ASSIGNMENT 2
Set+1_Descriptive+statistics+Probability+(2)

VIGNESH R BABU

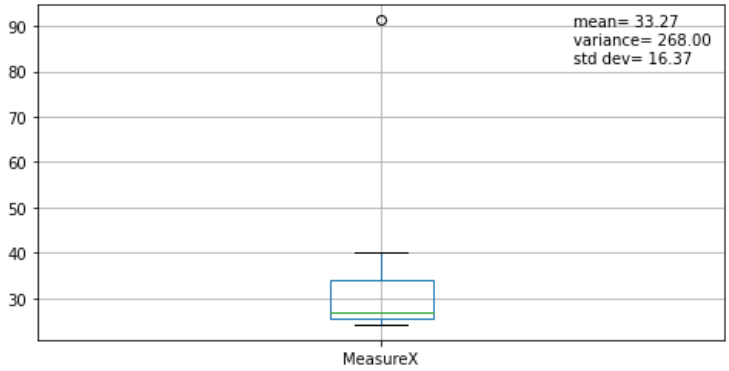
Topics: Descriptive Statistics and Probability

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%

Ans:

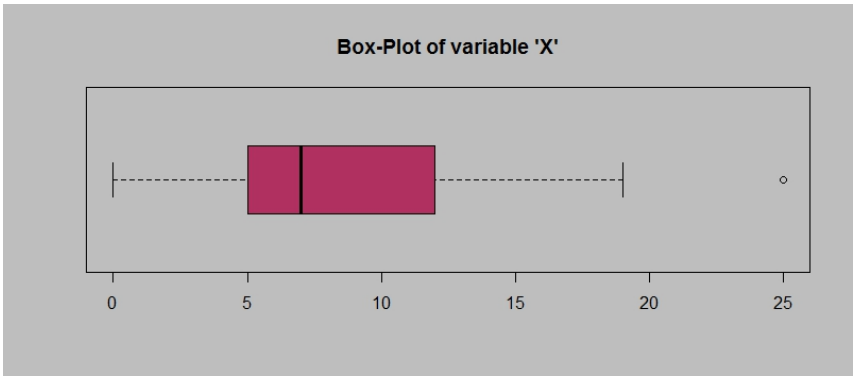
Mean= 32.27%,
Variance = 268.00,
Std dev = 16.37%
Outlier = Morgan Stanley with 91.36% is the outlier



Code:

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

#Converting values to a csv file for importing
x = pd.read_csv('C:\\Users\\Vignesh R
Babu\\Documents\\ExcelRPython\\Assignment Codes\\2\\2Q1.csv')

#Renaming columns
x.columns = ['NameOfCompany','MeasureX']

# Removing % symbol from MeasureX
x['MeasureX'] = x['MeasureX'].map(lambda x: x.rstrip('%'))

#MeasureX is a object datatype that needs to converted to float
x.MeasureX = pd.to_numeric(x['MeasureX'])

