STATISTICS ASSIGNMENT

Given --

1. The pharmaceutical company Sun Pharma is manufacturing a new batch of painkiller drugs, which are due for testing.

2. 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).

Things To Be Tested –

- **Time of Effect**

  Time taken for the drug to completely cure the pain

- Quality Assurance

Whether the drug was able to do a satisfactory job or not.

Let's start with our analysis –

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.

Solution --

- Given –

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.
2. A small sample of 10 drugs

To Find -- The theoretical probability that at most, 3 drugs are not able to do a satisfactory job.

Solution –

a) The probability distribution that portrays the above scenario completely is **Binomial Distribution.**

The three conditions which reflects this are –

- The experiment consists of *n* **identical trials** that is n=10.
- Each trial results in one of the **two outcomes**, called *success* and *failure* that is 0.2 and 0.8.
- The probability of success, remains the **same** from trial to trial.

b) From given, it is 4 times more likely that a drug is able to produce a satisfactory result than not.

Let us **assume** that 'P' is the probability of drug which are not able to do satisfactory job.

- 4P+P = 1
- 5P = 1
- P = 1/5
- P = 0.2

Using **C (n, r) = n! / (r! (n−r)!)**

n = 10 and r <= 3 that is r = 0,1,2,3

Hence the theoretical probability is –

⇨ P(r<=3) = P(r=0)+ P(r=1)+ P(r=2)+ P(r=3)

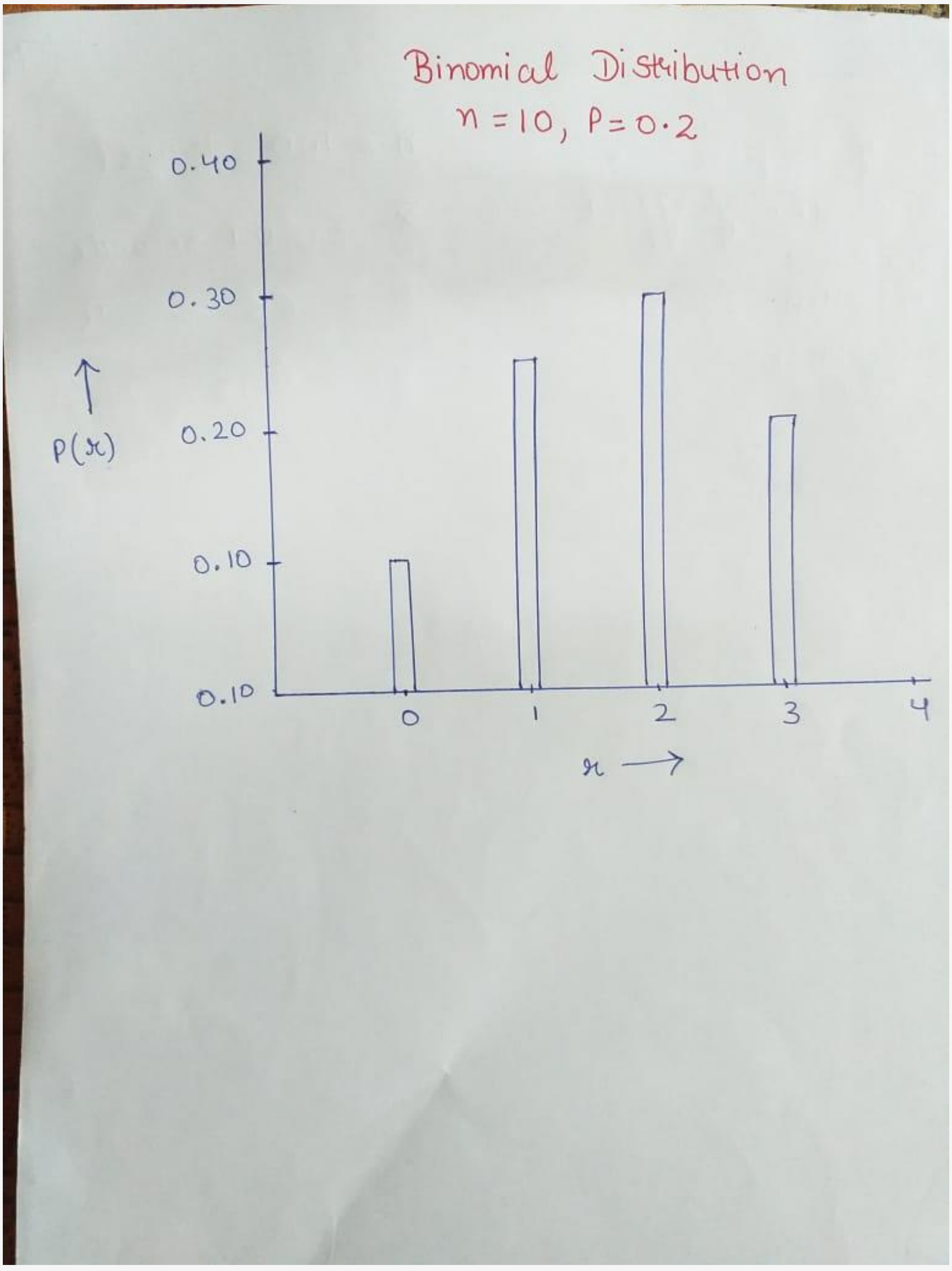
⇨ P(r<=3) = C(10,0) P^0 $(1- P)^{10}$ + C(10,1) P^1 $(1- P)^9$ + C(10,2) P^2 $(1- P)^8$ + C(10,3) P^3 $(1- P)^7$

