

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 1:

1.a.) I propose **Binomial Distribution**. The binomial distribution is a probability distribution that summarizes (**cumulative distribution function**) the likelihood that a value will take one of two independent values under a given set of parameters or assumptions. The three conditions that this distribution follows are:

1. The **total number** of trials is **fixed** at **n**.

2. Each trial is **binary**, i.e., it has **only two possible outcomes**: success or failure.
3. **Probability of success** is the **same** in all trials, denoted by **p**.

1.b.) Calculating required probability involves following steps:

It is 4 times more likely that a drug is able to produce a satisfactory result than not. Total probability should be always equal to 1.

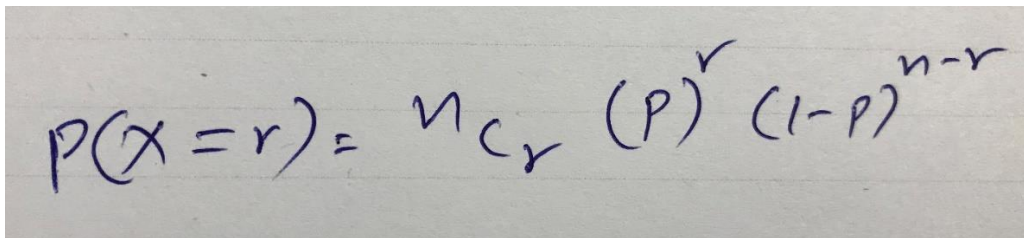
Total probability (X): $4X + X = 1 \rightarrow 5X = 1 \rightarrow X = 1/5 = 0.2$

$P(\text{Unsatisfactory}) = p = X = 1 \cdot 0.2 = 0.2$

$P(\text{Satisfactory}) = (1-p) = 4X = 4 \cdot 0.2 = 0.8$

2. Find the theoretical probability that at most, 3 drugs are not able to do a satisfactory job

Formula is



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

$$P(X=0) = {}^{10}C_0 (0.2)^0 (0.8)^{(10-0)} = 1 \cdot 1 \cdot 0.1073 = 0.1073$$

$$P(X=1) = {}^{10}C_1 (0.2)^1 (0.8)^{(10-1)} = 10 \cdot 0.2 \cdot 0.1342 = 0.2684$$

$$P(X=2) = {}^{10}C_2 (0.2)^2 (0.8)^{(10-2)} = 45 \cdot 0.04 \cdot 0.1677 = 0.3018$$

$$P(X=3) = {}^{10}C_3 (0.2)^3 (0.8)^{(10-3)} = 120 \cdot 0.008 \cdot 0.2097 = 0.2013$$

Answer is

$$P(X \leq 3) = P(X=0) + P(X=1) + P(X=2) + P(X=3) = \mathbf{0.8788}.$$

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 2:

2.a.) A confidence interval calculates the probability that a population parameter will fall between two set values. Confidence intervals measure the degree of uncertainty or certainty in a sampling method.

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

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

where, **Z^* is the Z-score associated with a y% confidence level**. In other words, the population mean and the sample mean differ by a **margin of error** given

by $\frac{Z^* S}{\sqrt{n}}$.

Some commonly used Z^* values are given below:

Confidence Level	Z^*
90%	± 1.65
95%	± 1.96
99%	± 2.58

2.b.) To find the required range from given parameters,

$$n = 100, \bar{X} = 207, S = 65, Y = 95\%$$

First, find the Z-score for 95% confidence level from the above table, i.e., 1.96.

Now, from the confidence interval formula, $(\bar{X} - \frac{Z^*S}{\sqrt{n}}, \bar{X} + \frac{Z^*S}{\sqrt{n}})$,

Required range is $(207 - ((1.96 * 65) / \sqrt{100}), 207 + ((1.96 * 65) / \sqrt{100}))$

Answer is (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 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 3:

3.a.)

Hypothesis Testing starts with the formulation of following two hypotheses:

1. **The null hypothesis (H_0)** always has the following signs: $=$ OR \leq OR \geq
2. **The alternate hypothesis (H_1)** always has the following signs: \neq OR $>$ OR $<$

So,

Null hypothesis (H_0): Painkiller drug needs to have a time of effect of at most 200 seconds (i.e., less than or equal to 200 seconds).

Alternate hypothesis (H_1): Painkiller drug needs to have a time of effect of more than 200 seconds.

Considering the problem $H_0: \mu \leq 200$ and $H_1: \mu > 200$, The ' $>$ ' sign means that the critical region is on the right-hand side (versus both the sides) and, therefore, you only need to use a **one-tailed test**. As the sign in the alternate hypothesis is ' $>$ ', it implies that the critical region is on the right-hand side and, thus, it would be an **upper-tailed test**.

