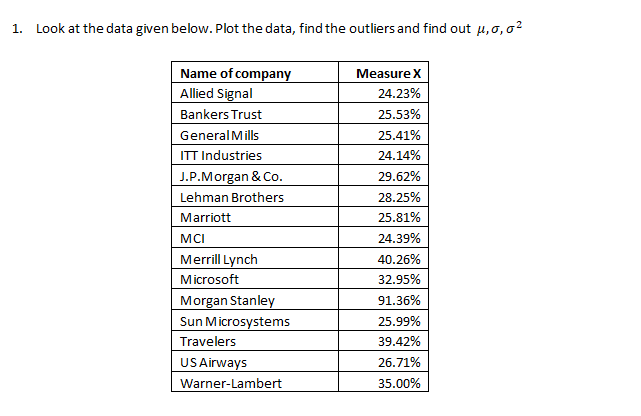
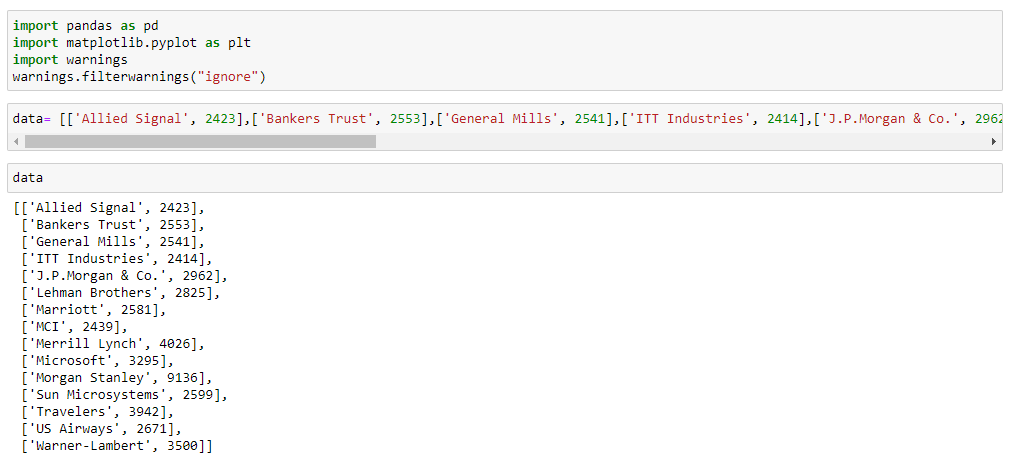
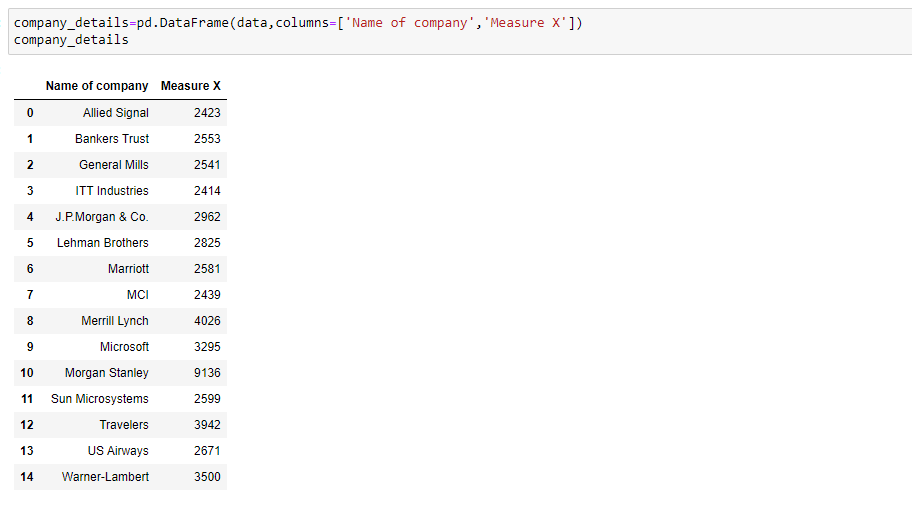
**Basic Statistics -2**

