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.

The **Binominal Distribution** can be used to calculate the probability of an event as this satisfies below three condition.

1) **The total number of trials is fixed** { n=10 sample of 10 drugs }
2) **Each trial is binary, has two possible outcomes, Success and Failure**
   P (drug Produce Satisfactory results) is:
   $4x + x = 1, x = 1/5$ or 0.2
   P (drug not able to Produce Satisfactory results) is:
   $1 - 1/5 = 4/5$ or 0.8
3) **The probability of success is the same for all the trials** {Probability that at most, 3 drugs are not able to do the satisfactory job $P(X \leq 3)$}

b.) Calculate the required probability.

1) The total number of trail n=10

2) The two possible outcome successes 0.2(i.e. drugs able to produce satisfactory job) and failure 0.8 (i.e. drugs unable to produce satisfactory job).

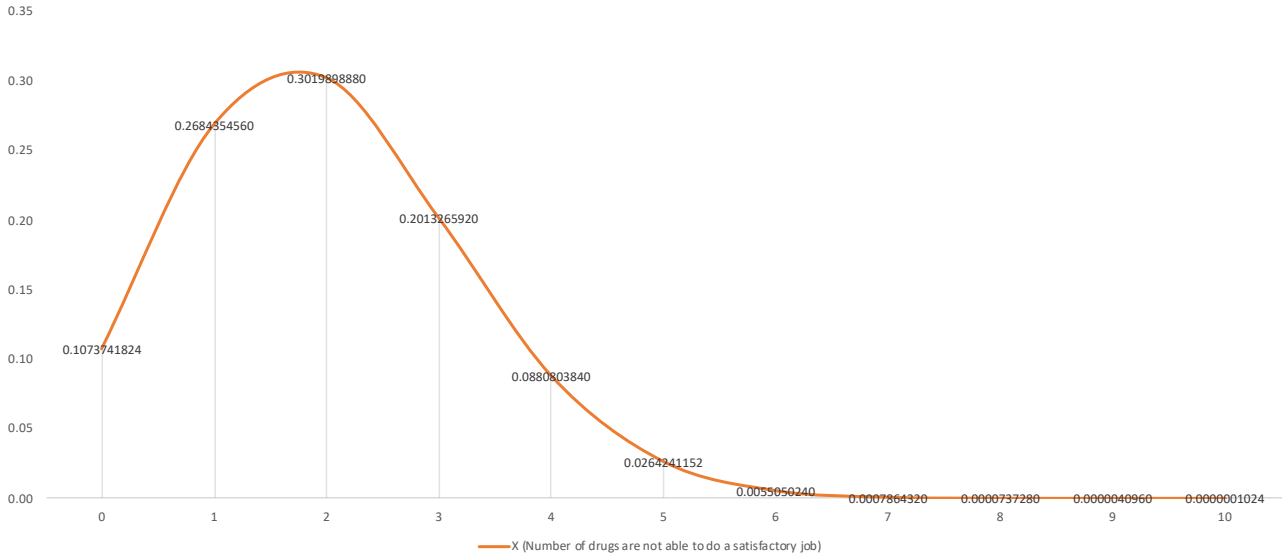
   The probability of producing satisfactory job in 1 trial is P = 0.2