In this problem, the area of the critical region beyond the only critical point, which is on the right side, is 0.05. So, the cumulative probability of the **critical point** (the total area till that point) would be $(1-0.05) = 0.950$.

The next step would be to find the Z_c , which would basically be the z-score for the value of 0.950. Looking at the z-table, 0.950 is not there in the z-table. So, look for the numbers nearest to 0.950. You can see that the z-score for 0.9495 is 1.64 (1.6 on the horizontal bar and 0.04 on the vertical bar), and the z-score for 0.9505 is 1.65. So, taking the average of these two, the **z-score** for 0.9500 is **1.645**.

So, the Z_c comes out to be 1.645. Now, find the critical value for the given Z_c and make the **decision to accept or reject the null hypothesis**.

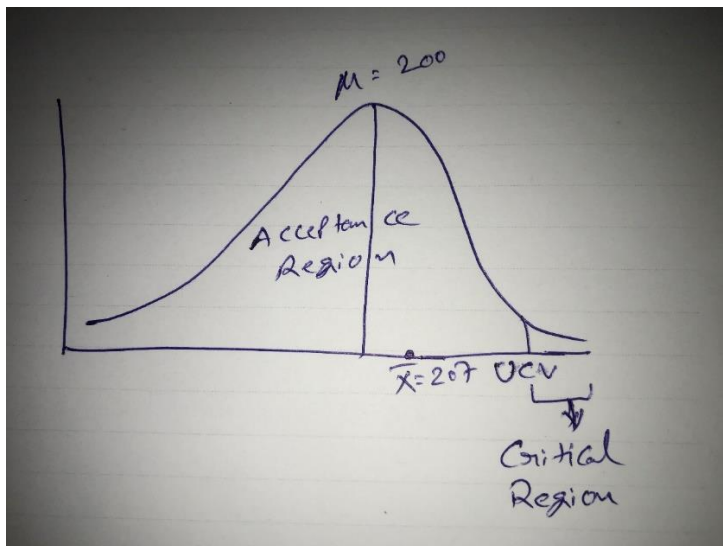
Available parameters: $n = 100$, $\bar{X} = 207$, $S = 65$, significance level $\alpha = 0.05$ (5%), $Z_c = 1.645$, $\mu = 200$

A sampling distribution, which is essentially the distribution of the sample means of a population, has some interesting properties, which are collectively called the **central limit theorem**. It states that no matter how the original population is distributed, the sampling distribution will follow these three properties:

1. **Sampling distribution's mean ($\mu_{\bar{X}}$) = Population mean (μ),**
2. **Sampling distribution's standard deviation (Standard error) = $\frac{\sigma}{\sqrt{n}}$,**
where σ is the population's standard deviation and n is the sample size,
and
3. **For $n > 30$, the sampling distribution becomes a normal distribution.**

I. Using critical value method for decision making:

The **critical value** can be calculated from $\mu + Z_c * (\sigma/\sqrt{n}) = \mu + Z_c * (S/\sqrt{n}) = 200 + 1.645(65/\sqrt{100}) = 210.69$. Since 207 (\bar{x}) is less than 210.69 (**UCV**), \bar{x} lies in the acceptance region and **you failed to reject the null hypothesis**.

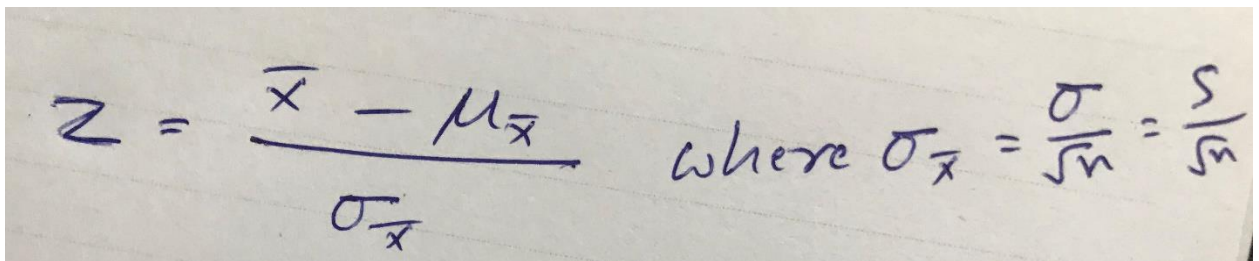


Answer is "Failed to reject the null hypothesis".

