upGrad

Statistics Assignment

Statistics & EDA

Module 11 Session 1

Version 1.0

TABLE OF CONTENTS

Table of Contents **2**
Documentation Control **2**

DOCUMENTATION CONTROL

Version	Date	Reason for issue	Issued By
V 1.0	06-Oct-2019	Initial Version	Chandan Singh

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 1(a):

As the test is about the probability of Success Or Failure of the drugs, in other words only two possible outcomes, Binomial Distribution is the probability distribution that accurately portray this scenario.

The three conditions, that a Binomial Distribution must meet are:

i) The no of observations or trial is fixed
ii) Each trial is binary, i.e., it has only two possible outcomes: Success or Failure
iii) The probability of success is the same for all the trials.

b.) Calculate the required probability.

Answer 1(b):

Let's X be No of Drugs not able to produce satisfactory result out of 10 Drugs
Then, X would follow a binomial distribution with n = 10 and P = ?

Let's find out the probability of not producing a satisfactory result => P(ns) = X

As per the previous quality checks, 4 times a drug is able to product satisfactory result, therefore, P(s) = 4X

Now, according to one of the probability rules :- The sum of the probabilities of all the outcomes in a sample space equals 1. Therefore:

$P(ns) + P(s) = 1$
$X + 4X = 1$
$5X = 1$
$X = 1/5$
$X = 0.2$

P(X) = 0.2

The probability of not able of produce a satisfactory result P(X) = 0.2

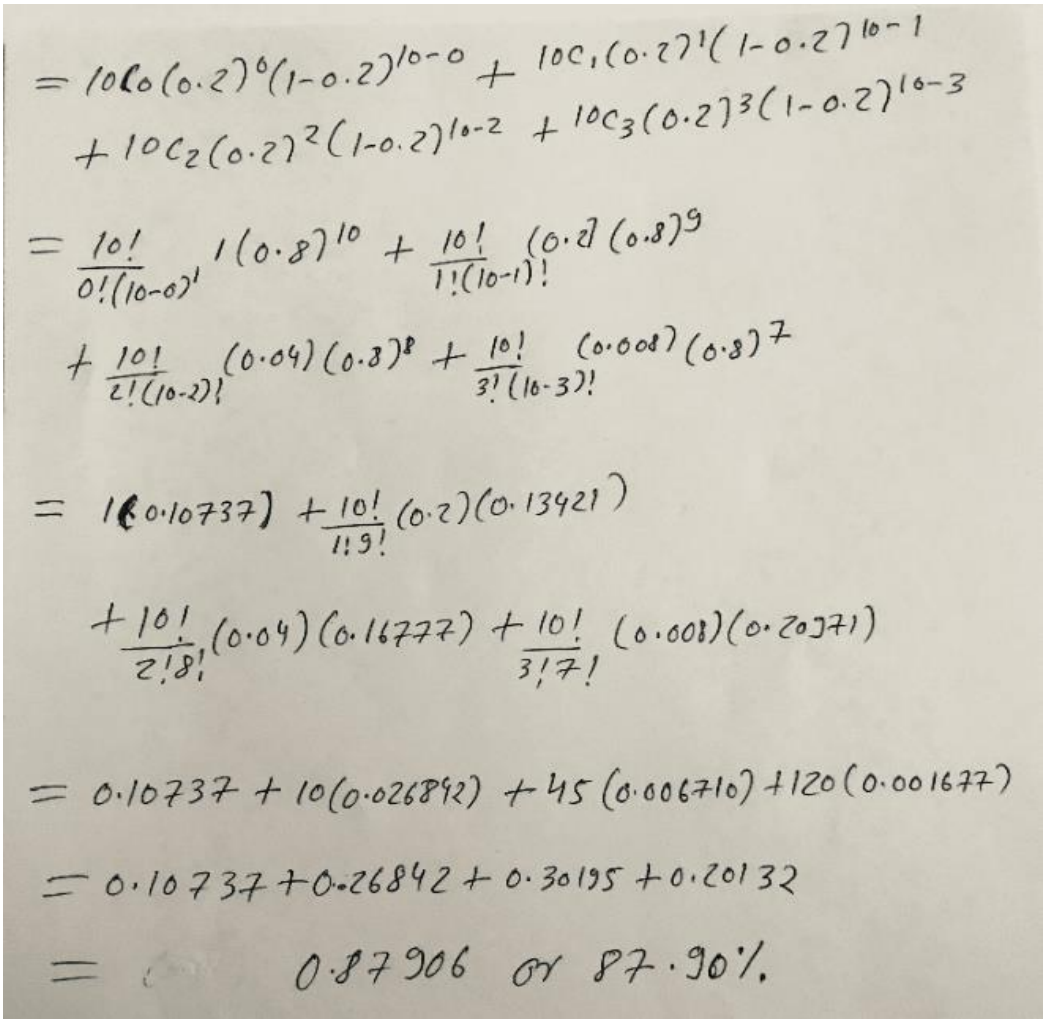
Now, the probability that at most, 3 drugs not able to do a satisfactory job can be found as:

