Name: B.R.Nivedha

SET:1

Topics: Descriptive Statistics and Probability

1. Look at the data given below. Plot the data, find the outliers and find out μ, σ, σ^2

Name of company	Measure X
Allied Signal	24.23
Bankers Trust	25.53
General Mills	25.41
ITT Industries	24.14
J.P.Morgan& Co.	29.62
Lehman Brothers	28.25
Marriott	25.81
MCI	24.39
Merrill Lynch	40.26
Microsoft	32.95
Morgan Stanley	91.36
Sun Microsystems	25.99
Travelers	39.42
US Airways	26.71
Warner-Lambert	35.00

SOL:

```
: import numpy as np
  import pandas as pd
  import matplotlib.pyplot as plt
```

executed in 31ms, finished 13:37:26 2021-01-18

```
: df=pd.read_csv('measure.csv')
  df
```

executed in 31ms, finished 13:37:27 2021-01-18

:

	Name of company	Measure X
0	Allied Signal	24.23
1	Bankers Trust	25.53
2	General Mills	25.41
3	ITT Industries	24.14
4	J.P.Morgan& Co.	29.62
5	Lehman Brothers	28.25
6	Marriott	25.81

```
df=df.rename({'Name of company':'CompanyName','Measure X':'measureX'},axis=1)
df
```

executed in 47ms, finished 13:37:28 2021-01-18

	CompanyName	measureX
0	Allied Signal	24.23
1	Bankers Trust	25.53
2	General Mills	25.41
3	ITT Industries	24.14
4	J.P.Morgan& Co.	29.62
5	Lehman Brothers	28.25
6	Marriott	25.81
7	MCI	24.39

```
x= df.CompanyName
y= df.measureX
```

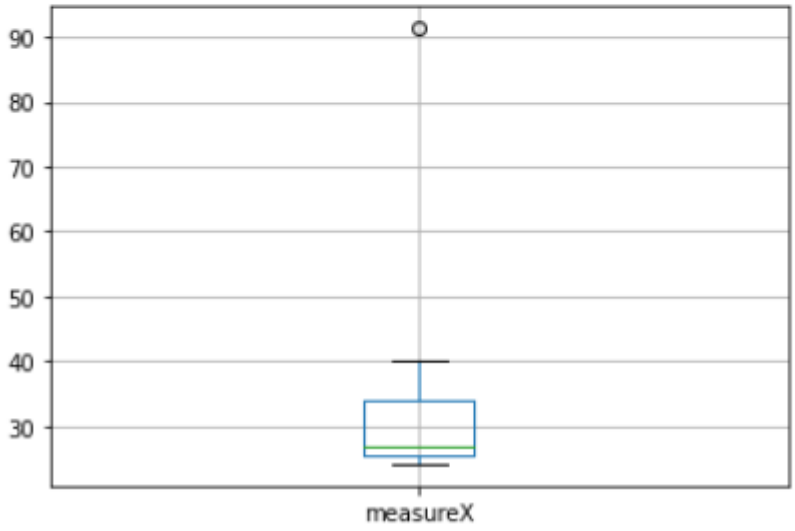
executed in 16ms, finished 13:37:58 2021-01-18

```
%matplotlib inline
```