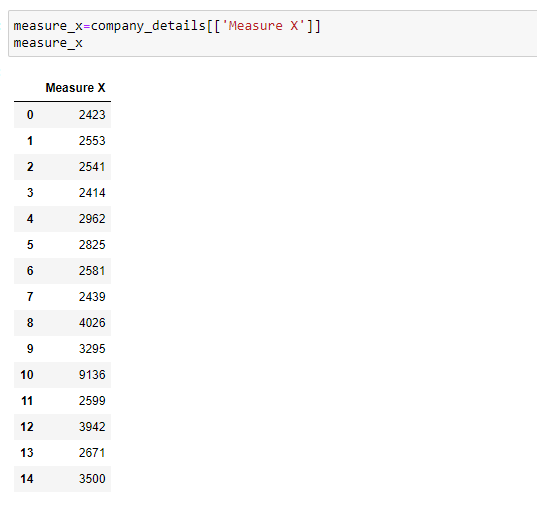
**SET-1**

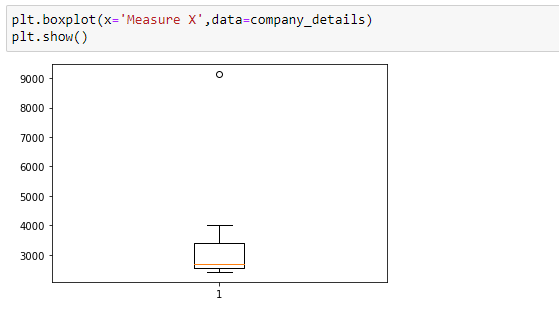
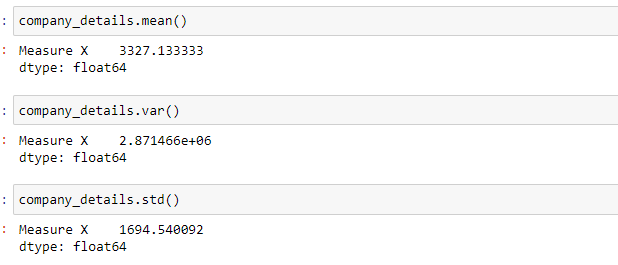
**Topics: Descriptive Statistics and Probability**