Let's, F(X) = P(X<=3)

P(X<=3) = P(X=0) + P(X=1) + P(X=2) + P(X=3)

Binomial Distribution Formula:

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



F(3) = P(X<=3) = 0.87906 or 87.90%

Therefore, probability that at most 3 drugs are unable to do satisfactory job is 87.90%

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.

Answer 1(a):

CLT(Central Line Theorem) is the methodology that would be used to solve this problem. According to CLT, the mean of a sample of data will be closer to the mean of the overall population in question, as the sample size increases. In other words, the data is accurate whether the distribution is normal or aberrant.

As a general rule, sample sizes equal to or greater than 30 are sufficient for the CLT to hold, meaning that the distribution of the sample means is fairly normally distributed. Therefore, the more sample one takes, the more the graphed results take the shape of a normal distribution.

The properties of CLT are:

I. *Sampling distribution's mean (μx) = Population mean (μ)*
II. *Sampling distribution standard deviation(standard error) = σ/√n, where σ is the population standard deviation.*
iii. *For n > 30, the sampling distribution becomes normal distribution.*

b.) Find the required range.

Answer 2(b):

Let's X defined as the time of effect, then for this sample of X:

Sample Mean $\bar{X}$ = 207 Secs
Sample Standard deviation S = 65 secs
Sample size n = 100

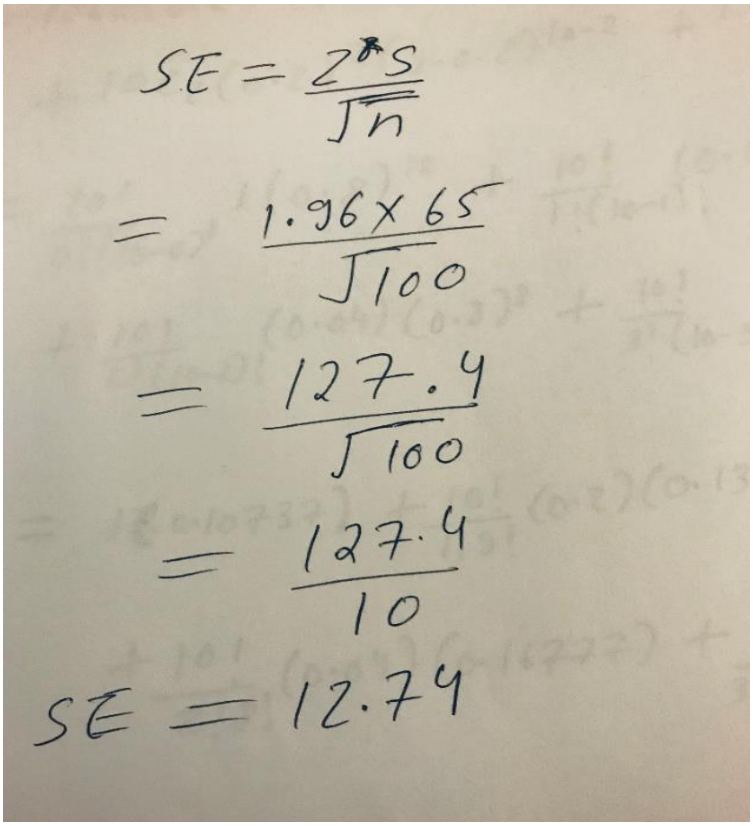
Confidence Level	Z*
90%	±1.65
95%	±1.96
99%	±2.58