#Boxplot with Mean Variance & Std Dev
x.boxplot(['MeasureX'])
plt.figtext(0.73, 0.75, '''mean= {0:.2f}
variance= {1:.2f}
std dev=
{2:.2f}'''.format(np.mean(x.MeasureX),np.var(x.MeasureX),np.std(x.MeasureX)))
plt.show
```

ASSIGNMENT 2



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.

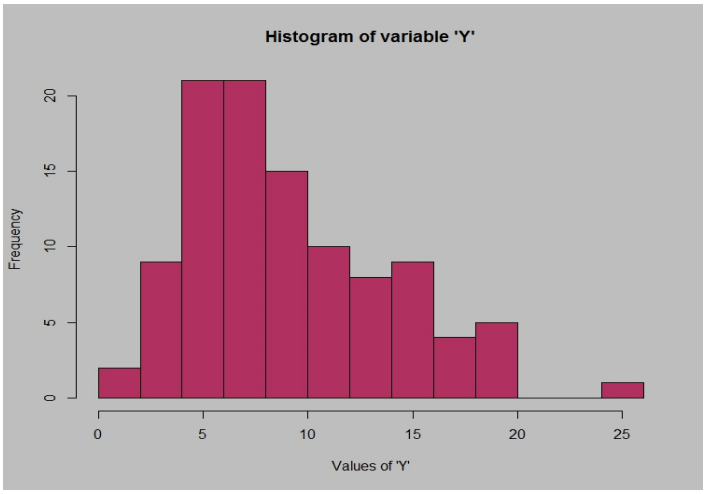
Ans: The IGR of the given data set is from 5 to 12 approximately. About 50% of all the data lies in this range and it doesn't not consider outliers or extreme values. Therefore, it could be considered to be more accurate than the range of the data.

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

Ans: The median in this case would be lesser than the mean. Therefore, this is a right tailed plot with extreme values and outliers to the right side of the graph. The skewness of the dataset must be positive skewness.

(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 outlier does not affect the box plot in any way. As per the question, if it is a 2.5, it comes in the lower fence of the boxplot. There could be a shift in the mean depending on the value. However the shift may not be large as 2.5 does not lie in the IQR range.



Questions referred to from *Aczel A., Sounderpandian J., Complete Business Statistics (7ed.)*

ASSIGNMENT 2

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

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

Ans: The mode would lie approx. between 4 and 8 where the frequency is greater than 20.

(ii) Comment on the skewness of the dataset.

Ans: It is a positive skewed right tailed distribution. There more positive values towards the right of the distribution.

(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.

Ans: Both histogram and box-plot enables us to understand the central tendency, the amount of variation in the data as well as the presence of gaps, outliers or unusual data points. Histogram provides a better visualization to grasp the probability distribution of a data. Box plots can be more useful when we have to compare several datasets.

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.)

Ans: Probability of call getting misdirected P(E) = 1/200
Prob of call not getting misdirected = 1 – P(E) = 199/200
No of attempts = 5, Prob that atleast one in 5 attempts misdirected calls P(X) = 1 – (prob that no calls misdirected in 5 attempts)
P(X) = 1 – ((199/200)*(199/200) (199/200)*(199/200) (199/200))
= 1 –((199/200)**5)
= 1 - 0.975248753121875
P(X) = 0.02475124687812502 or approx. => 0.025 **(Answer)**

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

ASSIGNMENT 2

1000	**0.2**
2000	**0.3**
3000	**0.1**

(i) What is the most likely monetary outcome of the business venture?

Ans: P(x) is highest for x = 2000. There it is the most likely monetary outcome of the business venture is x = 2000.

(ii) Is the venture likely to be successful? Explain

Ans: The venture is likely to be successful if P(x >0) is greater than P(x<=0).

P(x>0) = P(x = 1000) + P(x = 2000) + P(x = 3000)
= 0.2 + 0.3 + 0.1
= 0.6

P(x<=0) = P(x = 0) + P(x = -1000) + P(x = -2000)
= 0.2 + 0.1 + 0.1
= 0.4

P(x>0) > P(x<=0) . Therefore this venture is likely to be successful.

(iii) What is the long-term average earning of business ventures of this kind? Explain

Ans: It is required to calculate the expected value of the business venture.

E(x) = (3000*0.1)+(2000*0.3)+(1000*0.2)+ (0*0.2)+ (-1000*0.1)+ (-2000*0.1)
= 300 + 600 + 200 + -100 + -200

E(x) = 800

(iv) What is the good measure of the risk involved in a venture of this kind? Compute this measure

Ans: Standard Deviation is a good measure the calculate the risk involved in this case.

Std Dev = 1469.6938456699068

Code: This is same as manually calculating std dev.