executed in 32ms, finished 13:38:45 2021-01-18

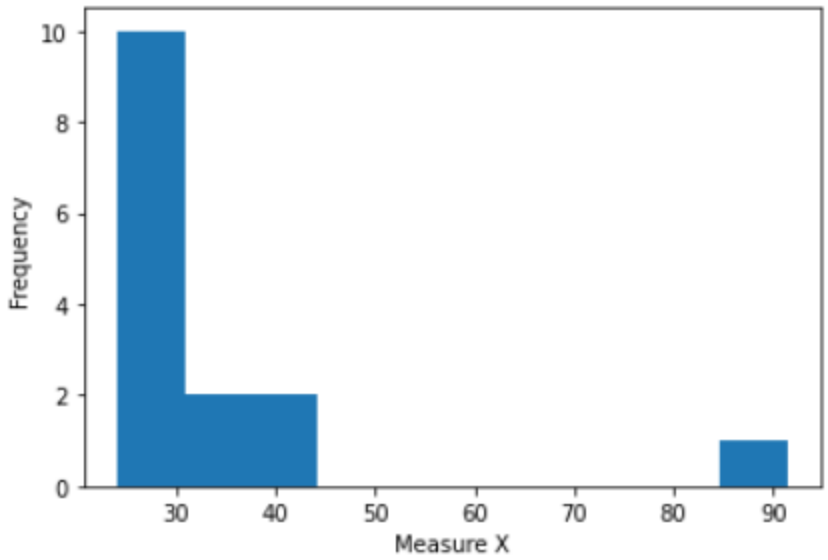
```
a = df.boxplot()
plt.show(a)
```

executed in 596ms, finished 13:42:41 2021-01-18



```
b = plt.hist(y)
plt.xlabel('Measure X')
plt.ylabel('Frequency')
plt.show(b)
```

executed in 272ms, finished 13:45:06 2021-01-18

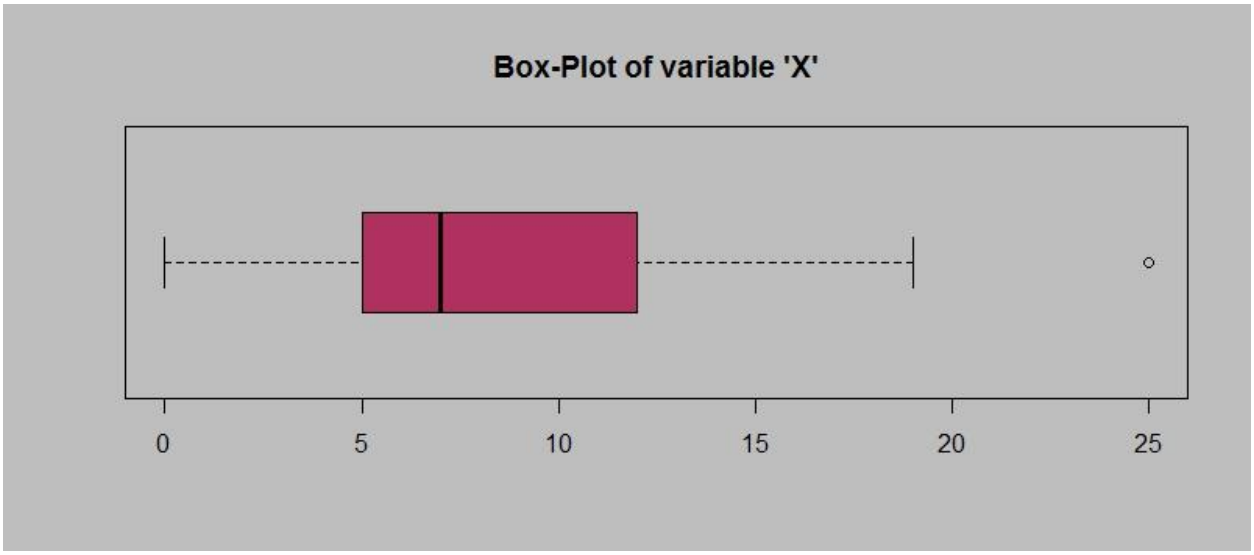



```
Result=df[['measureX']].agg(['mean','var','std'])
Result
```

executed in 140ms, finished 13:47:02 2021-01-18

	measureX
mean	33.271333
var	287.146612
std	16.945401

2.



Answer the following three questions based on the box-plot above.

(i) What is inter-quartile range of this dataset? (please approximate the numbers) In one line, explain what this value implies.

