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%

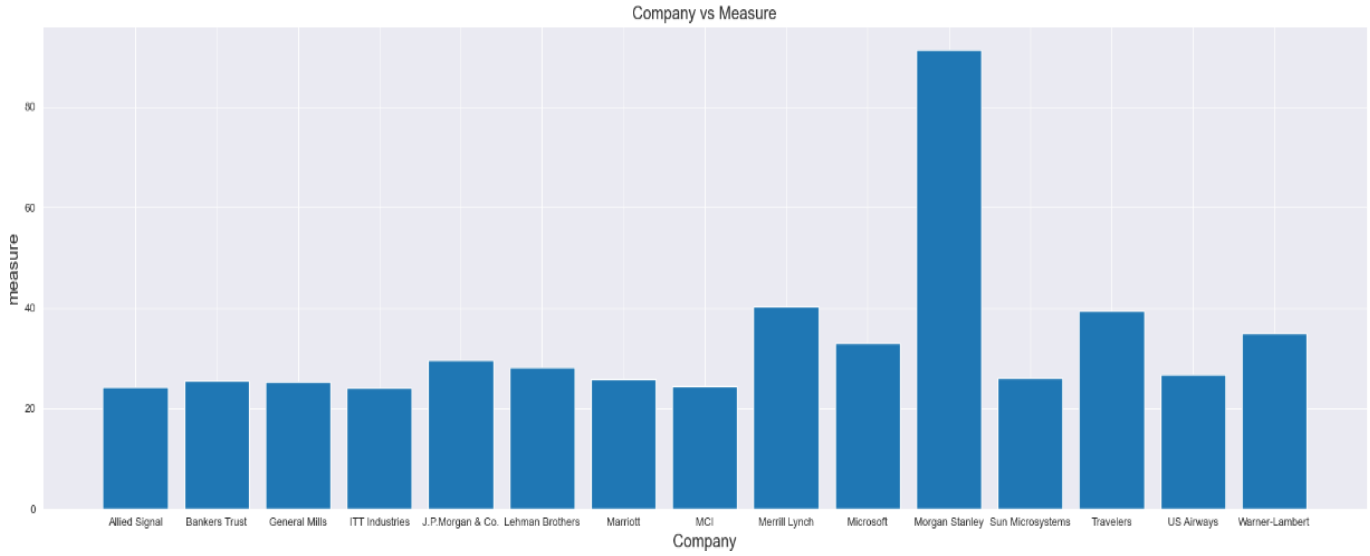
Ans:

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

x = [24.23,25.53,25.41,24.14,29.62,28.25,25.81,24.39,40.26,32.95,91.36,25.99,39.42,26.71,35.00]
names=['Allied Signal','Bankers Trust','General Mills','ITT Industries','J.P.Morgan & Co.','Lehman Brothers',
      'Marriott','MCI','Merrill Lynch','Microsoft','Morgan Stanley','Sun Microsystems','Travelers','US Airways',
      'Warner-Lambert']
df = pd.DataFrame({'company' : ['Allied Signal','Bankers Trust','General Mills','ITT Industries','J.P.Morgan & Co.','Lehman
                  'Marriott','MCI','Merrill Lynch','Microsoft','Morgan Stanley','Sun Microsystems','Travelers','US Airways',
                  'Warner-Lambert'],
                  'measure' :  [24.23,25.53,25.41,24.14,29.62,28.25,25.81,24.39,40.26,32.95,91.36,25.99,39.42,26.71,35.00]})
df
plt.figure(figsize = (24,7))
plt.xlabel('Company', fontsize = 15)
plt.ylabel('measure', fontsize = 15)
plt.title('Company vs Measure', fontsize = 15)
plt.bar(df['company'], df['measure'])