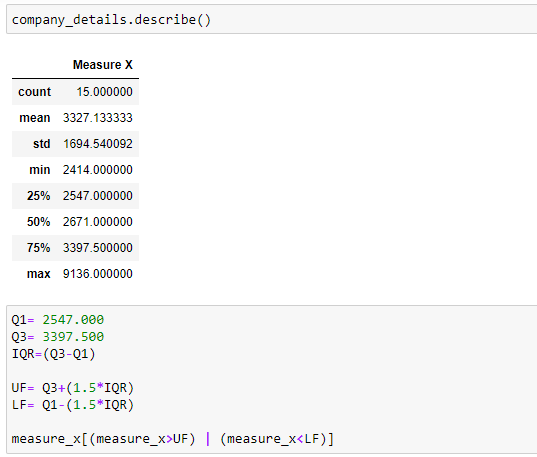


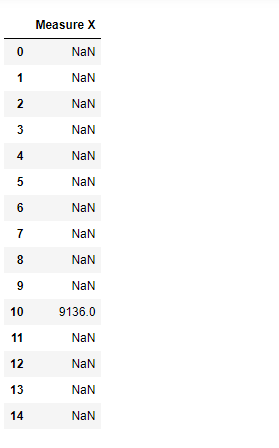


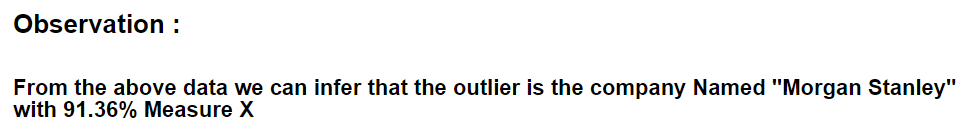


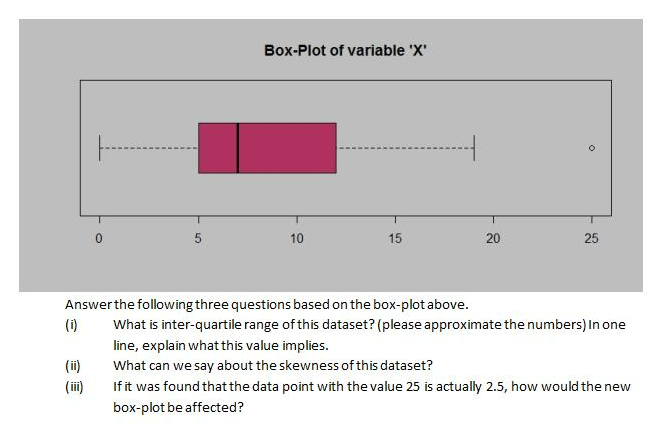










**2)**

1)The interquartile range of the dataset i.e., Q1-Q3 is (5,12)

Range= Max-Min

=12-5=7

>>It implies that most of the data lie between Q2 & Q3

2) The data is positively skewed, because Q1-Q2<Q2-Q3

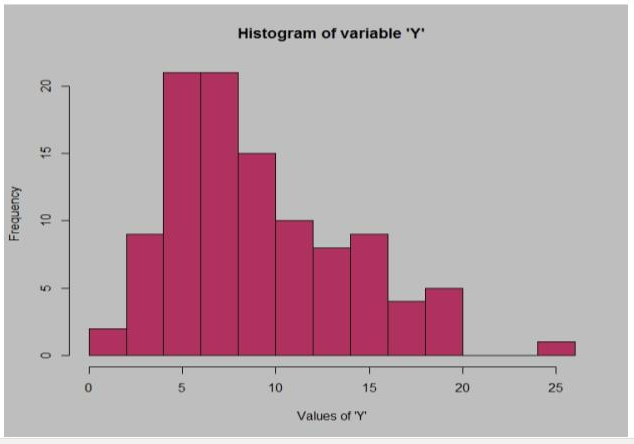
And the whisk on the right side is larger when compared to the Left.

3) If the data point with the value 25 is 2.5 then the boxplot would

be symmetric in nature,

The range Q1-Q2= Q2-Q3

>There would be no skewness, no outlier, it might be called as normally distributed data

**3)** 

**Answer the following three questions based on the histogram above.**

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

**(ii) Comment on the skewness of the dataset.**

**(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.**

1. As the max mass of the data lies left side of the distribution,

Mean > Median > Mode

The mode will be around 6

1. The data is positively skewed, the tail of the distribution is on the right side
2. Histogram and boxplot are very similar in that they both help to visualize. Although histograms are better in determining the underlying distribution of the data, box plots allow you to compare multiple data sets better than histograms as they are less detailed and take up less space.

It is easy to find the outliers in the boxplot than hist.

**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.)**

**>>** 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)= **nCx\*p^x\*q^(n-x)**

n = 5

p= 1/200

q= 199/200

atleast one in five attempted telephone calls reaches the wrong number

= 1- none of the call reaches the wrong number

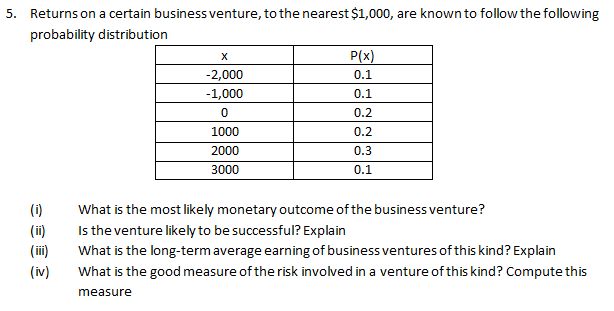
=1- p(0)

