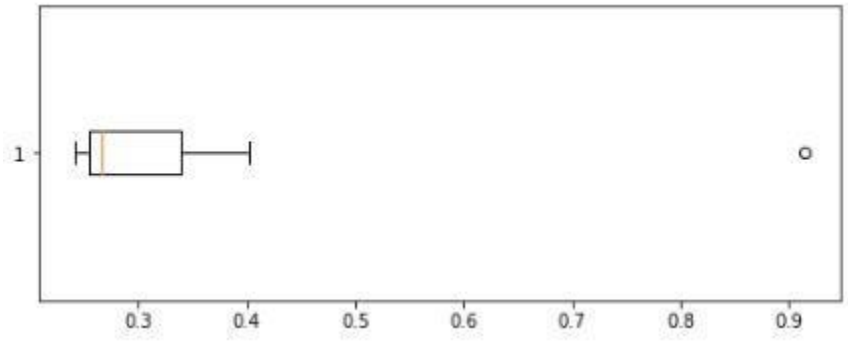
Topics: Descriptive Statistics and Probability

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%

Solution:

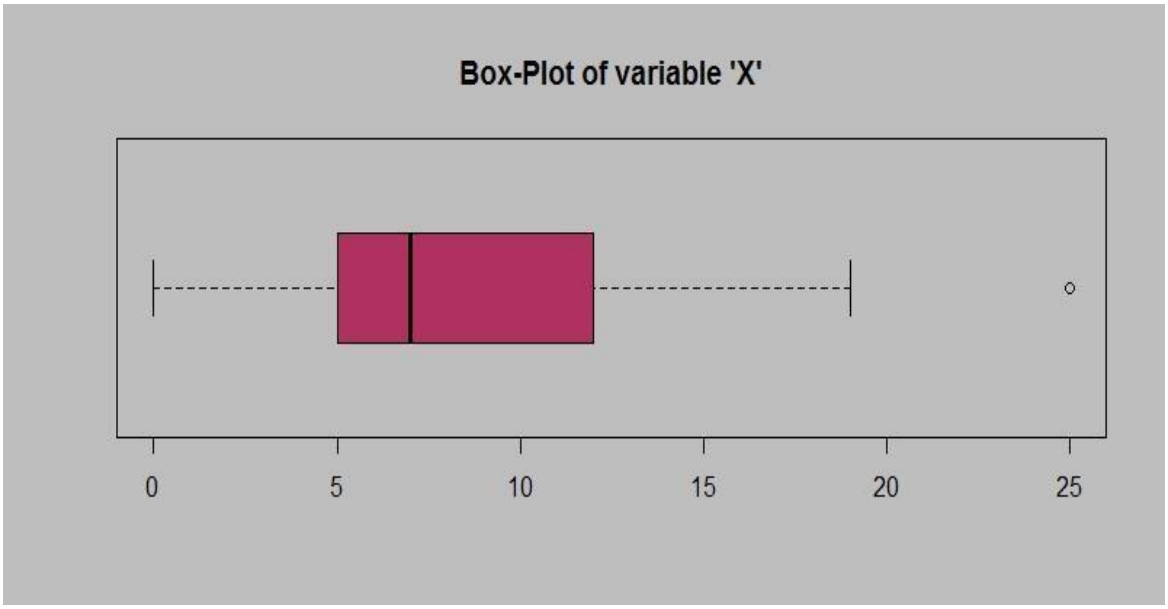
	Percentage
count	15.000000
mean	0.332713
std	0.169454
min	0.241400
25%	0.254700
50%	0.267100
75%	0.339750
max	0.913600