Also, for 95% confidence level Z^* is 1.96
*We know that margin of error or Standard Error SE = Z*S/ √n*

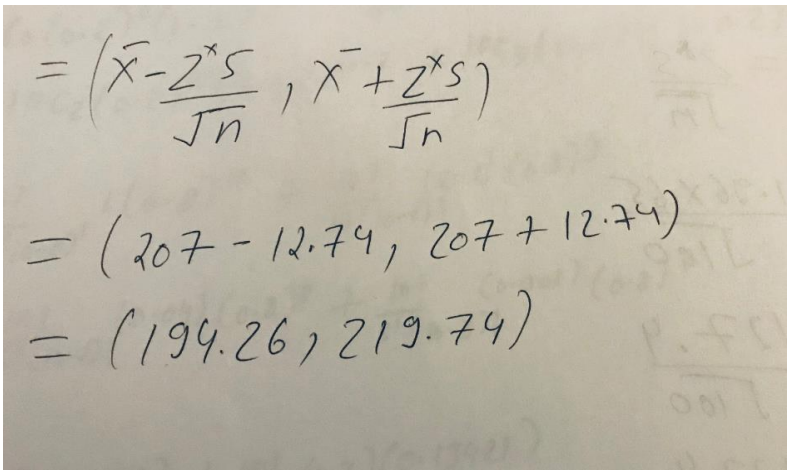
Formula to calculate interval is :

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

Let's first calculate the standard Error S.E



Now, calculate population mean Interval :



Interval/Range = (194.26 seconds, 219.74 seconds)

*Therefore, Margin of Error corresponding to 95% confidence level is 12.74 and the population mean lies between **194.26** seconds and **219.74** seconds.*

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 3(a):

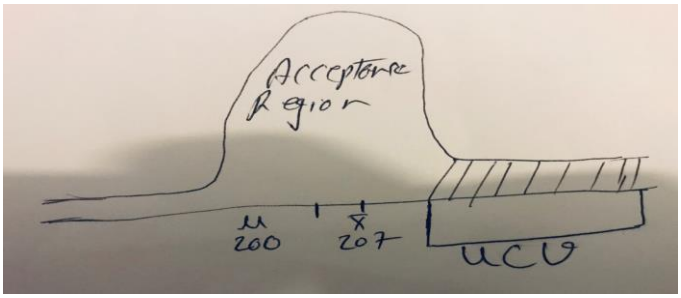
- **Critical Value Test**

The null hypothesis is your assumption about the population — it is based on the status quo. It always makes an argument about the population using the equality sign.

The null hypothesis in this case would be that the drug needs to have a time of effect less than or equal to 200 seconds. And the alternate hypothesis is that the time of effect for the drugs to be effective is greater than 200 seconds. Therefore, Null and Alternate Hypothesis can be represented as:

H_0 *: time of effect <= 200 seconds*
H_1 *: time of effect > 200 seconds*

Calculate the z-critical score for this test at 5% significance level.



This is a one tailed test. So, for 5% significance level, we would have only one critical region on the right hand side with the total area on 0.05. This mean the area till the critical point(the cumulative probability of that point) would be:

1 – 0.05 = 0.950.

Find the z-score value of 0.950 in the Z-Table.

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
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952

In The Z- Table, we 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.

Zc = 1.645

Now, let's find the critical value for "Zc = 1.645" and make the decision to accept or reject the null hypothesis.

Formula to calculate Critical Value, in this instance would be :

$$\mu + Zc * (\sigma/\sqrt{N})$$

Here, we have:

N = 100
$\mu_x = \mu = 200$,
$\sigma = 65$

Plug in these values in the formula.

$200 + 1.645 * (65/\sqrt{100})$

Critical Value (UCV) = $200 + 1.645 * (65/10)$
$= 200 + 1.645 * 6.5$
$= 200 + 10.6925$
UCV = 210.6925 seconds.