```
#taking both values in seperate lists
x =  [-2000, -1000, 0, 1000, 2000, 3000]
p = [0.1, 0.1, 0.2, 0.2, 0.3, 0.1]
#making E[x^2] and E[x]^2 lists
e_in_sqr = []
e_out_sqr = []
#multiplying and taking the expected values
for i in range(len(x)):
    e_out_sqr.append(x[i]*p[i])
    e_in_sqr.append((x[i]**2)*p[i])
#Taking variance from x, var(x) = E[x^2] - E[x]^2
variance_x = sum(e_in_sqr) - (sum(e_out_sqr)**2)
#finding std dev of x, std(x) = Sqrt(var(x))
print((variance_x)**0.5)
```

Questions referred to from *Aczel A., Sounderpandian J., Complete Business Statistics (7ed.)*

ASSIGNMENT 2

Set+2_Normal+Distribution+Functions+of+random+variables+(1)

VIGNESH R BABU

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

Ans: Let the prob of not meeting commitment be P(E).
We have to calculate the z-score first for the given scenario
Given : μ = 45 , σ = 8 , time = 60 – 10 = 50 Minutes
Z-Score at 50 => (time – mean time)/std dev => (50-45)/8 = 0.625
Corresponding probability from Z-table = 0.7324
P(E) = 1 – 0.7324 = 0.2676 **(Answer = Option B)**
Code:

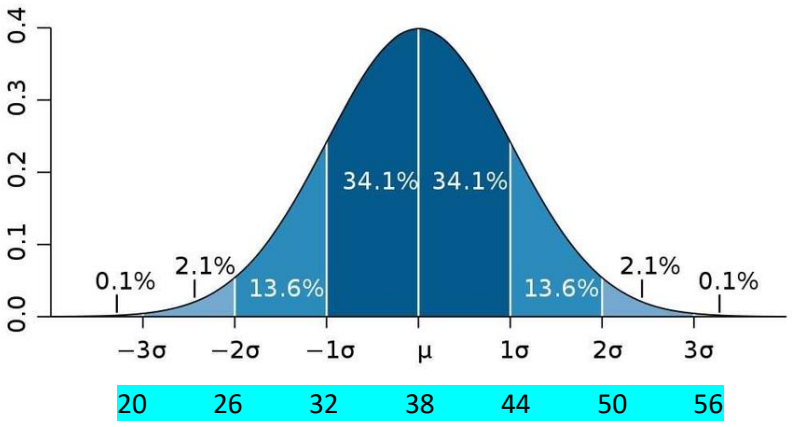
```
import scipy.stats as ss
xcritical = 50
mean = 45
stdev = 8
#to get z-score and probability in python
p = ss.norm.cdf(x=xcritical,loc=mean,scale=stdev)
z_score = ss.norm.ppf(p)
print(z_score)
print(p)
```

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.

Ans: False

Explanation: Consider the following Normal Distribution Graph. The range between ages 38 and 44 is within one standard deviation from the mean. This means that it contains about 34.1% of 400 approx. = 136 people. Going beyond age 44 will result in about



16% approx. = 64 people which is less than the former. Therefore, the answer is False.

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

Ans: True

Explanation: Finding the corresponding probability at age 30 gives approx. 36 people.

Code:

```
import scipy.stats as ss
xcritical = 30
mean = 38
stdev = 6
#to get z-score and probability in python
p = ss.norm.cdf(x=xcritical,loc=mean,scale=stdev)
z_score = ss.norm.ppf(p)
print(z_score)
print(p)
print(p*400)


Output:
-1.3333333333333333
0.09121121972586788
36.484487890347154 <= approx. 36
```

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.

Ans: iid stands for independent, identically distributed random variables. A good example is a succession of throws of a fair coin. As per the question, consider X1 and X2 be the outcomes of two die rolls. They iid normal random variables. Then X1+X2 is the sum of the numbers on the two dice and 2X1 is twice the number on the first die. These don't have the same distribution - for example, X1+X2 can be odd, and 2X1 is always even.

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

Ans: In case of 0.99 symmetric prob, to get symmetry about mean => (1-0.99)/2 = 0.005
z-score corresponding to the value is -2.57.
To find the a,b values => 20x(-2.57) +/- 100 would give approx. (48.6, 151.4)
Alternative Code:

```
import scipy.stats as ss
mean, std, p = 100, 20, 0.99
print([round(x,2) for x in ss.norm.interval(alpha=p, loc=mean, scale=std)])
Output: [48.48, 151.52]
```

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.

Ans: According sum of normal random variables rules, we can add up the profits.
Annual_profit ~ $N(5+7, 3^2 + 4^2)$ => $N(12, 5^2)$
Rupee Range = [99008103.48, 980991896.52]
Rupee Range ~ 99MillionRupees to 980MillionRupees (Answer)

Code:

```
import scipy.stats as ss
mean, std, p = 12, 5, 0.95
mean = mean*(10**6)*45
std = std*(10**6)*45
print([round(x,2) for x in ss.norm.interval(alpha=p, loc=mean, scale=std)])

Output: [99008103.48, 980991896.52]
```

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

Ans: We already have the upper and lower range of the Annual_profit.
We can calculate the 5th percentile using python.
5th percentile of profit = 143.1 Million Rupees (Answer)

Code:

```
import scipy.stats as ss
mean, std, p = 12, 5, 0.05
mean = mean*(10**6)*45
std = std*(10**6)*45
#to get z-score and rupee_value
y = ss.scoreatpercentile([99008103.48, 980991896.52], 5)
print(y)

Output: 143107293.132 ~ approx. 143.1Million
```

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

Ans: Division 1 will have larger probability for making a loss. (Answer)

Code:

```
import scipy.stats as ss
#prob of division1 to make profit less than 0
div1 = ss.norm.cdf(0,5,3)
print("P(div1 <0) = {:2f}".format(div1))
#prob of division2 to make profit less than 0
div2 = ss.norm.cdf(0,7,4)
print("P(div2 <0) = {:2f}".format(div2))
if div1 > div2:
    print("Division1 has larger prob for loss")
else: print("Division2 has larger prob for loss")


