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

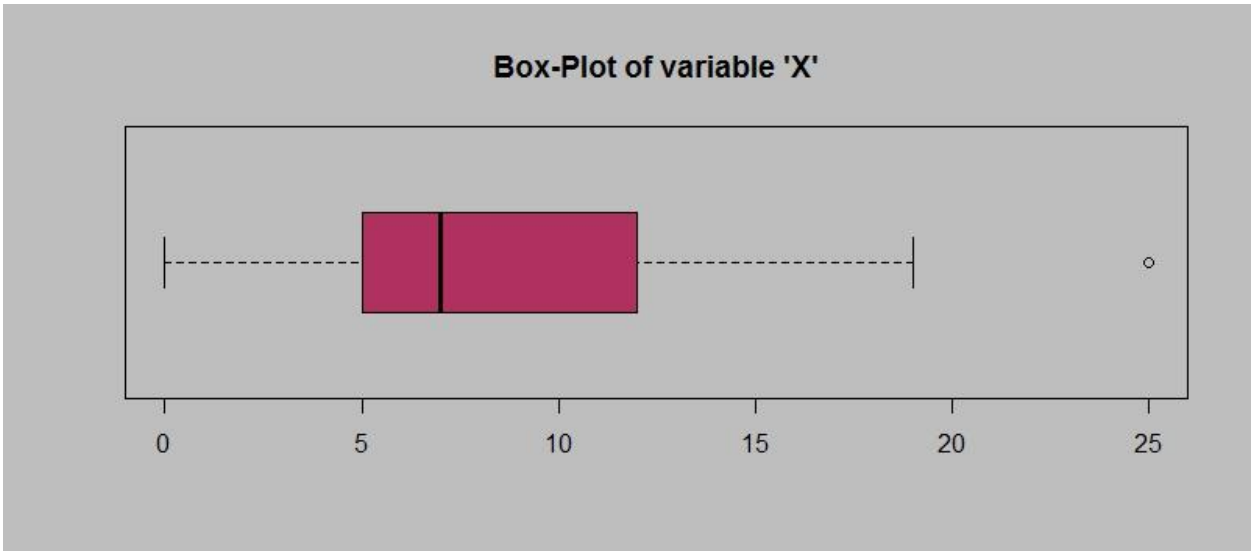
Name of company	Measure X
Allied Signal	24.23%
Bankers Trust	25.53%
General Mills	25.41%
ITT Industries	24.14%
J.P.Morgan & Co.	29.62%
Lehman Brothers	28.25%
Marriott	25.81%
MCI	24.39%
Merrill Lynch	40.26%
Microsoft	32.95%
Morgan Stanley	91.36%
Sun Microsystems	25.99%
Travelers	39.42%
US Airways	26.71%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

ANS:-

Plot the box-plot : box=plt.boxplot(level2.Measure)
plt.ylabel("Measure in %")

1. Outlier : [item.get_ydata() for item in box['fliers']]
91.36
2. Mean : level2.Measure.mean ()
33.271 %
0.3327
3. STD : A = level2.Measure.std ()
16.94 %
0.1694
4. (STD)^2 : B = A *A
287.14 %
2.8714

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?

