X}}$) = Population Mean (μ)
b. Sampling Distribution's Standard Deviation (Standard Error) = σ / $\sqrt{n}$, where σ is the population's standard deviation and n is the sample size.
c. For n > 30, the sampling distribution becomes a normal distribution

Sampling distribution is just a theoretical exercise. In real life scenarios, we do not create a sampling distribution; we just take a sample. By applying the CLT, we can assume that the sample mean is normally distributed for a sample size n > 30. First, we find the mean and standard deviation of the sample so that we can estimate the population mean in the form of an interval with some confidence level using the following formula.

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

Z^* is the Z-score associated with specific confidence interval.

Step 1: Take a sample of 100 drugs (n = 100)

Step 2: Find the mean ($\bar{X}$) and standard deviation (S) of the sample. $\bar{X}$ = 207 secs and S = 65 secs.

Step 3: Apply the formula to estimate the population mean at 95% confidence interval. For 95% confidence interval, Z-score is Z^* = 1.96.

Substituting the values in the formula, we get 207 ± (1.96 x 65 / √100)

Margin of Error = 12.74

The confidence interval is **(194.26, 219.74)** corresponding to 95% confidence level.

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 two 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.

A claim made about the entire population based on our intuition or through inference is called a hypothesis. We need to test our hypothesis so that we can conclude whether our hypothesis about the population parameter is true or not. The pain killer drug needs to have a time of effect of at most 200 seconds (beyond which it will be considered ineffective).

Null Hypothesis: H_0: $\mu \leq 200$ seconds (the status quo) i.e. the drug is effective
Alternate Hypothesis: H_1: $\mu > 200$ seconds (the challenge) i.e. the drug is not effective.

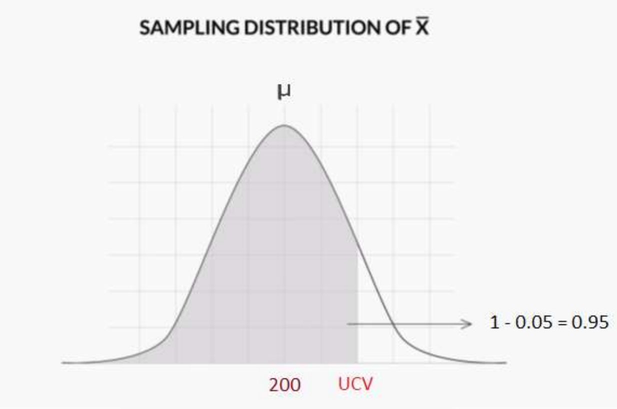
We perform hypothesis testing to see whether there is sufficient evidence to challenge the status quo so that we can reject the null hypothesis. If we do not have sufficient evidence to support the alternate hypothesis, then we fail to reject the null hypothesis.

We have taken a sample of 100 drugs (n = 100).
Sample Mean $\bar{X}$ = 207 secs and Sample Standard Deviation S = 65 secs

We need to test whether our claim that the drug is effective is true or not. Here, we will utilize two methods to arrive at our conclusion.

Method 1: Critical Value Method

We perform an **upper-tailed test** in this case i.e. the rejection region is towards the right side of the distribution. The region of acceptance is below the UCV (upper critical value) and is represented by the shaded area in the diagram below.



Critical value is computed as CV = $\mu_x + Z_c\, \sigma / \sqrt{n}$

The significance level is 5% (0.05)

Z_c for this significance level = 1.645

The cumulative probability is (1 – 0.05) = 0.95

Critical Value = 200 + 1.645 x 65 / 10

= 200 + 10.6925 = 210.6925 ≈ **210.69**

Sample mean 207 secs < 210.69 secs. Therefore, we **fail to reject the null hypothesis**.

Method 2: p-Value Method

The p-Value measures the strength of the evidence in support of a null hypothesis. If p-Value is less than the significance level (0.05), we reject the null hypothesis.

Step 1: The Z-score for the sample mean point on the distribution = ($\bar{X}$ - μ) / (σ / $\sqrt{n}$)

= (207 – 200) / (65 / 10) = 7 / 6.5 = 1.077

Step 2: Find the p-Value for this Z-score using Z-table. p-Value = 1 - 0.8599 (since the sample mean is on the right side of the distribution) = **0.1401**

Step 3: Here p-Value (0.1401) > significance level (0.05). So we **fail to reject the null hypothesis**.

Part B: In hypothesis testing, there is always a possibility of error i.e. making the wrong decision about our hypothesis. There are two types of errors:

Type 1: We reject the null hypothesis in spite of it being true.
Type 2: We fail to reject the null hypothesis in spite of it being false.

The probability of Type 1 error is denoted by α and that of Type 2 error by β.

α = 0.05 and β = 0.45

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. α = β = 0.15

Question 3 (b): 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).

α represents the probability of rejecting a true null hypothesis. This means that in spite of the drug being effective, we may reject it. The proposed alternate procedure increases this probability from 0.05 to 0.15. This will result in rejecting the batch. The alternate procedure is also reducing β the probability of failing to reject the null hypothesis in spite of it being false. This means that the drug is not effective, but still we may fail to reject it. Therefore, the alternate procedure proposed is **NOT PREFERRED** as it increases the probability of effective drug being rejected and decreases the probability of rejecting ineffective drug.

The **reverse scenario** is when having α = β = 0.15 is **PREFERRED** compared to α = 0.05 and β = 0.45. This is possible when the risk of rejecting a true null hypothesis is much lesser compared to the risk of failing to reject a false hypothesis. That means it is better to increase the probability of making a type 1 error. Suppose a patient has to be treated for a disease with a drug that is not having much of side effects. Consider the procedure for diagnosing the patient and ascertaining whether the patient has the disease. An increase in α means that in some cases the patient is wrongly diagnosed to be having the disease (false positive) and the treatment is administered. However, since the drug does not have much side effects it is not a major concern. Type 2 error in this case is that the patient has the disease but we rejected it by mistake (false negative). Decreasing the value of β i.e. decreasing the probability of type 2 error means that the possibilities of such error is reduced. The patient having the disease is not left out from being treated.

Question 4: 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.

A/B testing is an example of two-sample proportion test.

We randomly create two set of samples from the consumers and the sample set X is exposed to one specific campaign (say A) and the sample set Y is exposed to the alternate campaign (say B).

The objective is to study which option performs better.

We get two set of samples. The null hypothesis is that option A is as good as or better than option B. The alternate hypothesis is that the option A is not as good as option B. i.e. Proportion(A) < Proportion(B) or Proportion(A) – Proportion(B) < 0. We can decide the significance level as 0.01 or 0.05. The next step is to compute the Z-score or P-value.

Step 1: Identify the two options. Formulate the null hypothesis and the alternate hypothesis.

Step 2: Conduct the experiment and collect the data from the two samples.

Step 3: Probability of success for option A and B are computed as frequency / sample size.

Step 4: Compute Z score and p-Value for the proportion (two-sample mean test).

Step 5: Make the decision. If p-value < significance level, reject the null hypothesis (option B is better). If p-value > significance level, fail to reject the null hypothesis which means i.e. option A is as good as or better than option B.

We can use XLSTAT add-in to perform these calculations. The frequency (number of success values) and sample size for both the samples are provided as input. The alternate hypothesis and the significance level also can be keyed in.