= 1   -  ⁵C₀(1/200)⁰(199/200)⁵⁻⁰

= 1  -  (199/200)⁵

= 0.02475

The probability that at least one in five attempted telephone calls reaches the wrong number = 0.02475



1. As the probability (0.3) is more for 2000$ as compared to others,

Therefore, the most likely monetary outcome of the business venture = 2000$

1. Long term average = sum{P(xi)\*Xi} =

(-2000\*0.1) +(-1000\*0.1) +(0) +(1000\*0.2) +(2000\*0.3)

+(3000\*0.1) = 800$

As the long-term average gives positive numbers the business venture likely to be successful.

1. Long term average = sum{P(xi)\*Xi} =

(-2000\*0.1) +(-1000\*0.1) +(0) +(1000\*0.2) +(2000\*0.3)

+(3000\*0.1) = 800$

Means on an average return will be 800$

1. Investors can measure risk in many different ways

(EAR) Earnings at risk one of the best measures of risks will be the best for business venture.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*END of SET-1\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**SET 2**

**Topics: Normal distribution, Functions of Random Variables**

**1. The time required for servicing transmissions is normally distributed with m = 45 minutes and s = 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**

>> WE have a normal distribution with m=45 & x=8

Let X be the amount of time it takes to complete the repair on a customer’s car. To finish in one hour we must have

X so the question is to find p(X>50)

P(X>50)=1- p(X)

P(X) = p(Z(50-45)/8.0)

=p(Z0.625)

=73.4%

Therefore the probability that the service manager will not meet his demand will be = 100-73.4= 26.6%

= 0.2676 (option B)

**2. The current age (in years) of 400 clerical employees at an insurance claims processing center is normally distributed with mean m = 38 and Standard deviation s =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.**

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

>> We have a normal distribution with m= 38, & s=6

Let x be the number of employees, so according to question

1. Probability of employees greater than age of 44= p(X>44)

P(X>44) = 1- p(x 44)

p(x 44)= p(Z (44-38/6))

=p(Z1)

=84.1345%

Probability that the employee will be greater than age of 44 = 100-84.1345

= 15.86%

So, the probability of number of employees between38-34 years of age= p(X<44)-0.5

=84.1345-0.5

=34.1345%

Therefore the statement that “more employees at the processing center are olderthan 44 than between 38 and 44” is TRUE.

1. Probability of employees less than age of 30= p(X<30)

By using normal table

P(X30) = p(Z (30-38)/6)

=p(Z -1.333)

= 9.12%

So the number of employees with probability 0.912 of them being under age 30 = 0.0912\*400

=36.48

Therefore the statement B of the question is also TRUE.

1. **If X1 ~ N(μ, σ2) and X2 ~ N(μ, σ2) are iid normal random variables, then what is the difference between 2 X1 and X1 + X2? Discuss both their distributions and parameters**

**>>** As we know that if X~ N(**μ1, σ1ᶺ2),** and Y~N(**μ2, σ2ᶺ2)** are two independent random variables then X+Y ~N(**μ1+ μ2, σ1ᶺ2+ σ2ᶺ2)** and X-Y~N( **μ1- μ2,** **σ1ᶺ2+ σ2ᶺ2)**

Similarly if Z=aX+bY, where X and Y as defined above, i.e., Z is linear combination of X and Y, then Z- N(a**μ1+b μ2 , aᶺσ1ᶺ2 +bᶺσ2ᶺ2)**

Therefore in the question

2X1-N(2u,4σᶺ2)and

X1+X2 ~ N(μ+ μ, σᶺ2+ σᶺ2) ~ N(2u, **2**σᶺ2)

2X1-(X1+X2)= N(4u, **6**σᶺ2)

**4. Let X ~ N(100, 202). 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**

>>since 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,we have to work out in reverse order .The probability of getting values between a and b should be 0.99 so the probability going wrong ,are the probability outside the a and b area is 0.01 (I.e., 1-0.99)

The probability towards left from a= -0.005

The probability towards rightr from a= +0.005

Since we have the probability of a & b , we need to calculate X, the random variables at a & b which has got these probabilities

By finding the standard normal variable Z( Z value), we can calculate the X values.