Ans: IQR ranges from 5 to 12

Q3-Q1= 12-5= 7

The Inter Quartile Range is given by the difference between the Q3 and Q1 value

IQR of a data set used to measure how spread out the data points in dataset

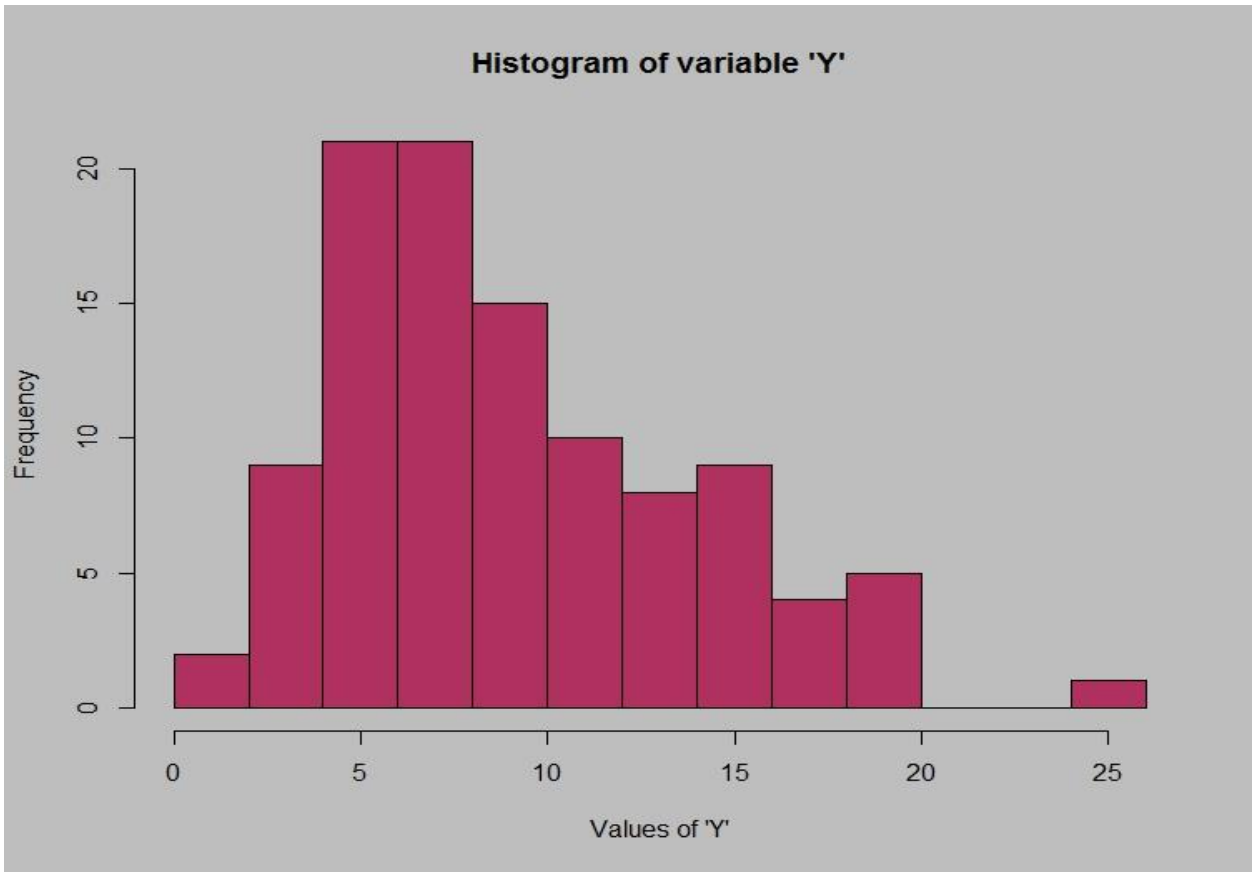
(ii) What can we say about the skewness of this dataset?

Ans: The box plot of the data is not symmetric and longer part of the box is right the median, therefore we can say that the data is skewed right

(iii) If it was found that the data point with the value 25 is actually 2.5, how would the new box-plot be affected?

Ans: The data point is a 25 is an outlier. Outlier does not effect the median of the data 2.5 is lies on the lower whisker region of the data. Q1 will be changed and also outlier of the boxplot is removed.

3.



Answer the following three questions based on the histogram above.

(i) Where would the mode of this dataset lie?

The Mode of the above dataset is approximately 4 to 8

(ii) Comment on the skewness of the dataset.

The distribution has a larger number of occurrence in the left side and few in the right side. Therefore this dataset is **Right skewed**, **Positively skewed**.

(iii) Suppose that the above histogram and the box-plot in question 2 are plotted for the same dataset. Explain how these graphs complement each other in providing information about any dataset.

Box plot of dataset gives minimum values, first quartile ,median, third quartile, maximum value of the data and also boxplot shows the outliers. But in histogram we get about mode value , skewness etc. Histogram doesnot provide idea about outliers, so we can consider that both are useful in giving the information of the data .