⇨ $P(r<=3) = C(10,0)\,(0.2)^0\,(0.8)^{10} + C(10,1)\,(0.2)^1\,(0.8)^9 + C(10,2)$

$(0.2)^2\,(0.8)^8 + C(10,3)\,(0.2)^3\,(0.8)^7$

⇨ $P(r<=3) = 0.107 + 0.268 + 0.301 + 0.201$

⇨ $P(r<=3) = 0.877$ ~ **0.88**



Conclusion – Hence, the theoretical probability that at most, 3 drugs are not able to do a satisfactory job is **88%.**

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 in estimate the range 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.

Solution—

Given --

1. Number of samples that is the sample size **(n)** = 100
2. Mean time of effect that is sample mean **(X bar)** = 207
3. Sample Standard deviation **(S)** = 65
4. Confidence level **(y %)** = 95%

To Find -- Estimate the range which the population mean might lie.

Solution –

a) Main methodology is **Central Limit Theorem.**

Properties –

- The mean of the sampling distribution of means is equal to the mean of the population from which the samples were drawn.
- The variance of the sampling distribution of means is equal to the variance of the population from which the samples were drawn divided by the size of the samples.
- If the original population is distributed normally (i.e. it is bell shaped), the sampling distribution of means will also be normal. If the original population is not normally distributed, the sampling distribution of means will increasingly approximate a normal distribution as sample size increases (i.e. when increasingly large samples are drawn).

Steps and **Formulas** that we'll be using further –

1. **Sampling Distribution's Standard Deviation**

Here sigma = S as population standard deviation is same as Sample standard deviation. Hence follows --

Sampling Distribution's Standard Deviation

$$(S.E.) = \frac{\sigma}{\sqrt{n}} = \frac{s}{\sqrt{n}} = \frac{65}{\sqrt{100}} = \frac{65}{10} = 6.5$$

Since the sample size (n) = 100 which is greater than 30, therefore the sampling distribution is a <u>NORMAL DISTRIBUTION</u>

2. Now we'll calculate the Z* using z value table.

As we are having confidence level y% = 95%

= 0.95

As it is a **two tailed test**, hence we'll divide y by 2, which implies

⇨ 0.95/2 = 0.475

After looking at the standard normal z table—

We observe that 1.96 has the value as 0.475

Hence the value of Z* = 1.96

3. Now finding confidence interval

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 & sample mean differ by a margin of error by $\frac{z^* s}{\sqrt{n}}$.

- (207 - 6.5 * 1.96 , 207 + 6.5 * 1.96)
- (207 - 12.74 , 207 + 12.74)
- (194.26 , 219.74)

- **Conclusion --** Hence the required range 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).

Solution –

Given –

1. The painkiller drug needs to have a time of effect of at most 200 seconds to be considered as having done a satisfactory job. That is **μ** <= 200.

2. Number of samples that is the sample size **(n)** = 100
3. Mean time of effect that is sample mean **(X bar)** = 207
4. Sample Standard deviation **(S)** = 65
5. The significance level at 5 %.
6. Standard error **(S.E.)** = 6.5

To find – To test the claim that the newer batch produces a satisfactory result and passes the quality assurance test

Solution –

(A) part

1. **Null Hypothesis** -- H_0 : μ <= 200

   We claim that null hypothesis is true that is painkiller drug have a time of effect of at most 200 seconds.