Z= (X- μ)/ σ

For probability 0.005 the Z value is -2.57 ( from z table)

Z\* σ+ μ = X

Z(-0.005)\*20+100 = -(-2.57)\*20+100 = 151.4

Z(+0.005)\*20+100 = -(-2.57)\*20+100 = 48.6

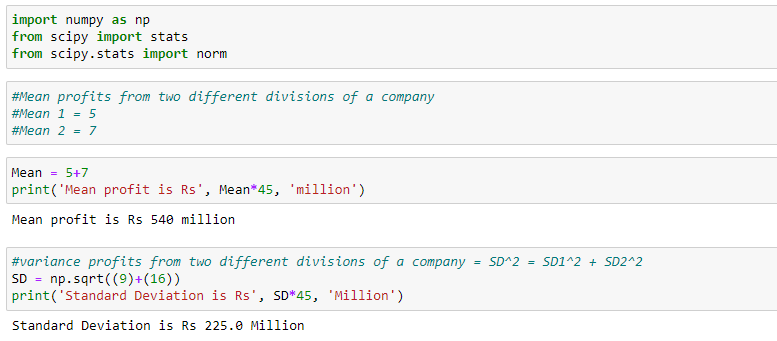
So option D is correct

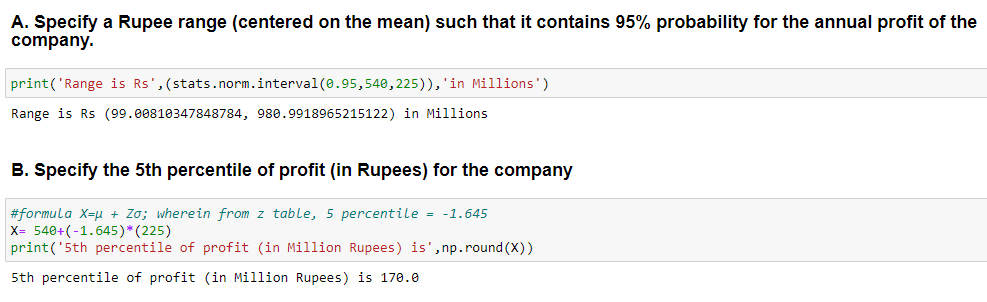
**5. Consider a company that has two different divisions. The annual profits from the two divisions are independent and have distributions Profit1 ~ N(5, 32) and Profit2 ~ N(7, 42) respectively. Both the profits are in $ Million. Answer the following questions about the total profit of the company in Rupees. Assume that $1 = Rs. 45**

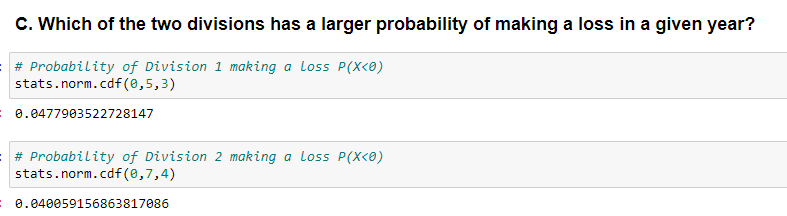
**A. Specify a Rupee range (centered on the mean) such that it contains 95% probability for the annual profit of the company.**

**B. Specify the 5th percentile of profit (in Rupees) for the company**

**C. Which of the two divisions has a larger probability of making a loss in a given year?**







\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*END of SET-2\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**SET-3**