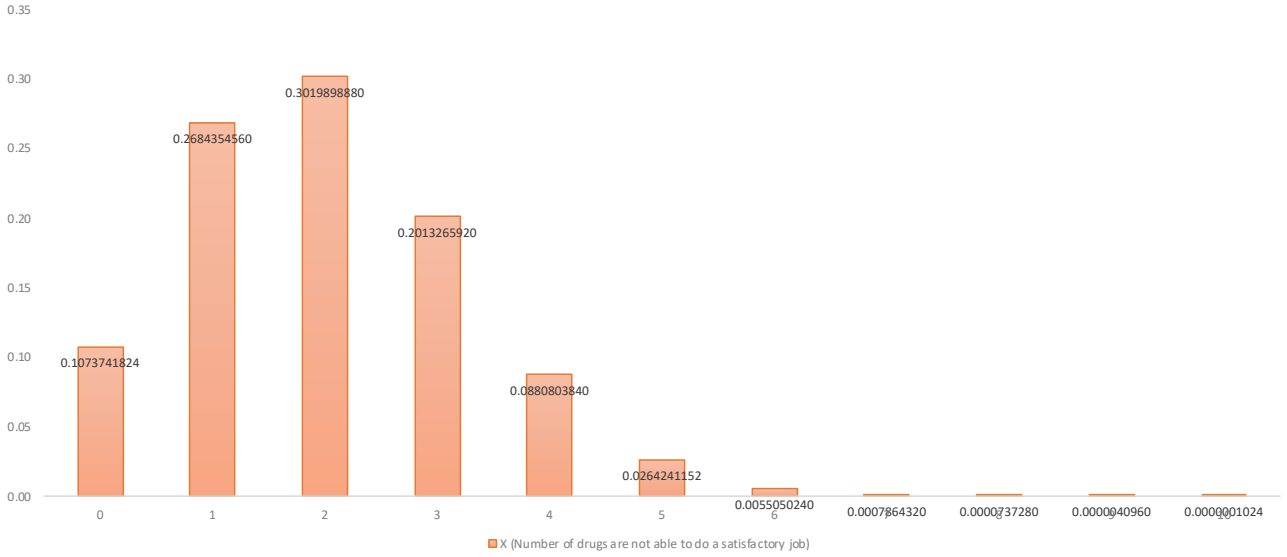
3) The probability of that at most, 3 drugs are not able to do a satisfactory job is given by –

$$P(X=x) = 10_{C_r}(0.2)^r(0.8)^{10-r}$$

x	P(X=x)
0	0.1073741824
1	0.2684354560
2	0.3019898880
3	0.2013265920
4	0.0880803840
5	0.0264241152
6	0.0055050240
7	0.0007864320
8	0.0000737280
9	0.0000040960
10	0.0000001024

P(X=0) = $10_{C_0}(0.2)^0(0.8)^{10}$ = 1 * 1 * 0.1073741824 = **0.1073741824**
P(X=1) = $10_{C_1}(0.2)^1(0.8)^9$ = 10 * 0.2 * 0.134217728 = **0.268435456**
P(X=2) = $10_{C_2}(0.2)^2(0.8)^8$ = 45 * 0.04 * 0.16777216 = **0.301989888**
P(X=3) = $10_{C_3}(0.2)^3(0.8)^7$ = 120 * 0.008 * 0.2097152 = **0.201326592**
P(X=4) = $10_{C_4}(0.2)^4(0.8)^6$ = 210 * 0.0016 * 0.262144 = **0.0880803840000001**
P(X=5) = $10_{C_5}(0.2)^5(0.8)^5$ = 252 * 0.00032 * 0.32768 = **0.0264241152**
P(X=6) = $10_{C_6}(0.2)^6(0.8)^4$ = 210 * 0.000064 * 0.4096 = **0.00550502400000001**
P(X=7) = $10_{C_7}(0.2)^7(0.8)^3$ = 120 * 0.0000128 * 0.512 = **0.000786432000000001**
P(X=8) = $10_{C_8}(0.2)^8(0.8)^2$ = 45 * 0.00000256 * 0.64 = **7.37280000000001E-05**
P(X=9) = $10_{C_9}(0.2)^9(0.8)^1$ = 10 * 0.000000512 * 0.8 = **0.000004096**
P(X=10) = $10_{C_{10}}(0.2)^{10}(0.8)^0$ = 1 * 0.0000001024 * 1 = **0.0000001024**





Cumulative probability of 3 drugs are not able to do a satisfactory job is $F(x) = P(X \leq x)$

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

$F(3) = 0.1073741824 + 0.268435456 + 0.301989888 + 0.201326592$

$F(3) = 0.8791261184$ or 87.91%

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.