In the given data the company Morgan Stanley is a outlier with 91.36%

Mean (μ) : 0.3327

Standard Deviation (σ) : 0.1695



Variance (σ^2) : 0.0287

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

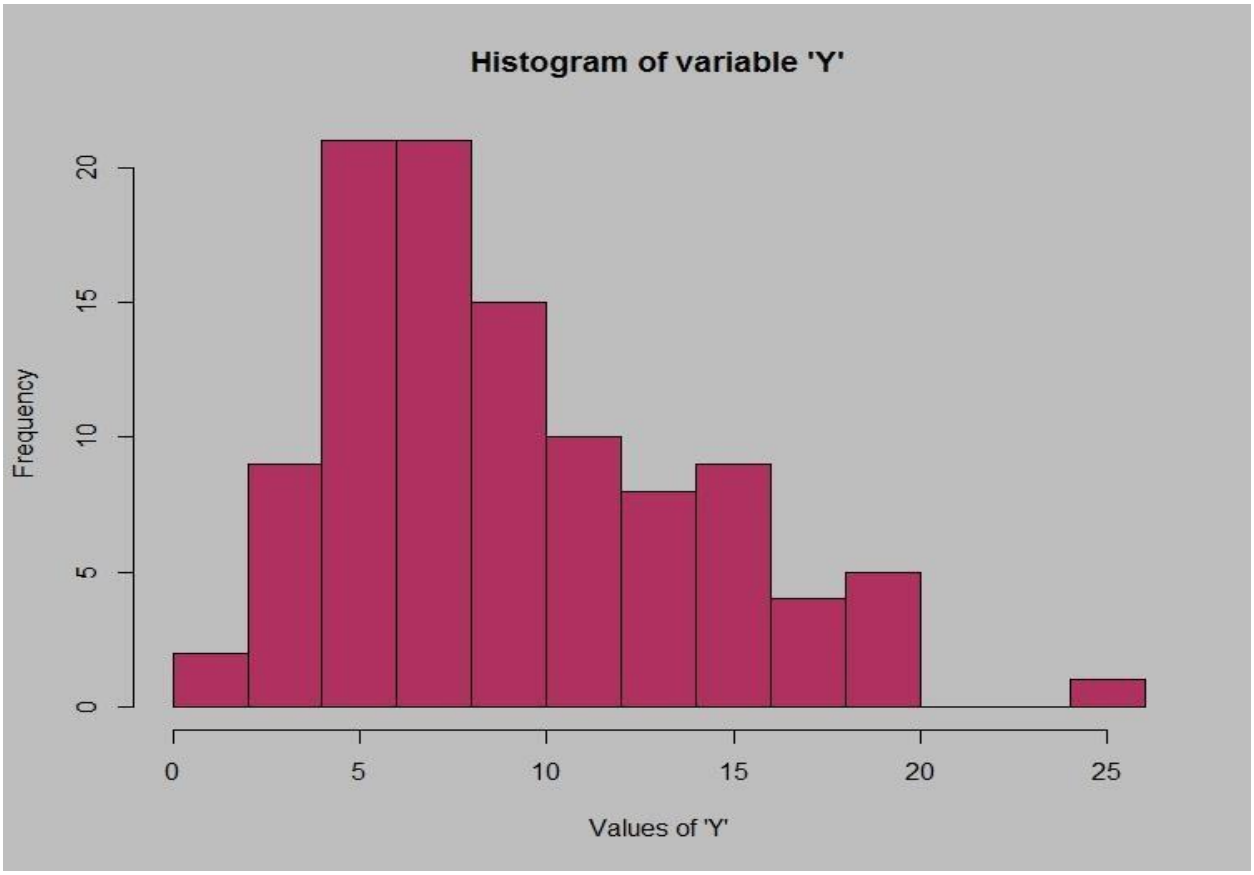
Solution:

(i) IQR = 12-5 = 7, this represents the range which contains 50% of the data points.

(ii) The given dataset is positively skewed.

(iii) The data point value 2.5 will not be considered as an outlier. The boxplot will range from 0 to 19.

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.

Solution:

(i) The mode of this dataset will lie between 4 to 8.

(ii) The dataset is positively skewed.

(iii) The Median in boxplot is 7 and Mode in histogram is lies between 4 to 8.

Histogram provides the frequency distribution so we can see how many times each data point is occurring. However Boxplot provides the quantile distribution i.e. 50% data lies between 5 and 12.

Boxplot provides whisker length to identify outliers, no such information can be concluded from the histogram. We can only guess by looking at the gap, that 25 may be an outlier.

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.)

Solution:

:

Number of Calls(n) = 5, p = 1/200, q = 199/200

P(x) = at least one in five attempted telephone calls reaches the wrong number

$P(x) = {}^{n}C_{x}\, p^{x}\, q^{n-x}$

$P(x) = ({}^{5}C_{1})\,(1/200)^{1}\,(199/200)^{5-1}$

P(x) = 0.0245037

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?

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

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

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

Solution:

(i) For Max. Value of P = 0.3 for P(2000). So most likely outcome is $2000.

(ii) P(x>0) = 0.6, implies there is a 60% chance that the venture would yield profits or greater than expected returns. So the venture is likely to be successful.

(iii) Weighted average = x*P(x) = 800. This means the average expected earnings over a long period of time would be 800(including all losses and gains over the period of time).

(iv) P(Incurring loss) = P(x= -2000)+P(x=-1000)=0.2. So the risk associated with this venture is 20%.