print(np.mean(x))
print(np.var(x))
print(np.std(x))
```

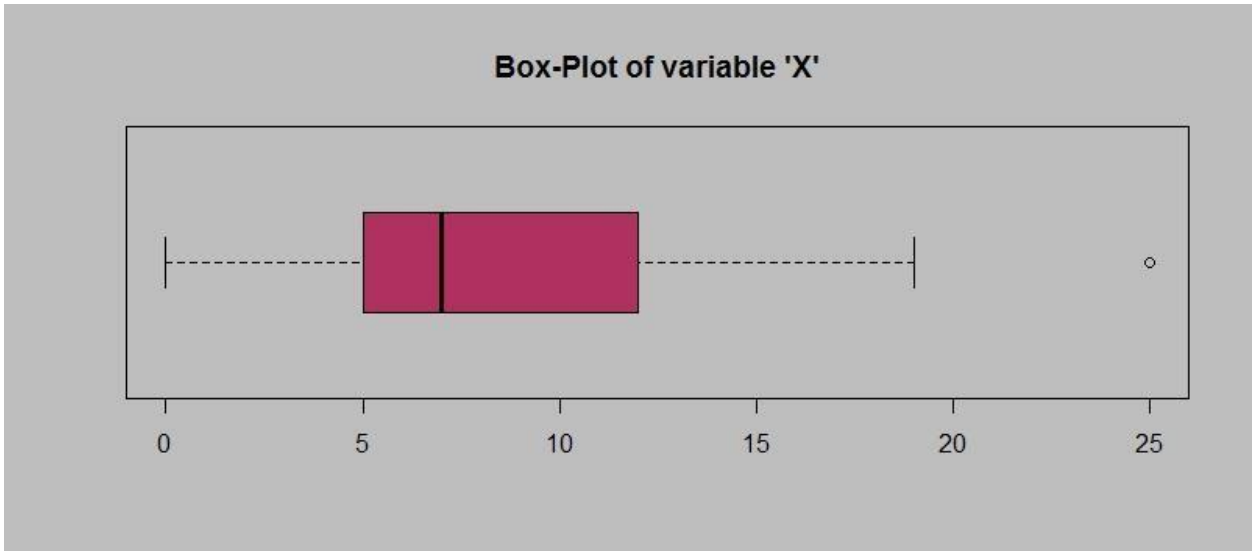
The outlier in the below histogram is Morgan Stanley 91.36%

Mean = 33.2713
Standard deviation = 16.3708
Variance = 268

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



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: Inter-quartile range = 12 – 5 = 7

So, the middle 50% of the data lies between 5 and 12

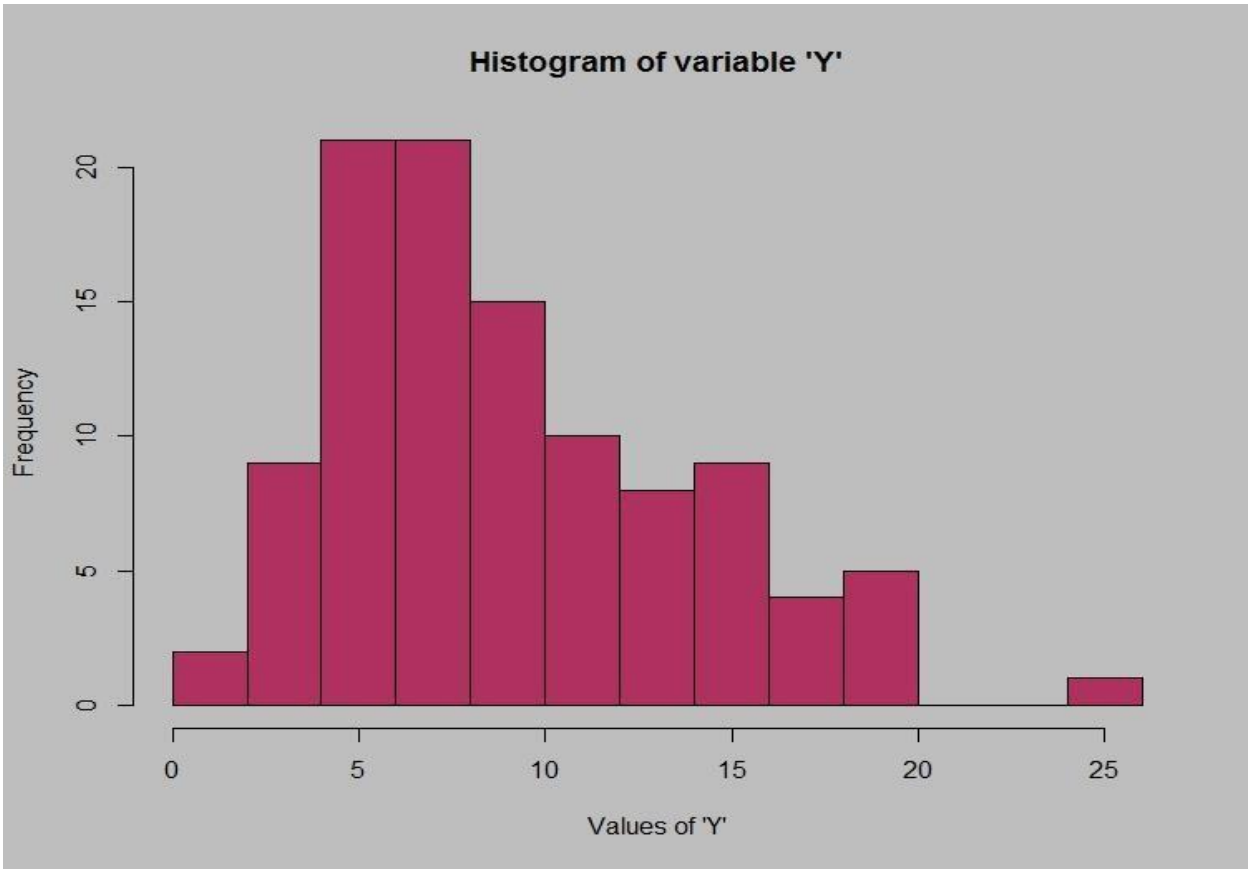
(ii) What can we say about the skewness of this dataset?

Ans: It is right skewed. The data in the given dataset is more concentrated towards The left region and the right tail is longer.

(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 interquartile range will change and the median value remain same and data will be normally distributed.

3.



Answer the following three questions based on the histogram above.

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

Ans: The mode of the dataset lies between 4 to 8

(ii) Comment on the skewness of the dataset.

Ans: It is right skewed. The data is more concentrated towards the left region and the right tail is longer

(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 plots are right skewed and the median can be easily calculated in both Graphs.

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 being misdirected (x) = (1/200) = 0.005
   Probability of call being not misdirected (1-x) = (1 – 0.005) = 0.995

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

   (5C1)*(0.005)*(0.995)^4 = 0.0245

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?

   Ans: The most likely monetary outcome of the business venture is 2000 as the probability is high for this i.e., 0.3

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

   Ans: It can be defined by return on investment = 0.2 + 0.3 + 0.1 = 0.6
   So, the venture likely to be successful as 60% probability can be achieved.