**Topics: Confidence Intervals**

1. **For each of the following statements, indicate whether it is True/False. If false, explain why.**
2. **The sample size of the survey should at least be a fixed percentage of the population size in order to produce representative results.**
3. **The sampling frame is a list of every item that appears in a survey sample, including those that did not respond to questions.**
4. **Larger surveys convey a more accurate impression of the population than smaller surveys.**
5. **TRUE:**

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

1. **FALSE:**

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

**3.TRUE:**

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

=================================================================

1. ***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:**
2. **The population**
3. **The parameter of interest**
4. **The sampling frame**
5. **The sample size**
6. **The sampling design**
7. **Any potential sources of bias or other problems with the survey or sample**
8. **The population:**

**Ans:**  p=x/n =225/9000=0.025

1. **The parameter of interest:**

**Ans:** Sample size, Average, Scale

1. **The sampling frame:**

**Ans:** 9000

1. **The sample size:**

**Ans:** 225

1. **The sampling design:**

**Ans:** This might be purposive sampling, also known as selective and subjective sampling.

In which the researchers rely on their own judgement as here the purpose of which selection of samples from a survey is filtered that is kodak compact digital camera.

1. **Any potential sources of bias or other problems with the survey or sample**

**Ans:** If the sample is not representative of the Population,

then the survey results are potentially biased. Failure to initially specify the Population, problems in selecting a sample, and poor response rate can all lead to sampling error and bias.

1. **For each of the following statements, indicate whether it is True/False. If false, explain why.**
2. **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.**
3. **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.**
4. **The 95% Confidence-Interval for *μ* only applies if the sample data are nearly normally distributed.**
5. TRUE:

100$ is a possible value for the population mean at 95% level of confidence interval 50$ to 110$.

1. TRUE:

If the 95% confidence interval for the number of moviegoers who purchase from the concession stand is 30% to 45%, Then fewer than half of all moviegoers do.

1. FALSE:

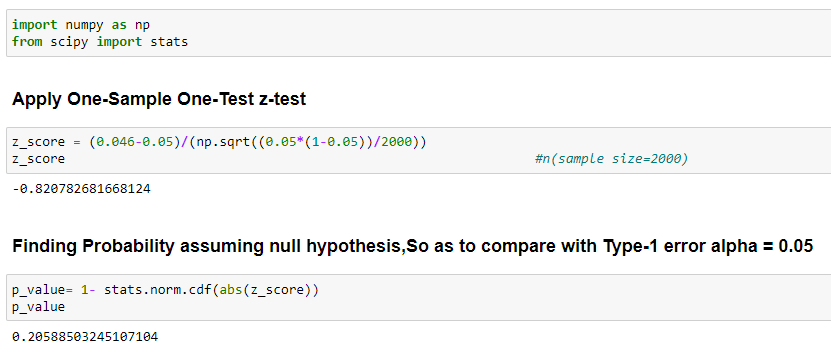
Confidence interval can be used with distributions which are not normal, that are highly skewed or non-normal.

1. **What are the chances that ?**
2. **¼**
3. **½**
4. **¾**
5. **1**

**Ans:** Option B (1/2)

Because sample mean is an unbiased estimator of population mean. By Central Limit Theorem we can see that the distribution of sample means can be normal around its true mean even though the means of population cant built normal distribution. So it depends we have equal chances.

1. **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.**
2. **If the sample were based on 2,000 users, could Microsoft conclude that Mozilla has a less than 5% share of the market?**
3. **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?**



1. **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?**
2. **All shipments are between 205 and 295 books.**
3. **95% of shipments are between 205 and 295 books.**
4. **The procedure that produced this interval generates ranges that hold the population mean for 95% of samples.**
5. **If we get another sample, then we can be 95% sure that the mean of this second sample is between 205 and 295.**
6. **We can be 95% confident that the range 160 to 340 holds the population mean.**

**Ans:**

The 95% confidence interval for the size of the shipment was 250 ± 45 books.

1. 95% of the shipments are between 205 & 295 books.

The option A is not correct, Not all the shipments fall between 205 & 295 only 95% confidence interval of size.