2. **Alternate Hypothesis –** H_1 : μ > 200

   That is opposite of null hypothesis, which is the painkiller drug have a time of effect less than 200.

3. As **significance level is 5%**

   That is **alpha** (**α) = 0.05, which means that in 5% of cases we'll reject the null hypothesis when it is true.**

⇨ **In other 95% of cases we fail to reject the null hypothesis.**

4. Critical region lies only on one side that is on the right hand, hence it is a **one tailed test**, also it is a **right tailed**

test as critical region lies on the right side, since $H_1 : \mu >$ 200.

5. We'll find the critical region values by finding **critical point (Z_c)** first using the Z table.

As **alpha (α) = 0.05**

⇨ 1 – 0.05 = 0.95

⇨ 0.95 is the **normal cumulative probability**

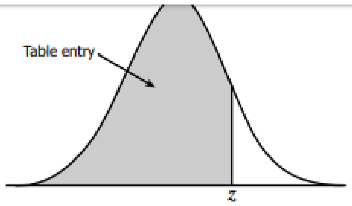


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857

⇨ After seeing in z table we get values Z_c as **1.65.**

⇨ Z_c = **1.65** is the **upper critical point**

6. Now we'll find the critical values –

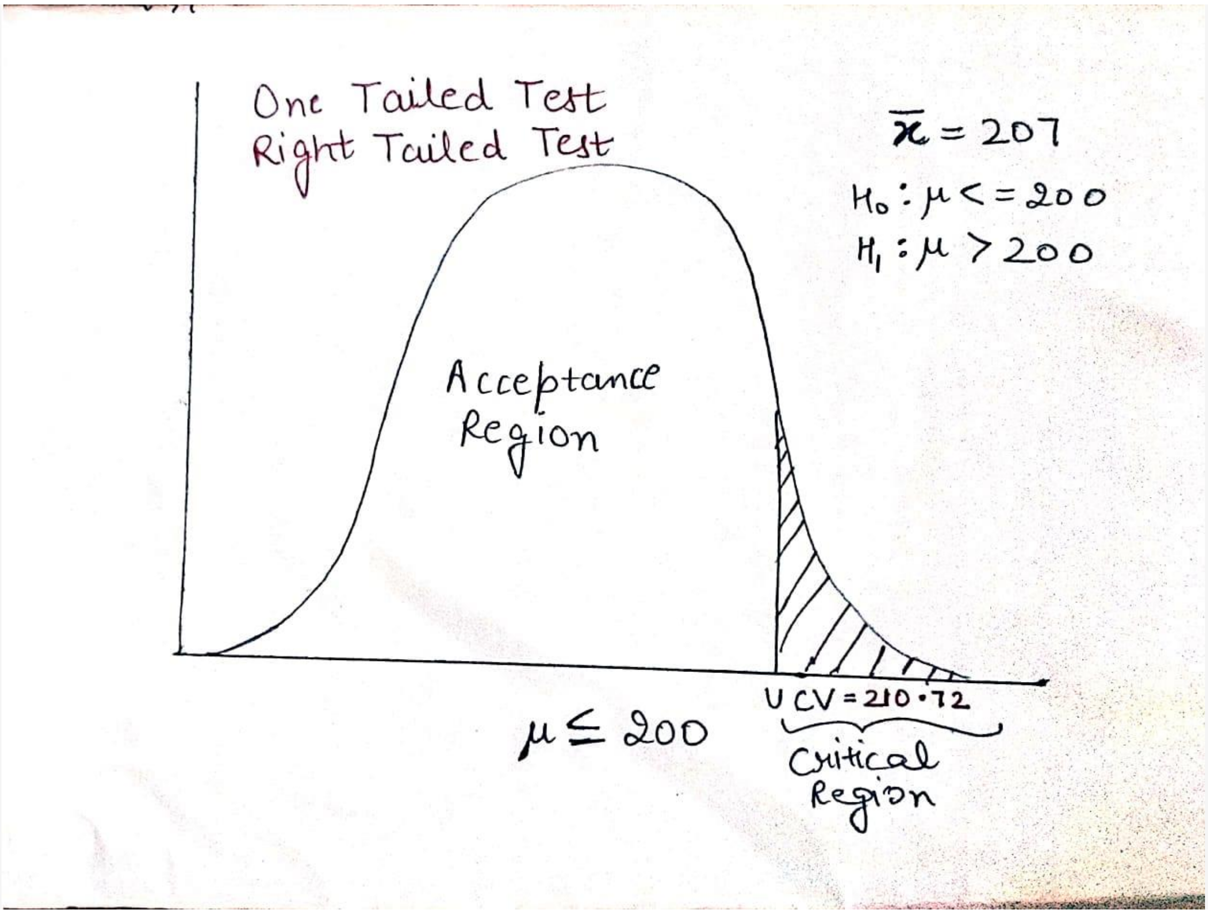
UCV = $\mu + (Z_c \times S.E.)$

= 200 + (1.65 x 6.5)