Central Limit Theorem (CLT) we can estimate the range in which the population mean might lie from the sample mean and sample distribution.

Estimate the population mean time of effect of 80,000 drugs, a sample of 100 drugs and found their mean time of effect.

- Sample mean $\overline{X}$ = 207 seconds
- Sample Standard deviation S = 65 seconds

Using CLT, the sampling distribution for mean commute time will have:

- Mean = μ {unknown}
- Standard error = $\frac{\sigma}{\sqrt{n}} = \frac{S}{\sqrt{n}} = \frac{65}{\sqrt{100}} = 6.5$
- Since n (100)>30, the sample distribution is a normal distribution

Using these properties, we can claim that the probability the population mean lies between:

P (μ-13 < 207 < μ+13)

= P (207-13 < μ <207+13)

= P (194 <μ < 220)

= P (((194-207)/6.5) < μ < ((207-220)/6.5))

P (μ-13 < 207 < μ+13)

= P (-2 < μ < 2)

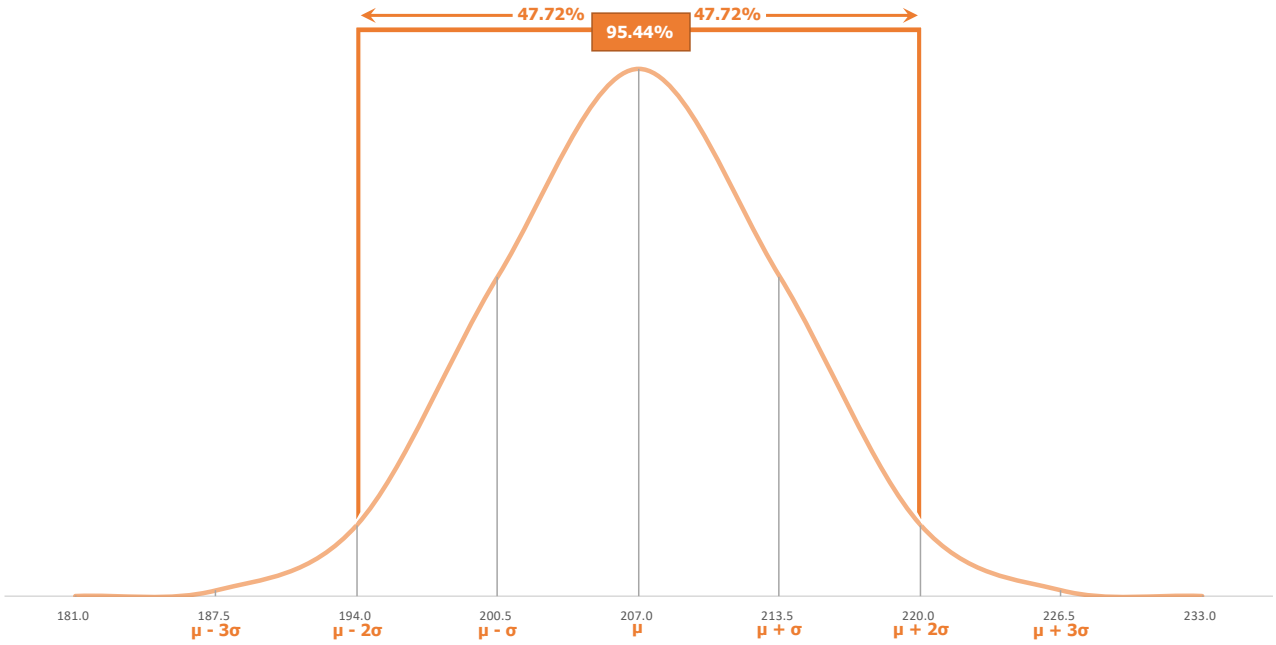
= P (Z<2) – P(Z<-2)

= 0.9772 - 0.0228

= **0.9544 or 95.44%**

The terminology related to the claim

- Confident level is **95.44%**
- Margin of error is **13 seconds**
- Confidence interval range **(194, 220)**



b.) Find the required range.