Since, the sample mean 207 seconds is less than the Critical Value(210.69 seconds), we fail to reject the Null hypothesis. This mean the drug is doing a satisfactory job and there is no need to raise any alarm.

- **P-Value Test**

The null hypothesis in this case would be that the drug needs to have a time of effect less than or equal to 200 seconds. And the alternate hypothesis is that the time of effect for the drugs to be effective is greater than 200 seconds. Therefore, Null and Alternate Hypothesis can be represented as:

H_0 : time of effect <= 200 seconds
H_1 : time of effect > 200 seconds

There are following three steps involved in order to make a decision using P-Value method.

1. *Calculate the value of the z-score for the sample mean point on the distribution.*
2. *Calculate the p-value from the cumulative probability for the given z-score using the z-table.*
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).*

Step#1: *Calculate the value of the z-score for the sample mean point of the distribution. Calculate the z-score for the sample mean ($\overline{X}$) = 207 seconds.*

Formula to calculate z-score for the sample mean $\overline{X}$:

$$\overline{X} - \mu / (\sigma/\sqrt{n})$$

Here, we have

$\overline{X}$ = 207
μ = 200
σ = 65
n = 100

Plug in these values in the formula

z-score = 207 – 200/(65/$\sqrt{100}$)
= 7/(65/10)
= 7/6.5
= 1.0769
So, z-score = 1.0769

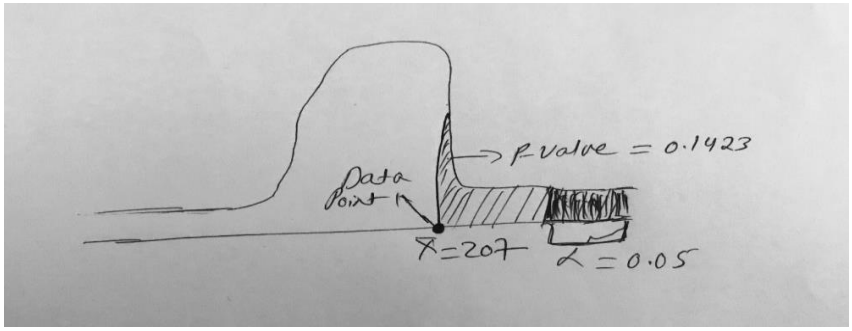
Step#2: *Find out the P-Value for the z-score 1.0769(corresponding to the sample mean on 207 seconds)*

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

As, can be seen, the value in the Z-table corresponding to 1.0 on the vertical axis and 0.07 on the horizontal axis is 0.8577. Since, the sample mean in on the right-hand side and this is a one-tailed test, the p-value would be

p-value = 1 – 0.8577
p-value = 0.1423

Step#3: Make A Decision



Here, the p-value comes out to be 0.1423. Here, the p-value is greater than the significance level (0.1423 > 0.05) and a greater p-value gives you greater evidence towards the null hypothesis. Therefore, we fail to reject the null hypothesis that the time of effect of the drug is less than 200 seconds.

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(b):

Case#1: *An instance where conducting the hypothesis testing with both α and β values fixed at 0.15 is preferred. In other words, increase in Type-1 error.*

Example:

A screening test for a serious but curable disease is similar to hypothesis testing. In this instance, screening test of stage-0 cancer in being tested on a person. Here, the null hypothesis would be that the person does not have the disease, and alternate hypothesis would be that the person has the disease. If the null hypothesis is rejected, it means the stage-0 cancer is detected and the treatment will be provided to the person. Otherwise, it will not. Assuming, the treatment has no serious effect.

Test #1

- *Type 1 error has 5% chance i.e. α = 0.05 of detecting stage-0 cancer, when the person does not have it.*
- *Type 2 error has 45% chance i.e. β = 0.45 of not detecting stage-0 cancer, when the person has it.*

Test #2

- *Type 1 error has 15% chance i.e. α = 0.15 of detecting stage-0 cancer, when the person does not have it.*
- *Type 2 error has 15% chance i.e. β = 0.15 of not detecting stage-0 cancer, when the person has it.*

Here, a type-I error would be providing treatment on false detection of the stage-0 cancer, when, in fact, the person does not have the stage-0 cancer. And a type-II error would be not providing treatment upon failing to detect the stage-o cancer, when, in fact, the person has it. Since the treatment has no serious side effects, a type-I error poses a lower health risk than a type-II error, as not providing treatment to a person who actually has the stage-0 cancer would increase his/her health risk.