= 200 + 10.725

= 210.725

7. As the sample mean is 207 which is greater than 200 and less than 210.725. Therefore, it is accepted.



8. **Decision making –** Hence, it is true that the newer batch produces a satisfactory result and passes the quality assurance test.

Now we'll do the same method by **p value method**

1. We'll calculate the **Z Score** for the sample mean –

Z = x̄ - μ / S.E.

= 207 – 200 / 6.5

= 7/6.5

= 1.076

= 1.08

Which is different from the critical value described above.

2. For one-tailed test → p = 1 - 0.8599 = 0.1401=14.01% ~ 14%
3. Now we remember that alpha = 0.05

⇨ However **$\bar{x}$** lies in the 14% of the region, which is more than the required 5 % region.

Conclusion – Therefore the sample mean does not lie in the critical region and we cannot reject the null hypothesis.

(B) Part

Given – alpha (α) = 0.05

Beta (β) = 0.45

Null Hypothesis -- H_0 : μ <= 200

We claim that null hypothesis is true that is painkiller drug have a time of effect of at most 200 seconds.

Explanation —

First Condition

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

Situation – When there is no competition in the market then even if the manufacturing cost is high, there is nothing to worry as people don't have any other compatible option. We can use labour also with machines though it is little bit costly and time consuming as compared to machines. Company is assuring that there is no side effects of the painkiller drug consumed.

Reasons – The alpha value (0.05) is very less but the total error (0.05 + 0.45 = 0.5) is quite high. Value of alpha is also quite less than the other value of alpha (0.15)

- ⇨ Since the alpha value is quite low, therefore we can infer that most of the time our null hypothesis will remain true when it is actually true.
- ⇨ Like in the above case even if the cost is high but due to no competition and painkiller drug giving relief to people on time that is time of effect is less than 200 seconds so people will obviously prefer our drugs only.
- ⇨ Though beta's value is high as compared to alpha which tells us that the chances of rejecting alternate hypothesis when it is actually true is quite high. That is drug will effect after 200 seconds but still it is safe as it is not having any side effects.

Second Condition

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.

Situation – Mixing of ingredients in excess quantity in the painkiller drug so that it can reduce the time of effect on patient.

Reasons – In this case alpha + beta = 0.15 + 0.15 = 0.3 is less as compared to 0.05 + 0.45 = 0.5.

Also alpha is quite more in this situation. Hence rejection of null hypothesis when it is actually true.

- ⇨ If the null hypothesis is true but it is rejected, it can harm to anyone's life as the painkiller having ingredient in excess quantity will give it is effect in less than 200 seconds but it will affect the health badly.
- ⇨ Both are equal but the combined error is less as compared to the other that is 0.3<0.5.

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.

Solution –

A/B testing is a method of comparing two versions of a webpage or app against each other to determine which one performs better. AB testing is essentially an experiment where two or more variants of a page are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal.

Reason for using A/B Testing –

1. It lets us ask focused questions about changes to our website or app (that is taglines in this case), and then collect data about the impact of that change.
2. It enables data-informed decisions that shift business conversations from "we think" to "we know."
3. It ensures that every change produces positive results.
4. A/B testing lets us know what words, phrases, images, videos, testimonials, and other elements work best.
5. It increases sales and reduces risks.

Procedure for using A/B testing –

1. **Choose what you want to test –** We have selected the two taglines, now we have to decide which one to use among them.

2. **Identify Goals** – This is to determine whether the variation is more successful than the original version. We need to plan an effective online ad campaign to attract new customers.
   Hence we'll post both the taglines and then will observe the response and behaviour of the users.

3. **Generate Hypothesis** – Create the page, keep everything same (including the font) except the taglines because we need to perform our test on taglines only. It will be easier to identify logical things to test and come up with the expected impact.
4. **A/B Testing tool** – We can use XLSTAT in Excel. By recording the observations we'll enter the range in dialogue box and proceed the testing.

5. **Analyse the statistics** -- Draw conclusions based on which variation won.

6. Once we understand which version our audience liked better — and by what margin, we can use that particular tagline.