4. AT&T was running commercials in 1990 aimed at luring back customers who had switched to one of the other long-distance phone service providers. One such commercial shows a businessman trying to reach Phoenix and mistakenly getting Fiji, where a half-naked native on a beach responds incomprehensibly in Polynesian. When asked about this advertisement, AT&T admitted that the portrayed incident did not actually take place but added that this was an enactment of something that "could happen." Suppose that one in 200 long-distance telephone calls is misdirected. What is the probability that at least one in five attempted telephone calls reaches the wrong number? (Assume independence of attempts.)

Sol: One in 200 long-distance telephone calls is misdirected

To find: probability that at least one in five attempted telephone calls reaches the wrong number

One in 200 long-distance telephone calls is misdirected

Probability of call misdirecting p=1/200

Probability of call not misdirecting =1-1/200=199/200

Number of calls =5

$P(X)= {}^{n}C_{x}P^{x}q^{n-x}$ n=5,p=1/200,q=199/200

At least one in five attempted telephone calls reaches the wrong number

=1- none of the call reaches the wrong number

=1-p(0)

$=1-{}^{5}C_{0}(1/200)^{0}(199/200)^{5-0}$

$= 1-1*1*(199/200)^{5}$

=0.02475

Probability that at least one in five attempted calls reaches the wrong number = 0.02475

5. Returns on a certain business venture, to the nearest $1,000, are known to follow the following probability distribution

x	P(x)
-2,000	0.1
-1,000	0.1
0	0.2
1000	0.2
2000	0.3
3000	0.1

(i) What is the most likely monetary outcome of the business venture?
Sol:2000 is the most likely monetary outcome since it has highest probability

(ii) Is the venture likely to be successful? Explain
Sol: we can say that the venture is success.
Basically the total probability is 1. Among this, probability of positive value of x is 0.2+0.3+0.1=0.6
Which is greater than the probability of negative values, which is 0.1+0.1=0.2 is likely to be not successful

(iii) What is the long-term average earning of business ventures of this kind? Explain
Sol: Long term Averge is the expected value of the venture

=(-2000*0.1)+(-1000*0.1)+(0*0.2)+(1000*0.2) +(2000*0.3)
+(3000*0.1)

= -200-100+0+200+600+300

=1100-300

=800

(iv) What is the good measure of the risk involved in a venture of this kind? Compute this measure
Sol: Variance is the good measure of the risk involved in the venture
var(x)=216000

SET:2

Topics: Normal distribution, Functions of Random Variables

1. The time required for servicing transmissions is normally distributed with μ = 45 minutes and σ = 8 minutes. The service manager plans to have work begin on the transmission of a customer's car 10 minutes after the car is dropped off and the customer is told that the car will be ready within 1 hour from drop-off. What is the probability that the service manager cannot meet his commitment?

   A. 0.3875
   B. 0.2676
   C. 0.5
   D. 0.6987

Sol: Let X be the time taken. Probability of work is completed within 50 mints is

P(X<= 50)=P(Z<=(50))

P(Z<=0.625)

=0.73401145

Probability that the service manger cannot meet his comment=1-P(X<=50)

=1-.734

=0.2659

2. The current age (in years) of 400 clerical employees at an insurance claims processing center is normally distributed with mean μ = 38 and Standard deviation σ =6. For each statement below, please specify True/False. If false, briefly explain why.
   A. More employees at the processing center are older than 44 than between 38 and 44.

      Sol: Given that age of employees are normally distributed with mean mu=38 and standard deviation=6

      Probability of employees greater than age of 44 = P(X>44)

      =1-P(X)

      =1-.841345

      =0.1586

      Probability of number of employees between 38 and 44 years of age

      P(38)=p(X)

=.841345 - 0.5
=0.3413545
Therefore probability of number of employees between 38 and 44years of age is greater age 44 so given statement is proven as false

B. A training program for employees under the age of 30 at the center would be expected to attract about 36 employees.

Probability of number employees less than age 30=P(X)
=P(Z<30)
=P(Z<-1.333)
=0.0912=9.12%
So the number employees less than 30 is 9.12% of the total number of employees which is equal to 0.0912*200=36.48=36 so statement can be considered as true

3. If $X_1 \sim N(\mu, \sigma^2)$ and $X_2 \sim N(\mu, \sigma^2)$ are *iid*normal random variables, then what is the difference between 2 X_1 and $X_1 + X_2$? Discuss both their distributions and parameters.

