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 …

   I.   Are nearly normal?

        ANS: C is normal

   II. Have a bimodal distribution? (One way to recognize a bimodal shape is a "gap" in the spacing of adjacent data values.)
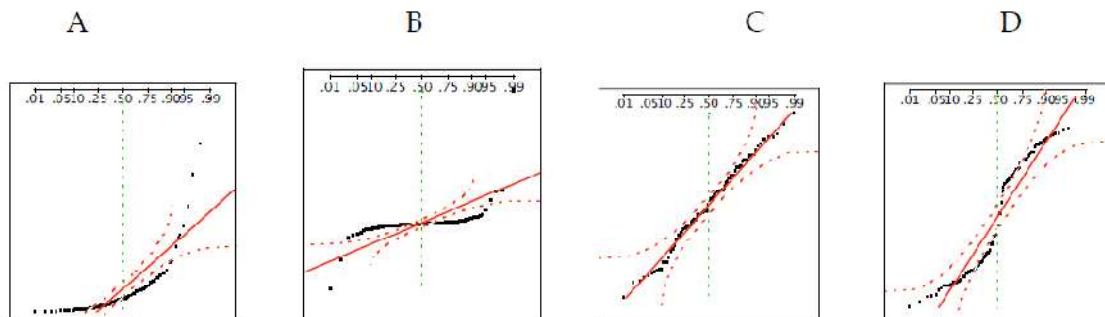
        ANS: B and D is bimadal distribution

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

        ANS: A,B and D

   IV. Have outliers on both sides of the center?

        ANS: A and B

2. For each of the following statements, indicate whether it is <u>True/False</u>. 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.

   (i) 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.

        ANS:

        False

        A sampling distribution is a probability distribution of a statistic obtained from a larger number of samples drawn from a specific population.

        In our case the samples contain 25 packages and the larger number of samples contain of each such 25 packages taken into different samples (25+25+25+25…and so on).

The mean for one these samples is 22lbs and standard deviation of 5lbs which means each individual package is having a weight varying between + or – 5lbs with respect to mean(22lbs).

Hence it is invalid to take a weight of individual packages and confirm that it follows normal distribution before using a normal model for the sampling distribution.

The Sample Central Limit Theorem states that the sampling distribution of the samples mean approaches normal distribution as the sample size is large enough.

(ii)   The standard error of the daily average SE($\bar{x}$) = 1.

   ANS:

   True

   Standard Error=Sample standard deviation/square root of(number of samples)

   $$SE=5/(25)^2=1$$

3. 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?

   A. 1.25%
   B. 2.5%
   C. 10.55%
   D. 21.1%
   E. 50%

   ANS: D

   MEAN=50,std=40,n=100,df=100-1=99

```
  #T-distribution
  t_forty_five=(45-50)/(40/np.sqrt(100))

[6] t_fifty_five=(55-50)/(40/np.sqrt(100))

[7] forty_five=stats.t.cdf(t_forty_five,df=99)

[8] fifty_five=stats.t.cdf(t_fifty_five,df=99)

[10] probability=fifty_five-forty_five

[45] #np.round(stats.t.interval(alpha = prob, df = d_f, loc = mean, scale = std/np.sqrt(n)),)
     confidence = 1 - probability
     np.round(stats.t.interval(confidence, df=99, loc=mean, scale=std/np.sqrt(n)),)

     array([49., 51.])

[44] print('The probability that in any given week, there will be an investigation is',np.round((confidence)*100,1),'%')

     The probability that in any given week, there will be an investigation is 21.4 %
```

4. 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.

   A. 144
   B. 150
   C. 196
   D. 250
   E. Not enough information
   ANS: D
   Sample=45
   Mean=50
   Std=40

```
[52] t = np.round(stats.t.ppf(0.025, df = 249),2)
     t

     -1.97
```

```
  # t_value = (sample - mean)/(std/n**0.5)
  # t = 45-50 or z = 55-50 z = +/- 5

  # t = 5/(40/n**0.5)
  # n = (sample_standard_deviation*tscore)/(sample_mean=population_mean)
  n = ((std*abs(t)) / (5))**2
```

```
[57] print('The Auditors would like to maintain the probability of investigation to 5%, they should sample',np.round(n,),'trnsactions')

     The Auditors would like to maintain the probability of investigation to 5%, they should sample 248.0 trnsactions
```

5. 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?

   A. The standard deviation of the scores within any sample will be 120.
   B. The standard deviation of the mean of across several samples will be 120.
   C. The mean score in any sample will be 720.
   D. The average of the mean across several samples will be 720.
   E. The standard deviation of the mean across several samples will be 0.60
   ANS: E
   Standard deviation=120,Sample size =40000
   Standard error=standard deviation/(sample size)^0.5=120/40000^0.5=120/200=0.6