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

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

Ans: The long term average is the expected value

$E(x) = \sum x.P(x) = -200 -100 + 0 + 200 + 600 + 300 = 800$. So, the returns on an average will be more than 800

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

Ans: Variance can be defined in this case, as variance is high the chances of risk is Also high.

Var(x) = E(x^2) – (E(X))^2

= 2800000 – 640000

= 2160000

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: **The average time as the work begin after 10 minutes will be 45+10 = 55**
Z = (60-55)/8 = 0.625
P(x>60) = 1-p(x<=60)

Code : 1 – stats.norm.cdf(60,55,8)
0.26598

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: **P(X > 44) = 1 - P(X ≤ 44) = 15.865%**
      Code: 1-stats.norm.cdf(44,38,6)

      P(38<x<44) = 34.134%
      Code: stats.norm.cdf(44,38,6) – stats.norm.cdf(38,38,6)
      Therefore, more employees at the processing center are older than 44 than between 38 and 44 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: **P(x<30) = 9.121%**
      Stats.norm.cdf(30,38,6)
      Therefore, the number of employees under the age of 30 having probability of 9.121% = 0.09121*400 = 36.48.
      So, it is TRUE that the center would be expected to extract about 36 employees.

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:

   $\boldsymbol{X_1 + X_2 \sim N(2\mu, 2\sigma^2)}$

   $\boldsymbol{2\,X_1 = N(2\mu, 4\sigma^2)}$

   $\boldsymbol{2X1 - (X1+X2) \sim (2\,\mu - 2\,\mu, 4\,\sigma 2 + 2\sigma 2\,)}$

   Distribution will remain same and the parameters deviate slightly for every sample data.

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: D

   The 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 are:

   [48.5, 151.5]

   Code: stats.norm.interval(0.99, 100,20)

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

Ans: Python code :