Sol: As we know that if $X \sim N(\mu 1, \sigma^2)$ and $Y \sim N(\mu, \sigma^2)$ are two independent random variables then Given that $X1 \sim N(\mu 1, \sigma^2)$ and $X\,2 \sim N(\mu 1, \sigma^2)$
Therefore $2X1 \sim N(2\mu, 4\sigma 2)$ and $X\,1 + X\,2 \sim N(2\mu, 2\sigma 2)$

4. Let $X \sim N(100, 20^2)$. Find two values, *a* and*b*, symmetric about the mean, such that the probability of the random variable taking a value between them is 0.99.

A. 90.5, 105.9
B. 80.2, 119.8
C. 22, 78
D. 48.5, 151.5
E. 90.1, 109.9

Sol: we need to find out the values of a and b which are symmetric about the mean, such that the probability of random variables taking a value between them is 0.99

The probability of getting value between and b should be 0.99, so the probability of going wrong or the probability outside the a and b area is 0.01(ie.1-0.99).

The probability towards left from a=-0.005(ie.0.01/2).
The probability towards right from b =+0.005(ie. 0.01/2)
So since we have the probabilities of a and b, we need to calculate X thus we want to find the 0.5^{th} and 99.5^{th} percentiles
By finding the standard normal variable Z(Zvalue), we can calculate the X values.
$Z=(X-\mu)/\sigma$
For probability 0.005 the Z value is -2.57(from ZTable).
0.5^{th} and 99.5^{th} the percentile value is
$Z^* \sigma+ \mu = X$
a=Z(+0.005)*20+100=(-2.57)*20+100 =48.6
b=Z(-0.005)*20+100=-(-2.57)*20+100=151.4
so option D is correct

5. Consider a company that has two different divisions. The annual profits from the two divisions are independent and have distributions $Profit_1 \sim N(5, 3^2)$ and $Profit_2 \sim N(7, 4^2)$ respectively. Both the profits are in $ Million. Answer the following questions about the total profit of the company in Rupees. Assume that $1 = Rs. 45

    A. Specify a Rupee range (centered on the mean) such that it contains 95% probability for the annual profit of the company.
    Sol: Total profit of the company =(profit 1+ profit 2)~ N((5+7),(3 2+ 4 2))= N(12,25)
    Here we want to find rupee range (centered on the mean) such that it contains 95% probability for the annual profit of the company
    Total area under the normal curve is 1
    Therefore excluded area is1-0.95=0.05
    Excluded area is 0.025 in each of the left and right tails of the normal curve
    Thus,
    We want to find the 2.5^{th} and 97.5 percentile
    To find the 2.5^{th} percentile value of profit we have to find Z .025=-1.96
    To find the 97.5^{th} the percentile value of profit we have to find Z.975=1.96

    Therefore lower rupee range=X=(1.96*5)+12
    2.2$=2.2*45=99rs

Upper rupee range =X= (1.96*5)+12
21.8$=21.8*45=981 rs

B. Specify the 5th percentile of profit (in Rupees) for the company
Sol: to find the 5th percentile value of profit we have find Zα =Z 0.05= -1.644584
Therefore 5th percentile value of profit=
=(-1.64458*5)+12
3.77$
3.77*45=170rs

C. Which of the two divisions has a larger probability of making a loss in a given year?
Division 2
Probability of making loss means that there is a 0 profit
Zvalue for first division is Z== -5/3 = -1.667
Zvalue for second division is Z== -7/4= -1.75
Since second division has the least value of Z , it has smaller probability of making loss.

SET:3

Topics: Confidence Intervals

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

   I. The sample size of the survey should at least be a fixed percentage of the population size in order to produce representative results.

   Ans: False

   The representation of the survey results should have a sample size. The sample size must be a fixed percentage of the total population size of the survey. The size of the sample should have atleast 30 observations

   II. The sampling frame is a list of every item that appears in a survey sample, including those that did not respond to questions.

   Ans: False

   The sampling frame refers to a list of an item which responds to the question and not the ones which do not respond to the questions.

   III. Larger surveys convey a more accurate impression of the population than smaller surveys.

   Ans: True

   The larger conveys a more accurate impression of the population as larger surveys involve large sample size which reduces the chances of error.