In order to generalize the entire process, we have sample population size (n=100), sample mean ($\overline{X}$ = 207 seconds) and Standard deviation (S = 65 seconds), the y% confidence interval:

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

$\boldsymbol{Z}^*$ is the z-score associated with a y% confidential level. The 95% confidence interval for the mean time of effect will be –

$$\mu = \left(\overline{X} - \frac{Z^* S}{\sqrt{n}} \, , \, \overline{X} + \frac{Z^* S}{\sqrt{n}}\right)$$

Here:

- $\overline{X}$ = 207 seconds
- S = 65 seconds
- n=100
- Z^* = ±1.96 (Z-score corresponding to 95% confidence Level)

$$\mu = \left(207 - \frac{1.96 * 65}{\sqrt{100}}\ ,\ 207 + \frac{1.96 * 65}{\sqrt{100}}\right)$$

$$\mu = \left(207 - \frac{1.96 * 65}{\sqrt{100}}\ ,\ 207 + \frac{1.96 * 65}{\sqrt{100}}\right)$$

$$\mu = \left(207 - \frac{127.4}{10}\ ,\ 207 + \frac{127.4}{10}\right)$$

$$\mu = (207 - 12.74\ ,\ 207 + 12.74\ )$$

$$\mu = (194.26 \text{ seconds}\ ,\ 219.74 \text{ seconds})$$

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.

Hypotheses testing is used to confirm the confirmation, can determine whether there is enough evidence to conclude if the hypotheses about a popular parameter is true or not.

Hypotheses Testing I (Critical Value Method)

1. Formulating hypotheses:

- Null Hypothesis

$$H_0 : \mu \leq 200 \text{ Seconds}$$

- Alternate Hypothesis

$$H_1 : \mu > 200 \text{ Seconds}$$

The sign ">" in the Alternate Hypothesis and the Upper tail test on right side of the distribution.

> $> H_1 \rightarrow$ Upper Tail test $\rightarrow$ Rejection region on the right side of distribution

2. Making a decision – Critical Value Method:

The first step of critical value method is to find Zc, calculate the cumulative probability of UCV from the value α of which is further used to find the z-critical value (Zc) for UCV

$\mu_{\bar{x}} = \mu = 207$

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{65}{\sqrt{100}} = 6.5$

$\alpha = 0.05$

Z-table for 0.9500 = 1.64 (0.9495) and 1.65 (0.9505)

