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.

Answer: The event here can have two outcomes, the drug doing a satisfactory job and the drug not doing a satisfactory job. This scenario can be accurately portrayed by a binomial distribution. The three conditions that this distribution follows are:

- The total number of trials is fixed
- Each trial is binary, i.e. has only two possible outcomes, success and failure.
- The probability of success is the same for all the trials.

b) Calculate the required probability.

Answer: Let **P** be the probability of **the drug not doing a satisfactory job**. The Probability of the drug doing a satisfactory job will be given by **(1-P)**.

It is given that 4*P = (1-P).

Solving the above equation, we get **P = 0.2**.

Let us take **X** to be the variable which denotes **the number of drugs that do not do a satisfactory job**.

Per the generalization of the binomial probability distribution,

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

Here **n** is the number of trials and **r** is the number of times the outcome is a success.

The probability that the random variable **X** takes a value less than or equal to **x**, the cumulative probability, can be denoted by F(x)

Now, $F(3) = P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$

$= {}^{10}C_0 * (0.2)^0 * (0.8)^{10} + {}^{10}C_1 * (0.2)^1 * (0.8)^9 + {}^{10}C_2 * (0.2)^2 * (0.8)^8 + {}^{10}C_3 * (0.2)^3 * (0.8)^7$

$= \mathbf{0.879}$

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.

Answer: Here we can infer the population mean using sample mean. A population mean can be estimated using a sample mean and calculating the margin of error. Population Mean = Sample Mean ± Margin of Error

Sample mean $\bar{X}$, Sample Standard Deviation **S** and Sample size **n**

Confidence interval (y% confidence level) = $(\bar{X} - Z^*S/\sqrt{n}, \bar{X} + Z^*S/\sqrt{n})$

Where Z* is the Z score associated with the y% confidence level.

Central Limit Theorem lets us assume that sample mean would be normally distributed, with mean μ and standard deviation σ/√n (approx. S/√n). CLT states that no matter how original population is distributed, sampling distribution will follow these properties:

i. Sampling distribution's mean ($\mu_{\bar{x}}$) = Population mean (μ)
ii. Sampling distribution's standard deviation (Standard error) = σ/√n, where σ is population's standard deviation and n is sample size
iii. For n > 30, the sampling distribution becomes a normal distribution

b.) Find the required range.

Answer: $\bar{X} = 207$

n = 100

S = 65

Z* for 95% confidence interval = ±1.96

Margin of error = $Z^*S/\sqrt{n}$

$= 1.96 * 65/\sqrt{100}$

$= 1.96 * 65/10$

= 12.74

Range for population mean with 95% confidence level = Sample Mean ± Margin of Error

= 207 ± 12.74

Hence, for 95% confidence level, the population mean will lie between **194.26 and 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.

Answer: To perform hypothesis testing for this problem, we first need to state the null hypothesis and the alternate hypothesis.

Null Hypothesis or H_0 : $\mu \leq 200$

Alternate Hypothesis or H_1 : $\mu > 200$

I. **Critical Value Method**: α (significance level 5%) is given as 0.05. 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 = **0.95**.

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, we can see that the z-score is equal to **1.645**.

Next step is to find the critical value for this Z_c.

Also, the standard deviation of the sample is given as 65, instead of the population's standard deviation. In such a case, we can approximate the population's standard deviation to the sample's standard deviation, which is 65 in this case.

Critical Value = $\mu + Z_c \times (\sigma/\sqrt{N})$

Putting the values $\mu = 207$, $Z_c = 1.645$, $\sigma = 65$ and $N = 100$

Critical Value = $207 + 1.645 \times (65/\sqrt{100})$

= **217.6925**

Since the sample mean, **207** is less than the critical value calculated above, we **fail to reject the null hypothesis**.

II. **p-value method**:

First step is to calculate the z-score for the sample mean point on the distribution.

z-score = $(\overline{x} - \mu)/(\sigma/\sqrt{N})$

= (207 – 200)/(65/√100)

= 1.08

Second step is to calculate the p-value for the z-score of 1.08 (corresponding to the sample mean of 207). This will be calculated using value in the z-table corresponding to 1.0 on the vertical axis and 0.08 on the horizontal axis.

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

Cumulative probability using the z-score of 1.08 = 0.8599

Since the z-score is positive, and this is a one-tailed test, p-value = 1-0.8599

= **0.1401**

Since the p-value, 0.1401 is greater than α (0.05), we fail to reject the null hypothesis.

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: The probability of Type 1 error is low for the first method and the probability of Type 2 error is low for the second method.
Null Hypothesis: The drug is effective ($\mu \leq 200$)
Alternate Hypothesis: The drug is ineffective ($\mu > 200$)

1. The first method would be preferable over the second method in those cases where a Type 1 error would be more dangerous to make (and/or have more serious consequences) than a Type 2 error. Imagine a scenario where the drug being developed is an only low cost alternative to a very costly drug. In this scenario if a Type 1 error is made and the drug is deemed to be ineffective when in fact it is effective, many poor people who cannot afford the existing costly drug will be deprived of a cheap and possible method of treatment. However, if a Type 2 error is made, and the drug is deemed effective when in fact it is ineffective, the worst that could happen is that the drug may be rejected as a method of treatment.
2. The second method would be preferable over the first method in those cases where a Type 2 error would be more dangerous to make (and has more serious consequences) than a Type 1 error. Let us consider that the side-effects from the drug are very serious (possibly fatal in the cases where drug is administered wrongly or the diagnosis is incorrect). Also, let there be another drug with similar cost already available in the market. In this scenario if a Type 2 error is made and the drug is deemed to be effective when in fact it is ineffective, it could cause serious consequences for those patients on which the drug is administered. However, if a Type 1 error is made and the drug is deemed to be ineffective when it is in fact effective, the only consequences will be that people will switch over to the alternate drug for treatment.

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: A/B testing can be used here since it is an appropriate approach to compare performances of two (or more than two) variations of a product (or experiment). In A/B testing, two (or more than two) options are present to users at random and then statistical analysis is done to hypothesize which option performs better. As the users are given the options, their responses to each option is measured and that data is collected. This data is then analyzed to check which option generated a more positive response.

In this scenario, two sets of users can be taken and then presented with one tagline each. Sales of the product can be checked to see which tagline generated more sale numbers.

A two-sample proportion test can be conducted as follows:

1. Divide the users into two populations, P1 (first tagline) and P2 (second tagline).
2. From both P1 and P2, some proportion of customers will react positively (buy the drug) to the tagline they were presented with. Let these be $\overline{P1}$ and $\overline{P2}$. Positive response can be stored as 1 and a negative response as 0.
3. The aim is to use the difference between these proportions to arrive at a conclusion if there is any significant difference between them.
4. We can take the sizes and the respective frequencies of the two populations, and use these values to conduct a “Tests for two proportions” from the “Parametric Tests” section of the XLSTAT tab in MS Excel.
5. The Alternate hypothesis in the Options tab of the test parameters box can be set to “Proportion 1 – Proportion 2 < D” where D (Hypothesized difference) is set to 0. This means that the Population 2 is better than Population 1. This would imply that the null hypothesis is that the Population 1 is better than the Population 2.
6. In the test results, if p-value is greater than the significance level, null hypothesis cannot be rejected and Population 1 is indeed better than the Population 2. If p-value is lesser than the significance level, null hypothesis can be rejected and the Population 2 is better than Population 1.