**CBA: Practice Problem Set 2**

**Topics: Sampling Distributions and Central Limit Theorem**

1. Examine the following normal Quintile plots carefully. Which of these plots indicates that the data …
2. Are nearly normal?

**C :  In a normal quantile plot, if the data points closely follow a straight line without any significant deviations or bends, it suggests that the data is nearly normally distributed**

1. Have a bimodal distribution? (One way to recognize a bimodal shape is a “gap” in the spacing of adjacent data values.)

**B: A bimodal distribution will have two distinct peaks or modes in the plot, indicating that the data has two different groups or sub-populations.**

1. Are skewed (i.e. not symmetric) ?

**A,C,D: A skewed distribution will have a longer tail on one side of the plot, suggesting that the data is not symmetric around the center**

1. Have outliers on both sides of the center?

**A: Outliers are data points that significantly deviate from the overall pattern in the plot. If there are outliers on both sides of the center, it indicates that the data has outliers in both the lower and upper tails**



1. For each of the following statements, indicate whether it is True/False. If false, explain why.

The manager of a warehouse monitors the volume of shipments made by the delivery team. The automated tracking system tracks every package as it moves through the facility. A sample of 25 packages is selected and weighed every day. Based on current contracts with customers, the weights should have *μ* = 22 lbs. and *σ* = 5 lbs.

1. Before using a normal model for the sampling distribution of the average package weights, the manager must confirm that weights of individual packages are normally distributed.

**False:** The statement is incorrect. Before using a normal model for the sampling distribution of the average package weights, the manager must confirm that the weights of individual packages are normally distributed.

ii) The standard error of the daily average SE() = 1.

**False:** The statement is incomplete. The standard error of the daily average depends on the standard deviation of the population, the sample size, and the distribution of the population. Without knowing the sample size and the population distribution, we cannot determine the standard error of the daily average.

1. Auditors at a small community bank randomly sample 100 withdrawal transactions made during the week at an ATM machine located near the bank’s main branch. Over the past 2 years, the average withdrawal amount has been $50 with a standard deviation of $40. Since audit investigations are typically expensive, the auditors decide to not initiate further investigations if the mean transaction amount of the sample is between $45 and $55. What is the probability that in any given week, there will be an investigation?
2. 1.25%
3. 2.5%
4. 10.55%
5. 21.1%
6. 50%

**Code:**

import scipy.stats as stats

import scipy.stats as stats

import numpy as np

mean = 50

std = 40

n = 100

d\_f = 100 - 1

t\_forty\_five = (45-50)/(40/np.sqrt(100))

t\_fifty\_five = (55-50)/(40/np.sqrt(100))

forty\_five = stats.t.cdf(t\_forty\_five, df = d\_f)

fifty\_five = stats.t.cdf(t\_fifty\_five, d\_f)

prob = fifty\_five - forty\_five

np.round(stats.t.interval(confidence = prob, df = d\_f, loc = mean, scale = std/np.sqrt(n)))

print('The probability that in any given week, there will be an investigation is',np.round((1-prob)\*100,1),'%')

**Output : 21.4%**

1. The auditors from the above example would like to maintain the probability of investigation to 5%. Which of the following represents the minimum number transactions that they should sample if they do not want to change the thresholds of 45 and 55? Assume that the sample statistics remain unchanged.
2. 144
3. 150
4. 196
5. 250
6. Not enough information

**Code:**

import scipy.stats as stats

import numpy as np

# Import the correct function from scipy.stats

from scipy.stats import t

# Define variables

x\_bar = 45

s\_std = 40

mew = 50

# Calculate the t-score

t\_score = stats.t.ppf(0.025, df=249)

# Calculate the sample size

n = ((s\_std \* abs(t\_score)) / (mew - x\_bar))\*\*2

# Print the sample size

print('The Auditors would like to maintain the probability of investigation to 5%, they should sample', np.round(n), 'transactions if they do not want to change the thresholds of 45 to 55')

# Calculate the degrees of freedom

df = n - 1

# Calculate the confidence interval

confidence\_interval = stats.t.interval(confidence=0.95, df=df, loc=mew, scale=s\_std / np.sqrt(n))

# Print the confidence interval

print('The 95% confidence interval for the population mean is:', np.round(confidence\_interval, 2))

**Output : 248**

1. An educational startup that helps MBA aspirants write their essays is targeting individuals who have taken GMAT in 2012 and have expressed interest in applying to FT top 20 b-schools. There are 40000 such individuals with an average GMAT score of 720 and a standard deviation of 120. The scores are distributed between 650 and 790 with a very long and thin tail towards the higher end resulting in substantial skewness. Which of the following is likely to be true for randomly chosen samples of aspirants?
2. The standard deviation of the scores within any sample will be 120.

**True**, as the standard deviation of the population is given as 120. In statistics, the standard deviation measures the amount of variation or dispersion in a set of values. So, within any sample, the standard deviation would remain 120.

1. The standard deviation of the mean of across several samples will be 120.

**Incorrect,** The standard deviation of the mean across several samples is calculated by dividing the population standard deviation by the square root of the sample size. It's essentially the standard error of the mean. So, the standard deviation of the mean across several samples will be less than 120 and will decrease as the sample size increases

1. The mean score in any sample will be 720.

**False**, While the population mean is given as 720, the mean score in any sample may vary due to random sampling. However, the mean of all samples combined (the sample mean) would tend towards the population mean as the sample size increases due to the central limit theorem

1. The average of the mean across several samples will be 720.

**True,** As mentioned earlier, the sample mean tends towards the population mean as the sample size increases. Therefore, the average of the mean across several samples is expected to be close to the population mean

1. The standard deviation of the mean across several samples will be 0.60

**Incorrect,** The standard deviation of the mean across several samples, also known as the standard error of the mean, is calculated as the population standard deviation divided by the square root of the sample size. It won't be 0.60 unless the sample size is given