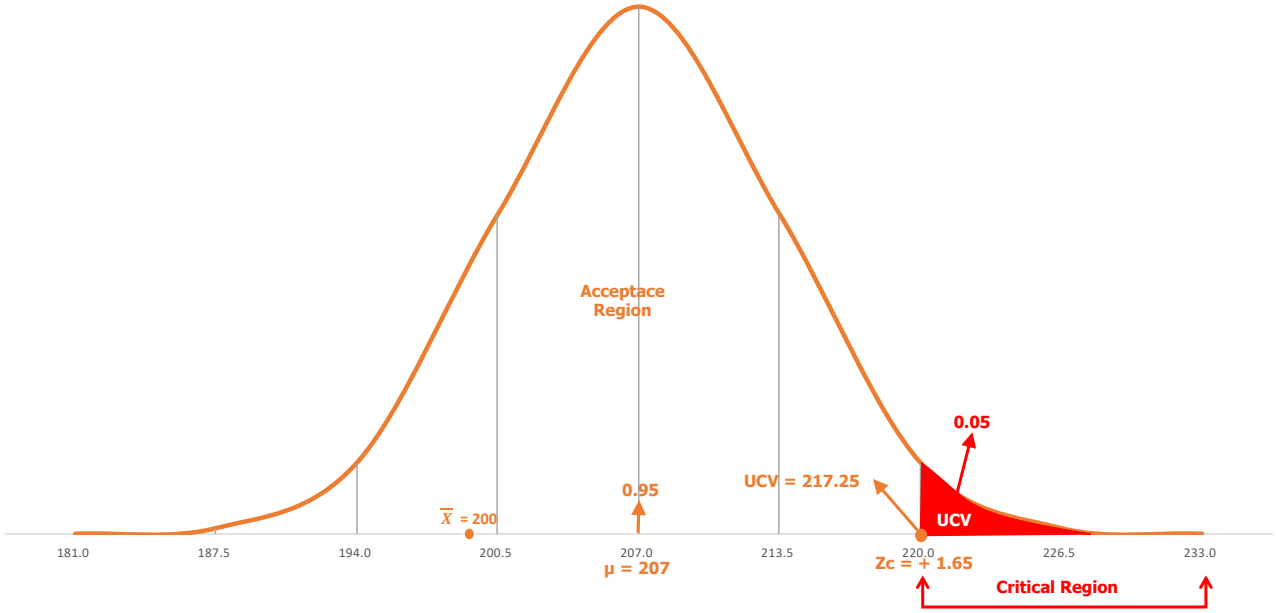
$Zc = (1 - 0.05) = 0.95 = \frac{1.64+1.65}{2} = 1.645$

$CV = \mu + (z_c * \sigma_{\bar{x}})$

UCV = 207 + (1.645 * 6.5) = 217.25

3. Decision:

As sample mean 200 lies is less than UCV i.e. it lies within the acceptance region



Decision: Fail to reject the Null Hypothesis

Hypotheses Testing II (p-value Method)

1. Calculate the value of z-score:

The first step is to calculate the value of z-score for the sample mean point on the distribution

$$\mu_{\bar{x}} = \mu = 207$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{65}{\sqrt{100}} = 6.5$$

$$\alpha = 0.05$$

$$Z = \frac{\overline{X} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{200-207}{6.5} = \frac{-7}{6.5} = -1.08$$

