STATISTICS ASSIGNMENT

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:

a.) I have considered Binomial distribution to solve the above problem statement due to the reasons listed below:

: - The number of outcomes is fixed that is 2 possibilities for each trial one success and failure.

: - The number of trials is fixed and independent.

: - The probability of success is same for all the trials.

b). As per the question: 4 drugs are able to produce a satisfactory result

Let S be the event produces satisfactory result and U be the event which does not produce the satisfactory result.

As per the question the equation would be:

$$4P(U) = P(S)$$

As we know probability of any event is always 1:

$$P(S) + P(U) = 1$$

From the above equations we get:

$$P(S) = \frac{4}{5} = .8\,, P(U) = \frac{1}{5} = .2$$

Accordingly, if we consider X be the event that out of 10 trials, we pick x amount of drugs that do not produce as good of results as other drug, X is a random variable following "Binomial Distribution", where sample size n=10; and (probability of success as not producing satisfactory result) p =1/5 = 0.2.

So, $P(X = x) = \binom{10}{x}\left(\frac{1}{5}\right)^{x}\left(\frac{4}{5}\right)^{10-x}$ for 0≤x≤10

Now considering 3 drugs are not able to produce the satisfactory results and p[of success] = 0.2 and n=10;

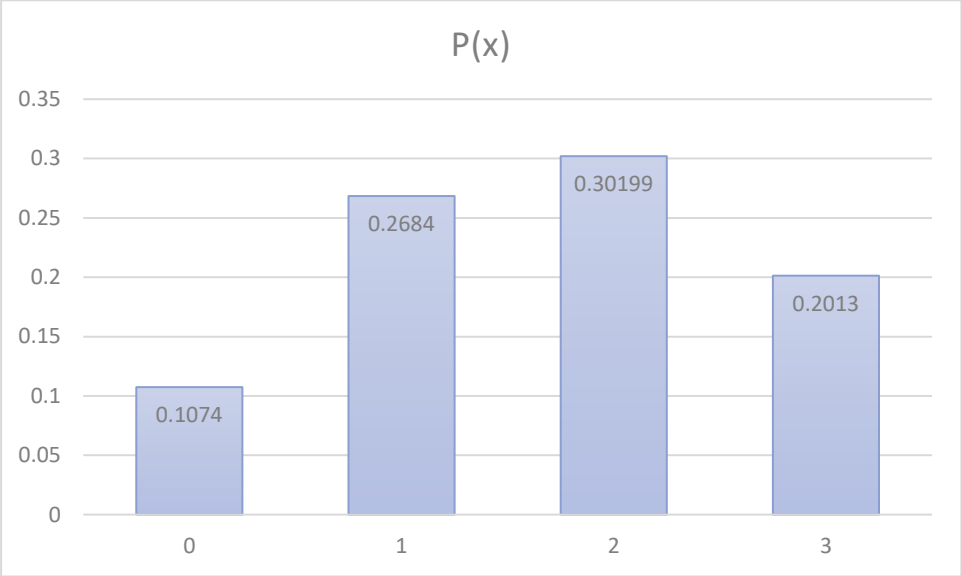
So probability P(X) will follow as below equation:

$$P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) = \sum_{i=0}^{3} \binom{3}{i} \frac{1}{5x}^{i} \frac{4}{5}^{10-i} = .8791$$

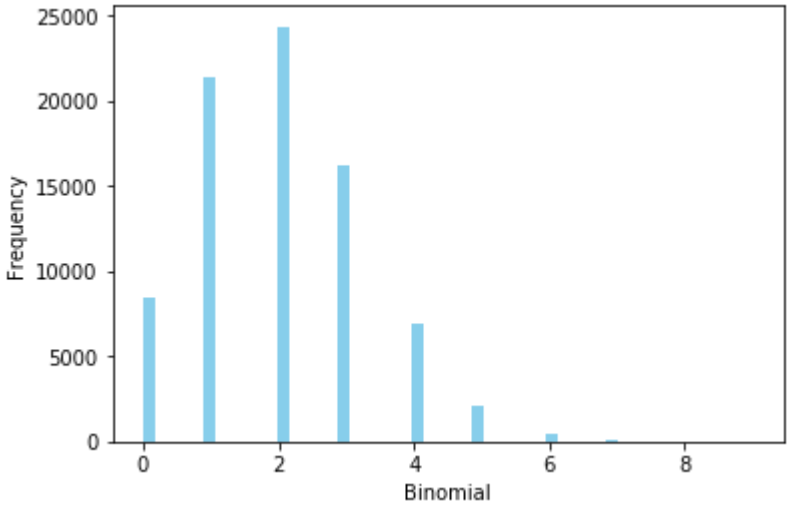
Tabular representation of Probability values when X takes different (0,1,2,3) values:

X	P(x)
0	.1074
1	.2684
2	.30199
3	.2013

Probability Distribution:



Graphical representation of "Binomial Distribution" p(success) = 0.2 size=80000, n=10



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.

Solution:

a.) **Credit Limit Theorem** will be used to solve this problem as data to be analysed is large and it can be performed on any random samples.

Credit Limit Theorem describes the characteristics of distribution of values we would obtain if we were able to draw an infinite number of random samples of a given size from a given population and we calculate mean of each sample.

Credit Limit Theorem has 3 important properties:

: - The mean of the sampling distribution of means is equal to the mean of the population from which the samples were drawn.