II. Using p-Value method for decision making:

After formulating the null and alternate hypotheses, the steps to follow in order to make a decision using the p-value method are as follows:

1. Calculate the value of the z-score for the sample mean point on the distribution.


$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \quad \text{where } \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{S}{\sqrt{n}}$$

$$Z \leq 1.07.$$

2. Calculate the p-value from the cumulative probability for the given z-score using the z-table.

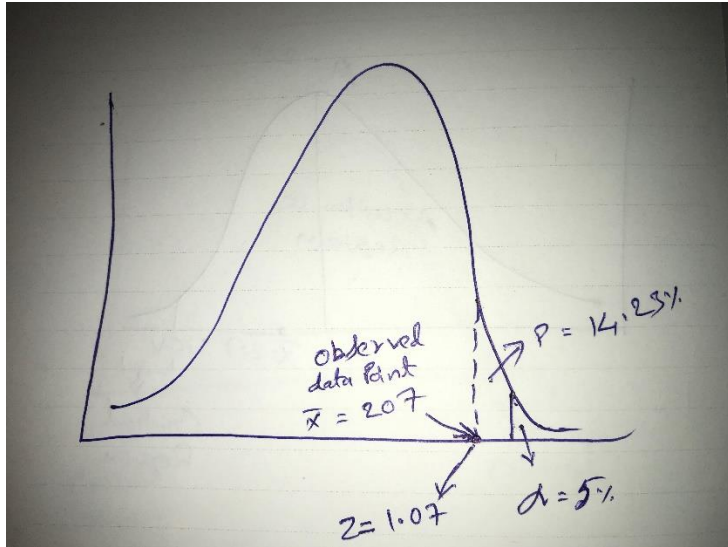
$$P = 1 - 0.8577 = 0.1423 \text{ (14.23\%).}$$

3. Make a decision on the basis of the p-value (multiply it by 2 for a two-tailed test) with respect to the given value of α (significance value).

As it is a one-tailed test, no need to multiply the p-Value by 2.

The higher the p-value, the higher is the probability of failing to reject a null hypothesis. And the lower the p-value, the higher is the probability of the null hypothesis being rejected.

Hence, p-Value 14.23% is greater than α (significance value) 5%, that means sample mean does not lie in critical region so, **you failed to reject null hypothesis.**



Answer is “Failed to reject the null hypothesis”.

3.b.)

We have following two sampling procedures for hypothesis testing:

Method-1: The value of α and β are 0.05 and 0.45 respectively.

Method-2: Both α and β values are fixed at 0.15

2 Types of errors in simple words:

Type-I error: Rejecting true (null hypothesis). Its probability is represented by α .

Type-II error: Accepting false (fail to reject a false null hypothesis). Its probability is represented by β .

Example - 1: A thoroughly tested rocket is scheduled for launching. But mission was postponed due to suspected failures. Truth is program was thoroughly tested. This is Type-I error i.e., rejecting the true. Despite, no issues in program, no problem in postponing the mission.

Since, α value in Method-1 which is less than in Method-2 i.e., probability of rejecting the truth is lesser in Method-1, **Method-1 is preferred over Method-2.**

Example - 2: An improperly serviced vehicle was chosen for travelling whose break may fail in middle of the travel. False is vehicle was serviced properly. This is Type-II error i.e., accepting the false. It is life threatening danger choosing such vehicles for travelling.

Since, β value in Method-2 which is less than in Method-1 i.e., probability of accepting the false is lesser in Method-2, **Method-2 is preferred over Method-1.**

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 4:

Why to use A/B Testing?

A/B testing provides a way for us to test the given two taglines to see which one performs better. A/B testing can be used to demonstrate the impact of each tagline on a user experience and to attract more customers.

How A/B Testing works?

The **two-sample proportion test** is used when you want to compare the given two taglines. After you prepare your variations, you present each tagline to half of your visitors. The test will tell you which tagline proved most popular among your audience based on specific metrics, such as conversion rate or time on page.

Stepwise procedure:

Before the A/B Test

1. Pick one tagline to test.

2. Identify your goal.
3. Create a 'control' and a 'challenger.'
4. Split your sample groups equally and randomly.
5. Determine your sample size (if applicable).
6. Decide how significant your results need to be.
7. Make sure you're only running one test at a time on any campaign.

During the A/B Test

8. Use an A/B testing tool.
9. Test both taglines simultaneously.
10. Give the A/B test enough time to produce useful data.
11. Ask for feedback from real users.

After the A/B Test

12. Focus on your goal metric.
13. Measure the significance of your results using any A/B testing calculator.
14. Choose a tagline on your results (Decision making).