2. Calculate p-value

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

Z-table for -1.08 = 0.1401

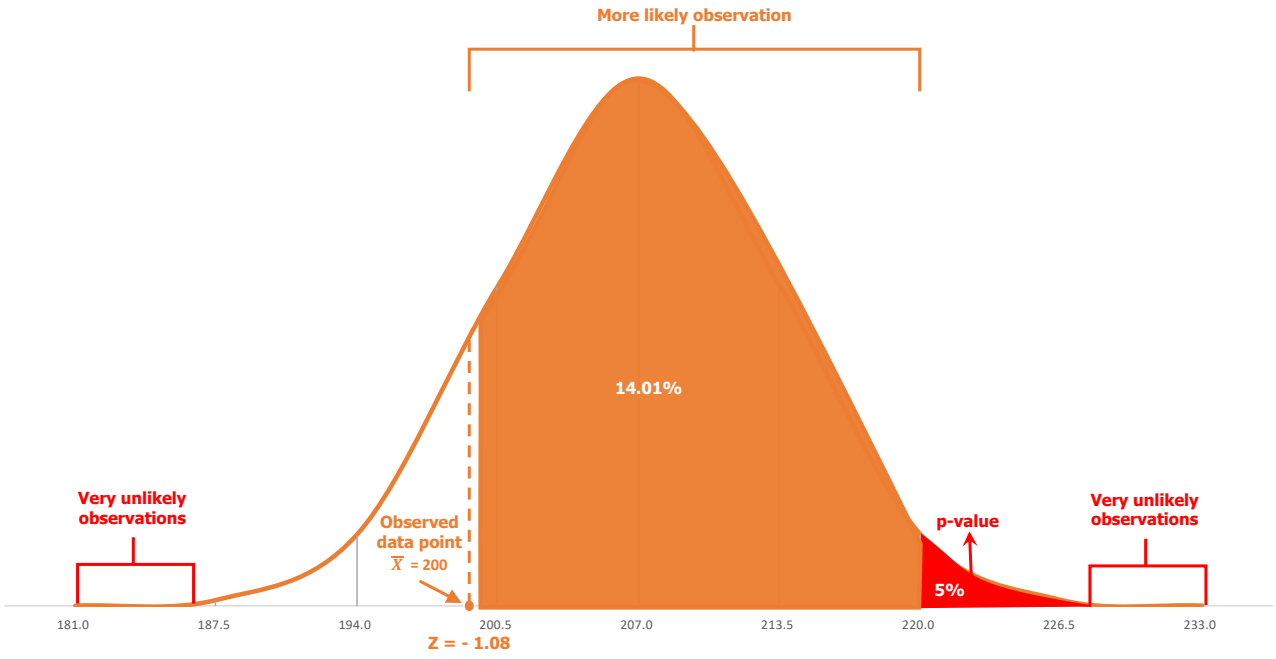
p-value = 0.1401 or 14.01%

For one tail test → p-value = 0.1401

3. Decision

As p-value (0.1401) is greater than the value of α (0.05)

p-value > α



Decision: 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).

Type I error occurs when the null hypothesis is true, but we reject it, **Type II error** occurs when the Alternate hypothesis is true, but we fail to reject it

Use Cases:

	Case I	Case II
α	0.05	0.15
β	0.45	0.15

Higher α or Higher β are depends upon the situations

Use Case I:

college is curious if they have to build cafeteria that the student interested in building cafeteria is 8%. The college asked about 36 students, show strong evidence interested in meal menu came out to be 7%. The population standard deviation is 3% with Significance level $\alpha = 0.05$ and $\beta = 0.45$

$\mu = 8$

$\sigma = 3$

$\overline{X} = 7$

n = 36

Significance level $\alpha = 0.05$ and $\beta = 0.45$

$H_0 : \mu = 8$

$H_1 : \mu \neq 8$

Two tail test

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{36}} = 0.5$$

$$Z = \frac{\bar{X} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{7-8}{0.5} = \frac{-1}{0.5} = -2$$

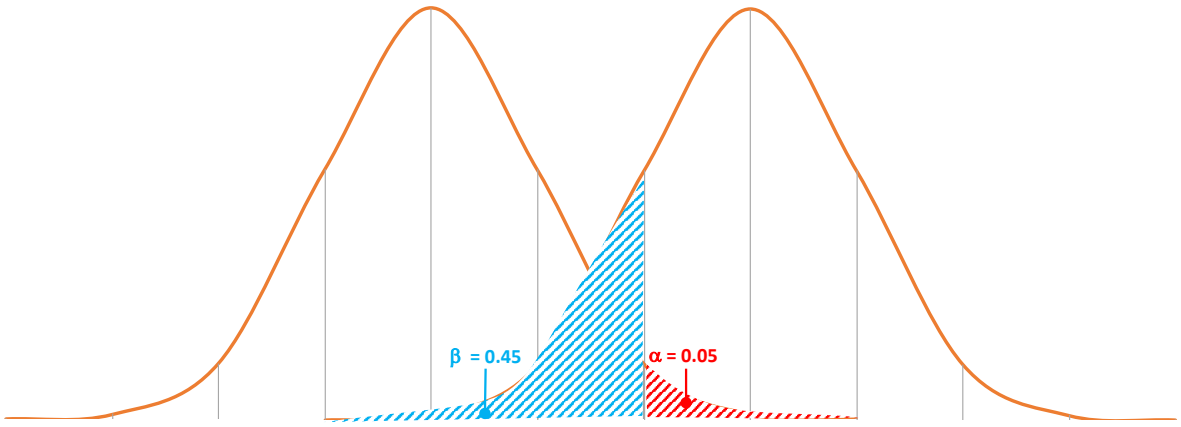
Z-table for -2 = 0.228

p-value (for two tail test) = 2 * 0.456 or 45.60%

p-value > α

Decision: Fail to reject the Null Hypothesis

	Don't consider building cafeteria	Consider building cafeteria
Student interested in meal plan	Type I error rejecting a true null hypothesis α	Correct decision
Student not interested in meal plan	Correct decision	Type II error fail to reject a false null hypothesis β