$\mathrm{E}(\overline{X}) = \mu$, where μ is population mean

: - 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.

$$var\left(\overline{X}\right) = \frac{\sigma^2}{n}$$

: - The distribution of means will increasingly approximate a normal distribution as the size N of samples increases. In other words, we can say that if the sample size is greater than 30 it follows normal distributions.

b.) In order to calculate the range, we need to calculate the confidence interval
Confidence interval is given by the formula:

$$(\bar{X} - \frac{Z^* S}{\sqrt{n}}, \bar{X} + \frac{Z^* S}{\sqrt{n}})$$

As per the question:

Mean (X) = 207
Confidence level = 95%
Z*(at confidence level 95% from the Z- table) = ± 1.96
Standard deviation σ = 65
Sample size (n) = 100
Standard Error = $\frac{\sigma}{\sqrt{n}}$ = 6.5
Margin of Error is given by : Z*$\frac{\sigma}{\sqrt{n}}$ = 1.96*6.5 = 12.74
Required Range = (207 -12.74), (207+12.74) = (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:

a.) To conduct hypothesis test with the given data we would consider two methods listed below:

: - Critical Value Method

: - P-value Method

Considering the given claim – the maximum time taken for the drug to be done a satisfactory job is 200 seconds, So the hypothesis statement would be formulated as

Null Hypothesis H_0: $\mu <= 200$

Alternate Hypothesis H_1: $\mu > 200$

Critical Value Method:

We can conclude that the test would be one tailed test with critical region lying on the right of the Z critical region

Given that: μ= 200,

$$\sigma_x = \frac{S}{\sqrt{n}} = \frac{65}{10} = 6.5$$

Significant level(α) = 5%= 0.05

Zc(Z critical) = 1.645

Critical Value = μ+Zc*σ_x = 200+1.645*6.5 = 210.6925

Conclusion: As the mean of the given experiment is 207 and lie on the left of Critical value(μ<CV) , we fail to reject the Null Hypothesis.

P-Value Method: In order to use this method to make a decision, we need to follow the below steps:

: - We need to calculate the Z-score from the sample mean point on the distribution.

: - Calculate the p-value from the cumulative probability for the given Z-score using Z table

: - Make a decision based on the given significance value

Given Mean X = 207 and sample mean = 200

$\boldsymbol{Zs_{sc}}$**= X- ux/**σ_x = 207 – 200/6.5 = 1.0779 = 1.08

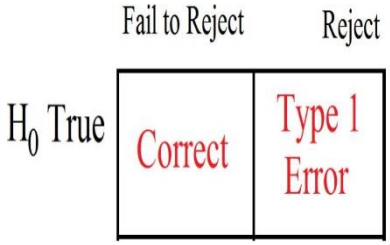
So considering the Z-table and confidence level 0.05

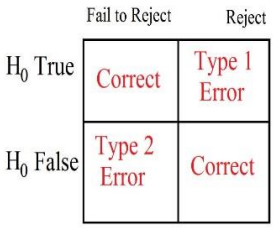
Cumulative Probability =0 .86

P -value = 1-0.86 =0.14 = 14 %

Conclusion: As p-value is greater than the given level of significance, we fail to reject the null hypothesis

b.) A **Type I error** is when we reject a true null hypothesis and **A Type II error** is when we fail to reject a false null hypothesis.
α: is Type I error rate where as **β**: is Type II error rate whereas **β** depends on other factors from sample, like **standard error** and **effect size.**





Let us consider the study of drugs Tamiflu and psychosis having an claim that the population who consumes Tamiflu shows psychosis behaviour.
Let consider **α = .05. & β=.45**
As per the inference the investigator has set 5% as the maximum chance of incorrectly rejecting the null hypothesis by inferring that the use of Tamiflu and psychosis incidence are associated in the population. This is the level of reasonable doubt that the investigator is willing to accept when he uses statistical tests to analyse the data after the study is completed.
When we chose small value of **α** that clearly signifies that we do not want to commit a Type I error.
β increases with standard error, and decreases with effect size.

If β is set at 0.45, then the investigator has decided that he/she is willing to accept a 45% chance of missing an association of a given effect size between Tamiflu and psychosis. This represents a power of 0.55, i.e., a 55% chance of finding an association of that size.
If both **α & β are set to .15** that is there is an equal opportunity of committing Type I & Type II error. As per the inference investigators are ready to accept 15% chance of finding an association and also missing the association of drugs Tamiflu and psychosis incidence on the population.
In general the investigator should choose a low value of alpha when the research question makes it particularly important to avoid a type I (false-positive) error, and he should choose a low value of beta when it is especially important to avoid a type II error.

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:

As per the, there are two taglines proposed by the company, and team is currently divided on which option to use. This is Two sample proportion test and we can do A/B testing.

Here we have to do comparative analysis, which tagline attract more customers. In other word which tagline is easy to catch an eye of customer and attract them. Reason of attraction of customers can be anything a phrase, be colour of text, font style of tagline, tagline sentence meaning etc.

Procedure:

: - Two groups of people are chosen and each of them is exposed to the only type of advertising.

: - Their conversion rates are considered.

: - The null hypothesis and the alternate hypothesis are defined accordingly. In this case, the null hypothesis would be that — the original tagline is more effective than the alternate tagline. And the alternate hypothesis would be the vice versa.

: - Choose the value of **α (significance level).**

: - After collecting the test data, use a suitable tool — such as XLSTAT, Optimizely, etc. — to run the hypothesis, and then calculate the p-value.

: -Make the decision on the basis of whether the p-value is larger than the given significance level.

Let us suppose that we have tagline 1 and tagline 2. Now we can divide them into two group of customers, which customer prefers which tagline either tagline 1 or tagline 2. Suppose

48% customers likes tagline 1 and 30% customers likes tagline 2 and rest of all don't line either tagline 1 or tagline 2.

So we come with conclusion tagline 1 is more attractive than tagline 2. If results are too closer, we can go for more sample collection for conclusion.