Case#2: *An instance where conducting the hypothesis testing with both α = 0.05 and* β = 0.45 *values is preferred. In other words, decrease in Type-1 error.*

Example:

Starbucks Coffee chain is planning to open new cafes across different cities in India. The owners are looking out to shortlist the cities for the same. They plan to conduct surveys among the residents of few cities and they decide to open the new cafes only in those cities, where the evidence of the demand is high. Here, the null hypothesis would be that the demand in the city is not high. While, the alternate hypothesis would be that the demand in the city is high. Two different companies (Company_1 and Company_2) have been lined up to conduct the surveys in the various cities.

The type-1 and type-2 errors for the surveys are as follows:

***Company_1**:*

- *Type 1 error has 15% chance i.e. α = 0.15 of selecting a city, where the demand is actually not high.*
- *Type 2 error has 15% chance i.e. β = 0.15 of not selecting a city, where the demand is actually high.*

***Company_2**:*

- *Type 1 error has 5% chance i.e. α = 0.05 of selecting a city, where the demand is actually not high.*
- *Type 2 error has 45% chance i.e. β = 0.45 of not selecting a city, where the demand is actually high.*

Here, type-1 error would be Starbucks selecting a city, where the demand was not actually high and type-2 error would be Starbucks not selecting the city, where there was actually high. In such cases, committing type-1 error would be costly for Starbucks as they end up losing a lot of money and low revenue by opening new cafes in cities, which do not have high demand. Therefore, they go with the company_2, which has lower significance level of α = 0.05, thereby reducing the probability of type-1 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.

Answer 4:

A/B testing (also known as split testing or bucket 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.

In this instance, there are two taglines proposed, one tagline can be shown for a specific set of audience (controlled version) and another tagline for another set of audience (variant version). As different set of audience are shown different set of ads, their responses to these 2 ads can be measured statistically and thereby used to determine which ad campaign gets converted into sales and generates revenue.

Why to Use A/B Testing

Following are the few reasons- why A/B testing should be used:

- *A/B testing allows individuals, teams, and companies to make careful changes to their user experiences while collecting data on the results. This allows them to construct hypotheses, and to learn better why certain elements of their experiences impact user behaviour.*
- *Get more conversion while investing less. ROI from A/B testing can be significant with minor changes resulting in a significant increase in conversions*
- *A/B testing is completely data driven with no room for guesswork.*
- *A/B testing allows for maximum output with minimal modifications, resulting in increased ROI.*

Steps involved in A/B Testing

Step#1: Research/Data Collection
The analytics will often provide insight into where you can begin optimizing. It helps to begin with high traffic areas of your site or app, as that will allow you to gather data faster. Look for pages with low conversion rates or high drop-off rates that can be improved.

Step#2: Form Hypothesis
Based on the data collection/insights, a hypothesis should be built. The hypothesis can be arrived at by determining what should be the final result, statistics on the user behaviour to which type of ad etc. The hypothesis in this case should be built with the main purpose of increasing conversions.

Step#3: Create Variations
Next step is to create a variation based on the hypothesis, and A/B test it against the existing version (control). A variation is another version of your existing version with changes that you want to test.

Step#4: Run Experiment
Start experiment and wait for visitors to participate. At this point, visitors to site or app will be randomly assigned to either the control or variation of the experience. Their interaction with each experience is measured, counted, and compared to determine how each performs.

Step#5: Analyze Results
Once the experiment is complete, it's time to analyze the results. A/B testing software will present the data from the experiment and show the difference between how the two versions of the page performed, and whether there is a statistically significant difference. if the test succeeds, deploy the winning variation. If the test remains inconclusive, draw insights from it and implement these in susequent test.