In this use case I - higher β (0.45) would increase would reduce Type I error. Consider building cafeteria however students are not interested in meal plan

Use Case II:

Student claims that the average life of bus pass is 36 months required renewal in 3rd year, as an auditor selects a sample of 49 students bus pass and calculate the average life to be 34.5 months. The population standard deviation is 4 months and Significance level $\alpha = 0.05$ and $\beta = 0.45$

$\mu = 36$

$\sigma = 4$

$\overline{X} = 34.5$

n = 49

Significance level $\alpha = 0.15$ and $\beta = 0.15$

$H_0 : \mu = 36$ months

$H_1 : \mu \neq 36$ months

Two tail test

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{49}} = 0.57$$

$$Z = \frac{\bar{X} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{34.5 - 36}{0.57} = \frac{-1.5}{0.57} = -2.63$$

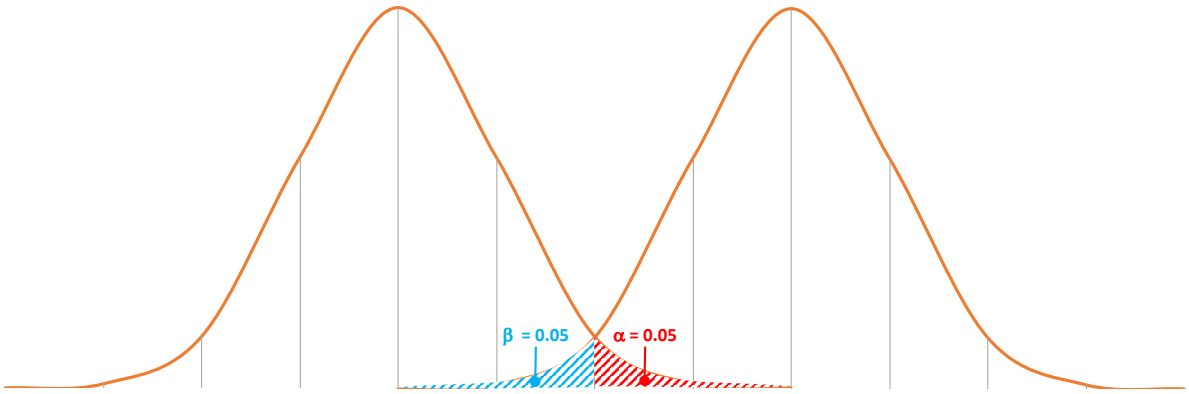
Z-table for -2.63 = 0.0043

p-value (for two tail test) = 2 * 0.0086 or 0.86%

p-value > α

Decision: Fail to reject the Null Hypothesis

	Bus pass renewal required	Bus pass renewal not required
Student Bus pass is valid for 36 months	Type I error rejecting a true null hypothesis α	Correct decision
Student Bus pass is not valid for 36 months	Correct decision	Type II error fail to reject a false null hypothesis β



In this use case increasing α would reduce the Type II error and increasing β would reduce Type I chance of have Type I and Type II is 50%

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.

A/B testing is direct industry application of the tow sample portion test sample.
Two sample test portion tests is used when Two taglines were proposed for the campaign, and the team is currently divided on which option to use.

- Take two webpage ad campaign have one tagline each webpage
- The half of sales from webpage 1 and another half from webpage 2
- By measuring sale impact that based on tagline metrics, you can ensure that which tagline produces positive results.