2. *PC Magazine* asked all of its readers to participate in a survey of their satisfaction with different brands of electronics. In the 2004 survey, which was included in an issue of the magazine that year, more than 9000 readers rated the products on a scale from 1 to 10. The magazine reported that the average rating assigned by 225 readers to a Kodak compact digital camera was 7.5. For this product, identify the following:

   A. The population

   Readers of the magazine=900

P=x/n=225/9000=0.025

B. The parameter of interest
Sample size, average, scale
Rating of the camera=7.5

C. The sampling frame
All readers of the issue where the survey was included. 9000

D. The sample size
225

E. The sampling design
Simple random sampling Voluntary response

F. Any potential sources of bias or other problems with the survey or sample
It is possible that only those who were particularly pleased or only who are displeased with the product participated in the survey which can makes result unreliable.

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

I. If the 95% confidence interval for the average purchase of customers at a department store is $50 to $110, then $100 is a plausible value for the population mean at this level of confidence.
It is correct to say that there is a 95% chance that the confidence interval you calculated contains the true population mean.

II. If the 95% confidence interval for the number of moviegoers who purchase concessions is 30% to 45%, this means that fewer than half of all moviegoers purchase concessions.
False
A 95% confidence interval is a range of values that you can be 95% certain contains the true mean of the population.

III. The 95% Confidence-Interval for μ only applies if the sample data are nearly normally distributed.

False

95% Confidence-Interval for μ only applies if the sample data are normally distributed that means sample size at least larger than 30

4. What are the chances that $\overline{X} > \mu$?

A. ¼
B. ½
C. ¾
D. 1

Sample means from the population will be approximately normal as long as sampl size is large

5. In January 2005, a company that monitors Internet traffic (WebSideStory) reported that its sampling revealed that the Mozilla Firefox browser launched in 2004 had grabbed a 4.6% share of the market.

I. If the sample were based on 2,000 users, could Microsoft conclude that Mozilla has a less than 5% share of the market?

(I) We conclude that Mozilla has more than or equal to 5% share of the market.

(II) Yes, we can conclude that Mozilla has a less than 5% share of the market.

In January 2005, a company that monitors Internet traffic (WebSideStory) reported that its

Let **p** = *population proportion share of the market by Mozilla*

So, **Null Hypothesis**, H0: p >5%{means that Mozilla has more than or equal to 5% share of the market}

Alternate Hypothesis, H1 : p < 5% {means that Mozilla has a less than 5% share of the market} The test statistics that will be used here is **One-sample z-test for proportions**;

T.S. = $\hat{p}$ -p/√p(1-p)/n ~ N(0,1)

where, $\hat{p}$ = sample proportion of the share of the market grabbed

by Mozilla in 2004 = 4.6% n = sample of users = 2,000

So, the test statistics =0.046-0.05/√0.05(1-0.05)/2000

= -0.821

The value of z-test statistics is -0.821.

Since ,in the question we are not given with the level of significance so we assume it to be 5%. **Now, at 5% level of significance the z table gives a critical value of -1.96 for left-tailed test.**

Since the value of our test statistics is more than the critical value of z, so we have *insufficient evidence to reject our null hypothesis* as it will not fall in the rejection region.

Therefore, we conclude that Mozilla has more than or equal to 5% share of the market.

II. WebSideStory claims that its sample includes all the daily Internet users. If that's the case, then

can Microsoft conclude that Mozilla has a less than 5% share of the market?

Sol:

We are given that WebSideStory claims that its sample includes all the daily Internet users. This means that the 4.6% share of the market represents the whole population.

Hence, we can conclude that Mozilla has a less than 5% share of the market

6. A book publisher monitors the size of shipments of its textbooks to university bookstores. For a sample of texts used at various schools, the 95% confidence interval for the size of the shipment was 250 ± 45 books. Which, if any, of the following interpretations of this interval are correct?

A. All shipments are between 205 and 295 books.
Incorrect

B. 95% of shipments are between 205 and 295 books.
Incorrect

C. The procedure that produced this interval generates ranges that hold the population mean for 95% of samples.
Correct

A 95% CI is a range of values that you can be 95% certain contains the true mean of the population

D. If we get another sample, then we can be 95% sure that the mean of this second sample is between 205 and 295.
Incorrect

E. We can be 95% confident that the range 160 to 340 holds the population mean.
Incorrect

7. Which is shorter: a 95% *z*-interval or a 95% *t*-interval for μ if we know that σ =s?

A. The z-interval is shorter
B. The t-interval is shorter
C. Both are equal
D. We cannot say

Questions 8 and 9 are based on the following: To prepare a report on the economy, analysts need to estimate the percentage of businesses that plan to hire additional employees in the next 60 days.