```
import numpy as np
import pandas as pd

# A.
mean_profit = (5+7)*45 # from two profits
print(mean_profit)

std_profit = (np.sqrt(9+16))*45
print(std_profit)

stats.norm.interval(0.95, mean_profit, std_profit)

# B.
# X=μ + Zσ
X= mean_profit + (-1.645)*(std_profit)
print('5th percentile of profit (inRupees) is',np.round(X,2))

# C.
# Probability of Division 1 making a loss P(X<0)
print(stats.norm.cdf(0,5*45,3*45))

# Probability of Division 2 making a loss P(X<0)
print(stats.norm.cdf(0,7*45,4*45))
```

A. Specify a Rupee range (centered on the mean) such that it contains 95% probability for the annual profit of the company.
Ans: for 95% probability for the annual profit of the company, rupee ranges in between [9.9 to 98.1] crores

B. Specify the 5th percentile of profit (in Rupees) for the company
Ans: The 5th percentile of profit for the company is 169.88 crore rupees

C. Which of the two divisions has a larger probability of making a loss in a given year?
Ans: Probability of 1st division making loss is 4.779%
Probability of 2nd division making loss is 4%
Division 2 has a larger probability of making loss in a given year

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 size of the sample should have at least 30 observations in order to produce representative results.

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

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

   Ans: True. As larger surveys contain larger sample size which results in more accurate values compared to smaller surveys.

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 readers of the Magazine i.e., 9000

   B. The parameter of interest

   Ans: sample size, average and scale

C. The sampling frame

Ans: 9000

D. The sample size

Ans: 225

E. The sampling design

Ans: Probability sampling or random sampling

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

Ans: Reach to the readers.

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. At 95% confidence levels, the population mean falls between $50 to $110, it means there is 95% availability for the data to fall between this range. So, $100 will be a plausible value for the population mean at this level of confidence.

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: False. At 95% confidence limit, the plausible value for the population mean is $100. So, it is not possible.

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

Ans: False. From the central limit theorem, it defines that distribution of the sample is normal regardless of the data itself.

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

   A. ¼
   B. ½
   C. ¾
   D. 1

   Ans: B. By the assumption, there is a 50%chance that the sample mean is greater than population 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: Yes, it can be concluded that Mozilla has a less than 5% share.
   code:

```
z_scores=(0.046-0.05)/(np.sqrt((0.05*(1-0.05))/2000))
p_value=1-stats.norm.cdf(abs(z_scores))
```

   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: False

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.

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

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

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

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

Ans : C

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: A.

As t-critical is greater than z-critical, 95% confidence interval for mean is shorter for z-interval. So, z-interval is shorter

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

Ans : A

Sample size = 600.23

code :

```
from scipy import stats
z = stats.norm.ppf(0.99)
moe = 0.04
p = 0.5
sample_size = (z/moe)**2 * p*(1-p)
sample_size
```

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

Sample size = 845.61

Code:

```
from scipy import stats
z = stats.norm.ppf(0.99)
moe = 0.04
p = 0.5
sample_size = (z/moe)**2 * p*(1-p)
sample_size
```

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 is nearly normal.

   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 representing bimodal distribution.

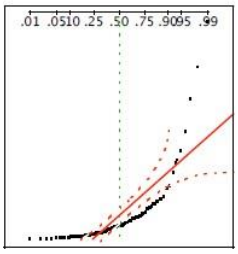
   III. Are skewed (i.e. not symmetric) ?

   Ans: Plot A is skewed.

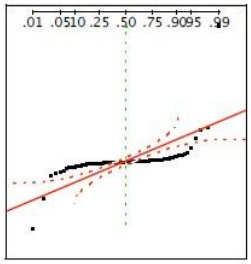
   IV. Have outliers on both sides of the center?

   Ans: Plot B having outliers on both sides of the center.

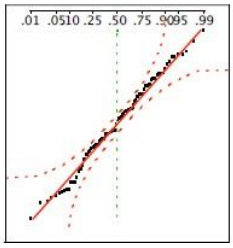
A



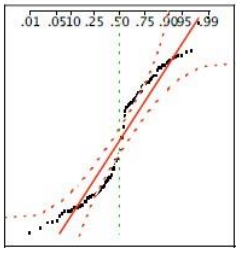
B



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: False. For sampling distribution, checking the weights of the individual packages is not needed.

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

   Ans : True

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. 21.13%

Code:

```
from scipy import stats
import math

# For no investigation = P(45<x<55)
# For investigation = 1 - P(45<x<55)

y = stats.norm.cdf(55,50,40/pow(100,0.5)) - stats.norm.cdf(45,50,40/pow(100,0.5))

# For investigation
print(f'Probability that investigation happens is {(1-y)*100} %')
```

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.

To maintain the probability at 5% i.e., stats.norm.ppf(0.975) will give z-score of -1.96 and 1.96.

Therefore, lower limit -> (45-50)/40/(sqrt(n)) = -1.96

upper limit -> (55-50)/40/(sqrt(n)) = 1.96

n = 245.86

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: A, D

As the data having large sample size i.e., 40000, so the sample means follows normal distribution and for this condition the sample mean equal to population mean, so, the average scores of the random samples will likely close to 720 and similarly the standard deviation of the scores will likely close to 120.