**CBA: Practice Problem Set 2**

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

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

Soln: C

If there is a one-to-one relationship between the data and the theoretical quantiles, then the data follow a nearly normal distribution.

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

Soln: B

A bimodal distribution has two peaks. In the context of a continuous probability distribution, modes are peaks in the distribution. The graph below shows a bimodal distribution.

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

Soln: A C D

If one tail is longer than another, the distribution is skewed. These distributions are sometimes called asymmetric or asymmetrical distributions as they don’t show any kind of [symmetry](https://www.statisticshowto.com/symmetry-of-a-function/).

A [**left-skewed distribution**](https://www.statisticshowto.com/probability-and-statistics/skewed-distribution/#SkewLeft) has a long left tail. Left-skewed distributions are also called negatively-skewed distributions.

A [**right-skewed distribution**](https://www.statisticshowto.com/probability-and-statistics/skewed-distribution/#SkewRight) has a long right tail. Right-skewed distributions are also called positive-skew distributions.

1. Have outliers on both sides of the center?

Soln: A

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.



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.

Soln: TRUE.

In this case, at least 30 sample packages must be selected and weighed everyday. Based on the central limit theorem, the sampling distribution of the sample mean approach normal distribution as the sample size become bigger (over 30).

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

Soln: TRUE. Standard error equal to standard deviation divided by square root of sample size = 5/sqrt(25) =1

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%

Soln: import numpy as np

from scipy import stats

from scipy.stats import norm

n=100, Pop Mean=50, Pop SD=40 As no. of samples is more than 30, we can consider it normal distribution

# For No investigation P(45)

# For Investigation 1-P(45)

# find z-scores at x=45; z=(s\_mean-P\_mean)/(p\_SD/sqrt(n))

z=(45-50)/(40/100\*\*0.5)

z = -1.25

# find z-scores at x=55; z=(s\_mean-P\_mean)/(p\_SD/sqrt(n))

z=(55-50)/(40/100\*\*0.5)

z = 1.25

# For No investigation P(45

stats.norm.cdf(1.25)-stats.norm.cdf(-1.25)

0.7887004526662893

stats.norm.interval(0.7887,loc=50,scale=40/(100\*\*0.5))

(45.00000495667348, 54.99999504332652)

# For Investigation 1-P(45)

1-0.7887

0.21130000000000004

21.1% is the probability that in any given week, there will be an investigation.

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

Soln: z=(x-mean)/sigma/sqrt(n); z-test, because sd is given for the long term  
 =+/- 5/40/sqrt(100)  
 =+/- 1.25

The probability of z between those values is 0.7887, so probability of an investigation is 1-0.7887, or 0.2113.

For 5%, z has to be +/-1.96

So 1.96=(5)\*sqrt(n)/40

Sqrt(n)=15.68

N=245.86 or 246

250 represents the minimum number transactions that they should sample if they do not want to change the thresholds of 45 and 55.

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?

Soln: The SEM is sd/sqrt(n)=120/sqrt(40000)=0.6

1. The standard deviation of the scores within any sample will be 120.

Soln: The standard deviation will not be 120 of scores in any one sample, since we don’t know the sample size.

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

Soln: The standard deviation of the mean of across several samples will also not be 120. It will be less probably about 0.6.

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

Soln: The mean score in any sample will be 720. Maybe but it can be less or more.

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

Soln: The average of the mean across several samples will be 720. This is possible but it requires mean of all the samples.

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

Soln: The standard deviation of the mean across several samples will be 0.60. This is likely, given the sample size, which even with a lot of skewness will tends towards normality given the sample size.