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

Solution:

B. 0.2676

μ=45 min,σ=8 min,

μ=10 min after =45+10=55

Car will be ready in 1 hr=60min(x)

Z=x-μ/σ

=60-55/8

=0.625

1=stats.norm.cdf(z)

1=stats.norm.cdf(0.625)

=0.2659

P value for z score is 0.2359 app. To 0.2676.

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

   Solution:

   p(X>44) = 0.1587

   p(38<X<44) = 0.3413

   The statement is False. The Probability of employees aged from 38 to 44 is more.

   A. N*P(X<30) = 36.49

   The statement is true. The no. of employees aged below 30 years attending training is 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.

   Solution:

   If $X_1 = N(\mu, \sigma^2)$ and $X_2 = N(\mu, \sigma^2)$

   Then, $2X_1 = N(2\mu, 4\sigma^2)$ and

   $X_1 + X_2 = N(\mu, \sigma^2) + N(\mu, \sigma^2) = N(2\mu, 2\sigma^2)$

   $\therefore 2X_1 - (X_1 + X_2) = N(2\sigma^2)$

4. Let X ~ 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

Solution:

Given: p(a<x<b) = 0.99 ,m ean =100,standardDeviation = 20

To Find:

Identify symmetric values for the standard normal distribution such that the area enclosed is .99

From the above details,we have to excluded area of .005 in each of the left and right tails.

Hence, we want to find the 0.5th and the 99.5th percentiles Z score values

Using Python

Z value is given as stats.norm.ppf(pvalue)

Z value at 0.5th percentile is given as

$$Z(0.5) = \text{stats.norm.ppf}(0.005) = -2.576$$

Z value at 99.5 percentile is given as

$$Z(99.5) = \text{stats.norm.ppf}(0.995) = 2.576$$

$Z = (x - 100)/20 => x = 20z+100$

$a = -(20*2.576) + 100 = 48.5$

$b = (20*2.576)+100 = 151.5$

Two values symmetric about mean for the given standard normal distribution are[48.5,151.5]

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.
B. Specify the 5^{th} 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?

Solution:

A) the range for 95% probability for the company is Rs.99 to 980.99 Million.

B) The 5^{th} percentile of profit (in million rupees)is 202.

C) Probability of division 1 making a loss P(X<0) 0.047

Probability of division 2 making a loss P(X<0)0.040

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. - **False**

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

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: **All the readers of PC Magazine**
   B. The parameter of interest: **Sample size, average, scale**
   C. The sampling frame: **9000 readers**
   D. The sample size: **225 readers**
   E. The sampling design: **μ=7.5**
   F. Any potential sources of bias or other problems with the survey or sample : **No potential sources of bias or other problems with the survey or sample**

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. - **True**

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**

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

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

   A. ¼
   B. ½
   C. ¾
   D. 1

   Solution: B. ½

5. 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.

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

6. 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

Solution: A. The 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.

7. 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

Solution: A) 600.

P=0.50

Z= 1.960 for 95% confidence interval

N=(z/M)^2(0.5)(1-0.5)

=600.25.

8. 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

Solution: C) 848.

Z= 2.576 for 98% confidence interval

0.04=2.326 *√0.5*0.5/n

N=(2.3262)^2*0.5*0.5/(0.04)^2

=1.3525/0.0016

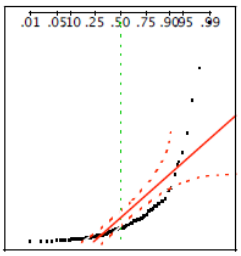
=845.35

=848.

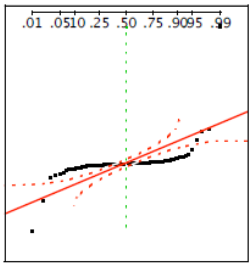
CBA: Practice Problem Set 2
Topics: Sampling Distributions and Central Limit Theorem

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

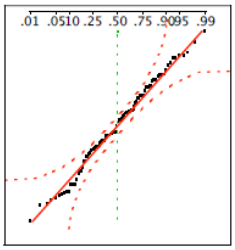
A



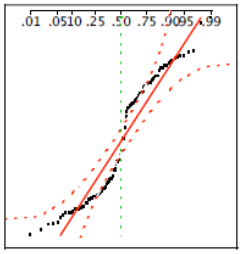
B



C



D



Solution:
1) C
2) B, D
3) A, B, D
4)A, B

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: No, there is no need to check the weight of individual packages are normally distributed.

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

Solution:
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%

   Solution:

   D. 21.1%

   pnorm(55,50,4) = 0.89435022633

   pnorm(45,50,4) = 0.105649773666855 = 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.

   A. 144
   B. 150
   C. 196
   D. 250
   E. Not enough information

   Solution:

   D.250

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

Solution:

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