1. **Which is shorter: a 95% *z*-interval or a 95% *t*-interval for *μ* if we know that σ =s?**
2. **The z-interval is shorter**
3. **The t-interval is shorter**
4. **Both are equal**
5. **We cannot say**

**Ans:**

1. The z-interval is shorter than t- interval because of the

z-value & t-value at 95% confidence interval.

**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)?**

1. **600**
2. **400**
3. **550**
4. **1000**

**Ans:**

1. is the correct option

Margins of error estimate is 1/sqrt(n)

If 0.04 = 1/25 is margin of error then

n= 25^2 = 625

So we can choose 600

1. **Suppose we want the above margin of error to be based on a 98% confidence level. What sample size (minimum) must we now use?**
2. **1000**
3. **757**
4. **848**

**D.543**

**Ans:**

­­­ Option (D) would be the correct answer

Given that we want above margin of error with 98% confidence level.

Margin of error =1/sqrt(n)

0.04 =1/sqrt(543)

=1/23.30

=0.0429

Therefore sample size(n) = 543 is the correct one.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*END OF SET-3\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**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 …**
2. **Are nearly normal?**
3. **Have a bimodal distribution? (One way to recognize a bimodal shape is a “gap” in the spacing of adjacent data values.)**
4. **Are skewed (i.e. not symmetric) ?**
5. **Have outliers on both sides of the center?**



**Ans:** 1. Graph (C ) is nearly normal

Because all the data points plotted on the graph perfectly lies on a straight line

1. Graph (B) and Graph (D) are having bimodal distribution, because of the space on both sides of the straight line.
2. Graph (A) is the only skewed distribution, because B& D as they are bi-modal they can be symmetrical too and C is normal.
3. Graph (A),(C),(D) are having outliers present in the data
4. **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.**
2. **The standard error of the daily average SE() = 1**

**Ans:**

1.FALSE:

Before using a normal model for the sampling distribution of the average package weights, the manager must need not be confirm that weights of individual packages are normally distributed

According to Central Limit Theorem:

The distribution of sample mean will be approximately normal even if the data in the population is not normal, if the sample size is very large.

2.TRUE:

Standard error equal to standard deviation divided by square root of sample size

Sigma = 5 (given)

Sample size = 25 (given)

= 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?**

1. **1.25%**
2. **2.5%**
3. **­­10.55%**
4. **21.1%**
5. **50%**

**Ans:**

Option (D) 21.1% is the correct option

It is a sample mean distribution

In this case the center is at mu = 50

And Standard error is SE = s/sqrt(n)

=40/sqrt(100)

=40/10

= 4

This distribution is normally distributed because of central limit theorem. The fact that n> 30 is true

The value of p(45<x<55)

Is 0.7887 (by using calculator)

Subtracting 1 from that gives 1-0.7887

= 0.2113

Which converts to 21.13%

That rounds to 21.1%

**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.**

1. **144**
2. **150**
3. **196**
4. **250**
5. **Not enough information**

**Ans:**

We have

From the above example

Population Mean (μ) = 50

Sample Mean (x̅) = 55

Sample Standard Deviation =40

For thresholds of 55 & 45

= x̅1- μ/(40/sqrt(n)) - x̅2 - μ/(40/sqrt(n))

= 55-50/(40/sqrt(n)) – 45-50/(40/sqrt(n))

n =250

Option (D)

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.**
3. **The standard deviation of the mean of across several samples will be 120.**
4. **The mean score in any sample will be 720.**
5. **The average of the mean across several samples will be 720.**
6. **The standard deviation of the mean across several samples will be 0.60**

**Ans:**

1. Data is distributed with leptokurtic kurtosis mean more information at the center and lesser information at the tail. This mean there is higher chance that average of mean of a aspirant that randomly chosen will be 720 that fall in between 650 and 790 at the center.

\*\*\*\*\*\*\*\*\*\*\*\*\*END OF SET-4\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*