Output:
P(div1 <0) = 0.047790
P(div2 <0) = 0.040059
Division1 has larger prob for loss
```

ASSIGNMENT2

Set +3 Topics: Confidence Intervals

VIGNESH R BABU

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: 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.

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

Ans: All the readers of the PC magazine

B. The parameter of interest

Ans: The population mean that rated the digital camera

C. The sampling frame

Ans: 9000

D. The sample size

Ans: 225

E. The sampling design

Ans: Sampling Design = n/N

Where n – number of units to be samples

N – number of units in total population

Sampling Design = 225/9000 = 0.025 (Answer)

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

Ans: Selection of the readers, Selection of the issue which will contain the survey

ASSIGNMENT2
Set +3 Topics: Confidence Intervals

Vr_babu

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.

Ans: True

Reason - The 95% confidence interval for the average purchase of customers at a department store is $50 to $110. Which means that there is a 95% chance that the population mean will fall between $50 and $110. Hence, as $100 falls between $50 and $110, it is a plausible value for the population mean at this confidence level.

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.

Ans: True

Reason - The 95% confidence interval for the number of moviegoers who purchase concessions is 30% to 45%, this means that there is a 95% chance that only 30% to 45 % of moviegoers purchase concessions, which is less than 50%. Hence, we can infer that fewer than half of all the moviegoers purchase concessions.

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

Ans: False

Reason – Most data we have are not normally distributed. But we can use central limit theorem to make a normal approximation and obtain an asymptotical confidence interval. Reference

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

A. ¼
B. ½
C. ¾
D. 1

Ans: D. 1

Reason: Due to the Central Limit Theory, the distribution of sample means will be normally distributed around the true mean.

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?

Ans: No. It does not clearly mention the type of users the sample was based on.

ASSIGNMENT2
Set +3 Topics: Confidence Intervals

Vr_babu

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?

Ans: Yes

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.

Ans: False

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

Ans: True

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

Ans: True

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

Ans: True

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

Ans: False

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

Ans: D. We cannot say.

It depends on estimate of standard deviation. A 95% t-interval for might be longer and a 95% z-interval for shorter due to z-critical value less than t-critical value and a 95% t-interval for might be shorter interval. The outcome depends on standard deviation instead of the sample because the standard deviation value is effect the length of an 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.

ASSIGNMENT2
Set +3 Topics: Confidence Intervals

Vr_babu

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

Ans: A. 600
We are required to find the value of n in order to create an estimate where we are 95% confident with a margin or error 4%.
Margin of error is given by

$$ME = Z^* \sqrt{\frac{\hat{p}\hat{q}}{n}}$$ ----------(1)

Let n represent the size of the sample.

Let p^ be the sample proportion and q^ = 1-p^
Since value of sample proportion p^ has not been given and then we can take p^ =0.5 (which implies q^ = 1 -0.5 = 0.5 and p^ x (1−p^) = 0.25), because this will result in the largest possible sample size and this will largest possible sample size will be appropriate for all sample proportions.

z-score corresponding to 95% = 1.96

Margin of Error, M.E = 0.04

Calculating n using eq(1)

n = 0.25/((M.E/z)^2) = 600.25 ~ approx. 600 **(Ans**0.25/((M.E/z)^2) **wer)**

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

Ans: C.848
Using above method,
z-score corresponding to 98% = 2.33 , M.E = 0.04
n = 0.25/((M.E/z)^2) = 848.265 ~ approx. 848 **(Answer)**

ASSIGNMENT 2

Set +3 VIGNESH R BABU

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

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

Ans: D

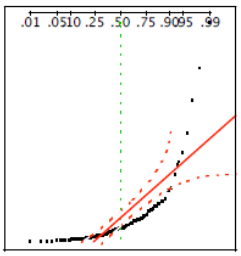
III. Are skewed (i.e. not symmetric) ?

Ans: A

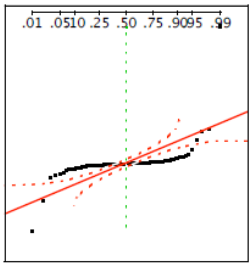
IV. Have outliers on both sides of the center?

Ans: A

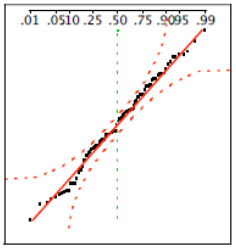
A



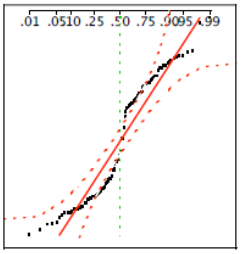
B



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.

Ans: True.

Reason: Sample size n<30. Therefore, it may or maynot follow Central Limit Theorem. Based on the central limit theorem, the sampling distribution of the sample mean approach

normal distribution as the sample size become bigger (over 30). Thereforem the manager must confirm if the weights are normally distributed.

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

Ans: True

Reason: 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%

Ans: D 21.1% **(Answer)**

It is given that n = 100, Mean = 50, std dev (pop) = 40, std dev(sample) = 40/sqrt(100)

We need to calculate z scores and find probability of x in range(45,55)

P(45<x<55) => Code:

```
import scipy.stats as ss
xvalue1, xvalue2 = 45, 55
mean = 50
stdev = 40/(10)
#find probability of x between 45 and 55
p = ss.norm.cdf(xvalue2,mean,stdev)-ss.norm.cdf(xvalue1,mean,stdev)
# 1 - p is our desired result
print("Percentage for investigation is {:.2f}%".format((1-p)*100))

Output:
Percentage for investigation is 21.13%
```

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. 250
We are required to find the n value. Mean and Population std dev remains the same.
P(E) = 0.05 => 0.05/2 => 1-0.025 = 0.975 [since it is two sided distribution]
Corresponding z-score = 1.644
(value – Mean)/std_sample = 1.644
Std_sample = (55-50)/1.64 = 3.041
Std_sample = Std_population/root(n)
n = (std_population/std_sample)^2
n = 15.68^2 = 246.05~ approx. 250 **(Answer)**

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. The standard deviation of the mean across several samples will be 0.60
Given n = 40000
Mean = 720
Std_pop = 120
Std_sample = Std_pop/root(40000)
Std_sample = 0.60 Therefore, Option E is the right answer. **(Answer)**

***