ANS:-

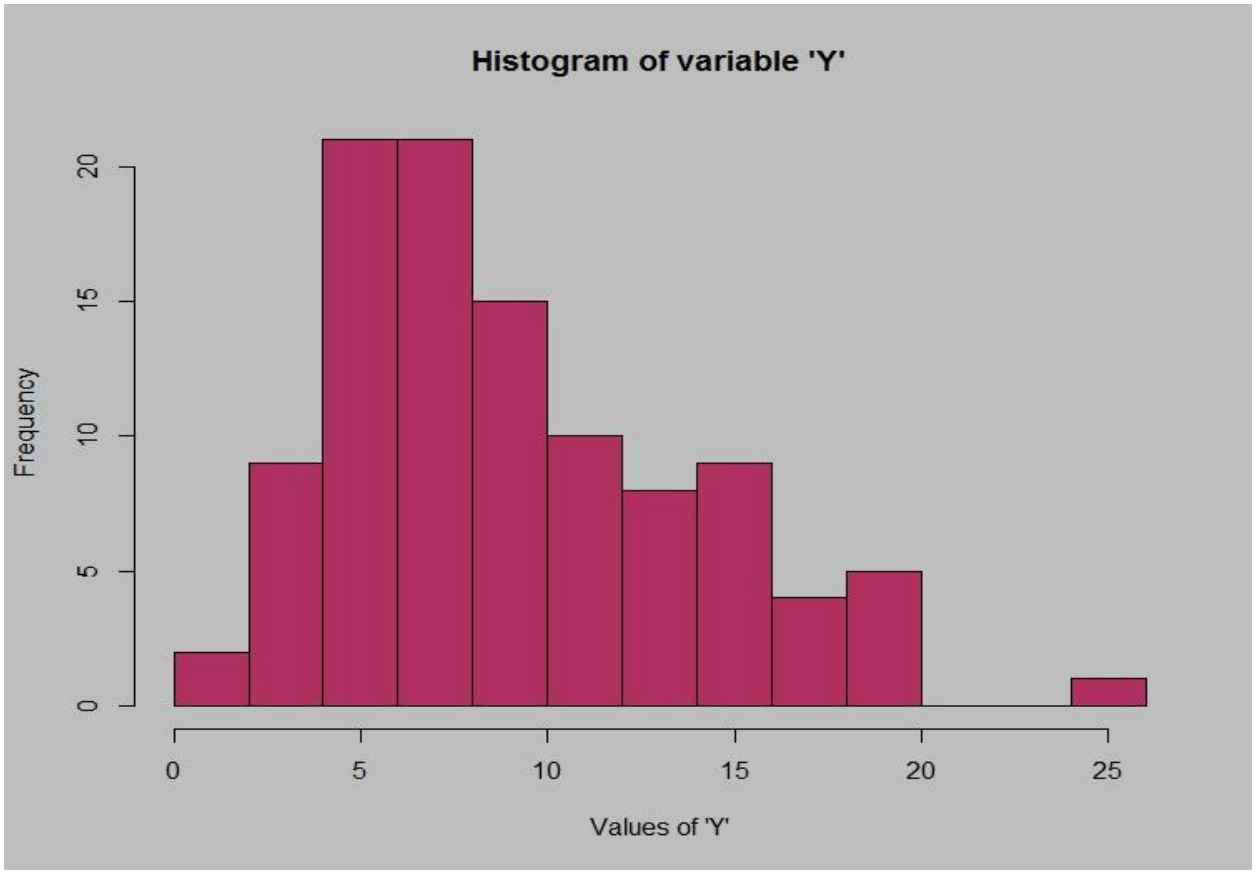
1) **inter-quartile range (IQR) = Q3 – Q2**
   = 12 - 5
   = 7

2) **From the above figure we can say that the these has positive skewness / Right skewed .**

3) **If it was found that the data point with the value 25 is actually 2.5, then 2.5 will not be Considered as outlier**

3) 2.5 will be not considered an outlier. The boxplot will start from 0 and send at 20 in representation

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.

ANS:-

1. The above Histogram we can say that the mode is lie between 4 to 8 .

2. As most of the data lies between left side of the graph we can say that it is positively skewed.

3. We can say that 50% of data lies in between 5 to 12 .Histogram provide frequency distribution and box plot is providing most of the body lies in between 5 to 12. From both figures we can say that 25 is a outlier

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

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:-

Out of 200 call one call is misdirecting

P (call is misdirecting) = 1/200

P (call is not misdirecting) = (1 - 1/200) = 199/200

We have formula p(x) = nCr * P(success)^r * P(failed)^(n-r)

Where ,

n=5

r= 0

P(call is not misdirecting) = [(5*4*3*2*1) / (5*4*3*2*1)] x (1/200)*0 x (199/200)*5

= (199/200)*5 = 0.975

P(call is misdirecting) = 1 – 0.975 = 0.025

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

ANS:-

1. **Most likely momentary outcome of the business venture is 2000 $ as it has maximum probability amongst all which is 0.3**

2. **-2000x0.1 + (-1000x0.1) + (0x0.2) + 1000x0.2 + 2000x0.3 + 3000x0.1**

   = 800 $

   As the average of the above is in positive we can say that the venture likely to be successful.

3. **As have already calculated above the long terms average earning will be 800 $**

4. **. 4. we will take loss of (-2000) and another one is (-1000).**
   There probability will be 0.1+0.1 = 0.2
   The risk involved will be 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

ANS : -

Given mean = 45 , standared deviation = 8

As per given condition the work will start after 10 min so mean will be 45+10 = 55

We have asked customer the car will be ready after 1 hour(x) = 60

Hence, Z = (60 – 55) / 8 = 0.625

From z table

Z value will be 0.73237

the probability that the service manager cannot meet his commitment

= 1 – 0.73237 = 0.267

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.

ANS:-

A. First find out the probability for employees older than 44 :

X = 44 , mean = 38, std = 6

1 - Stats.norm.cdf (44 , 38 , 6) = 1- 0.8413

= 0.1587

Now find the probability of employees between 38 and 44 :-

Stats.norm.cdf (44 , 38 , 6) - stats.norm.cdf (38,38, 6)

= 0.8413 - 0.5 = 0.3413

As we can clearly see that the probability of employees age between 38-44 is more than employees age more than 44.