8. How many randomly selected employers (minimum number) must we contact in order to guarantee a margin of error of no more than 4% (at 95% confidence)?

A. 600
B. 400
C. 550
D. 1000
Sol: Assume that P=0.5

Q=0.5

n= randomly selected no.of employers

Margin of error =0.04

The critical value for 95% CI is Z=1.96

$ME=Z*\sqrt{pq/n}$

$0.04=\sqrt{1.96*0.5*0.5/n}$

Squaring on both sides

$0.04^2=1.96*0.5*0.5/n$

$N=1.96*0.5*0.5*/0.0016$

$N=306.26\sim400$

So option B is true

9. Suppose we want the above margin of error to be based on a 98% confidence level. What sample size (minimum) must we now use?

A. 1000
B. 757
C. 848
D. 543

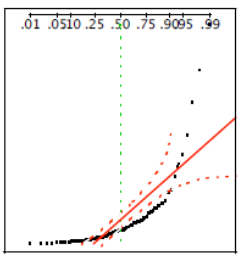
SET:4

CBA: Practice Problem Set 2

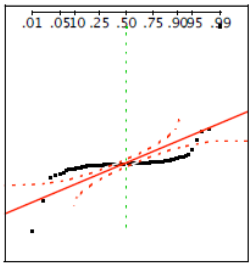
Topics: Sampling Distributions and Central Limit Theorem

1. Examine the following normal Quantile plots carefully. Which of these plots indicates that the data …
   I. Are nearly normal?
      Ans: Plot C
   II. Have a bimodal distribution? (One way to recognize a bimodal shape is a "gap" in the spacing of adjacent data values.)
      Ans: Plot D
   III. Are skewed (i.e. not symmetric) ?
      Ans: Plot A
   IV. Have outliers on both sides of the center?
      Ans: Plot B

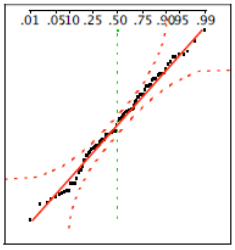
A



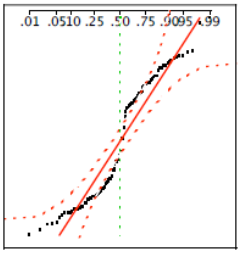
B



C



D



2. 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.

   (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.

FALSE 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).

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

Standard error equal to standard deviation divided by square root of sample size = 5/sqrt(25) =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%

The distribution (sample mean distribution, or distribution of sample means).

In this case the center is at mu = 50 and the standard error is SE = s/sqrt(n) = 40/sqrt(100) = 40/10 = 4 This distribution is normally distributed because of the central limit theorem.

The fact that n = 100 makes n > 30 true indicates that we can use this idea. The value of P(45 < x < 55) is roughly 0.7887

Probability that there will be an investigation = 1- P(45<x<55)=1-0.7887 = 0.2113 which converts to 21.13%

D is correct Answer

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:-
Using the definition of Z value for 5% , we know that Z =-1.96
Here $\bar{x}$ = 45 , mu=50 , σ=40 , Z =-1.96 we want to findout value of n,
- 1.96 =(40-50)/(40/√n)
n = 245.6801 =246

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?

D is True
The SEM is sd/sqrt(n)=120/sqrt(40000)=0.6

A. The standard deviation of the scores within any sample will be 120.
SD will not be 120 of scores in any one sample, especially since we don't know the sample size

B. The standard deviation of the mean of across several samples will be 120.
SD of mean across several samples will also not be 120. It will be less; indeed, probably about 0.6

C. The mean score in any sample will be 720.
The mean score in any sample will be 720. Maybe, but no reason it couldn't be less or more.

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

The average of the mean across several samples will be 720. This is certainly possible, but it requires the mean of all samples that sample size, which would be the case

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

The SEM will be 0.60. This is likely, given the sample size, which even with a lot of skewness will tend towards normality given the sample size. I would use this in calculations. The mean would have an expected value of 720, but in calculations, the SEM is 0.6.