So the given statement is False.

B. Lets calculate the probability of employees age under 30 :

X = 30 , mean = 38, std = 6
Stats.norm.cdf (30 , 38 , 6) = 0.0912

So the total number of emplyees age under 30 is 0.0912*400 = 36.48

So we can say that the statement is 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 .

ANS:-

Here, x1 and x2 are random variables which have same distribution and independent of each other

We have to find the sum of the mean and the variance

Sum of mean = 2 μ

Sum of the variance = 2 σ 2

There is no any difference between 2x1 and x1+x2 as both of them have same Distribution

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:-

Mean = 100 and std = 20

probability of the random variable taking a value between them is 0.99

hence,

Z value at 99 % = Stats.norm.ppf (0.995) = 2.5758

1 st value will be 2.5758*20 + 100 = 151.5
2 nd value will be (-2.5758)*20 + 100 = 48.484

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.
   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?

ANS:-

Total profit = profit 1 + profit 2

Mean = profit 1 (mean) + profit 2 (mean)

= 5 + 7 = 12

Std = sqrt (9+16) = sqrt (25) = 5

Mean in rs = 12*45 = 540

Std in rs = 5* 45 = 225

A) Range for 95 % :-

Stats.norm.interval (0.95 , 540 , 225)

Range is rs (99.008 , 980.991) in millions

B) the 5th percentile :-

From z score we need to find the value of 0.5000-0.050 = 0.4500

We are getting the value of -1.645

the 5th percentile of profit = mean + (-1.645)*std

= 540 – (1.645 * 225
= 540 – 370.125 = 169.87

= 170 in million

C) Probability of 1st division making loss = stats.norm.cdf (0, 5 , 3)

= 0.0479

Probability of 2nd division making loss = stats.norm.cdf (0, 7, 4)
= 0.04005

We can see that 1st division can make more loss compared to 1st division.

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:- TRUE

    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

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

    ANS:-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
    B. The parameter of interest
    C. The sampling frame
    D. The sample size
    E. The sampling design
    F. Any potential sources of bias or other problems with the survey or sample

ANS:-

A) All the readers who rated which is 9000.

B) Rating of Kodak compact digital camera (7.5)

C) All the readers 9000

D) 225

E) Readers rated the products on a scale from 1 to 10 .

F) It is possible that only those who who were particularly please or who are unpleased with the product participated in survey which can make the 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 confiden

ANS:-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.

ANS :- False, we have direction but we can not 100% biased on that data.

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

ANS :- FALSE , there is no need of sample data should be normally distributed , but the sample size must be greater than 30.

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

    A. ¼
    B. ½
    C. ¾
    D. 1

ANS :- It has 50% 50% chances that mean of sample can be greater than mean of population.

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?

    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 :-

1. **) h0 = mozilla has more than 5% share of the market ; h0 > 5%**
   ha = Mozilla has less than 5% share of the market ; ha < 5%

Apply one sample one tail Z-test :-

z-score = (0.045 – 0.05) / np.sqrt ((0.05 * (1-0.05))/2000)

z-score = - 0.82078

from z table we get the value of 1.96

so we will go with the null hypothesis

so we conclude that Mozilla has more than5% share of the market.

2. **) WebSideStory claims that its sample includes all the daily Internet users.**
   That means 4.6 % share of the market shows for entire population.
   So Microsoft **conclude that 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.

      ANS:-INCORRECT

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

      ANS:-INCORRECT

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

      ANS:-CORRECT

   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:-INCORRECT

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

      ANS:-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

ANS:- 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.

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 :-

z-value of 95% confidence is 1.96
Stats.norm.ppf (0.975)
Margin of error = 0.04
ME = z* sqrt ((p*q)/n)

We have to find the n
Hence,
N = p*q / ME^2 * z
Assume p= 0.5 and q = 0.5
Then n = (05*0.5)/ (0.04*0.04) * 1.96^2 = (0.25/0.0016) * 3.8416
= 600.25
Option A is correct which is 600.

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 :-

z-value of 98% confidence is 2.326
N = (0.25/0.0016) * 5.41 = 1.352 / 0.0016 = 845.35

Option C is correct which is 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 …

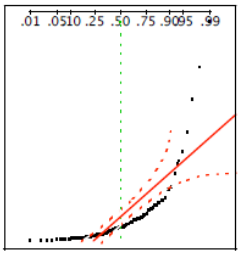
   I. Are nearly normal?
   ANS:-plot C

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

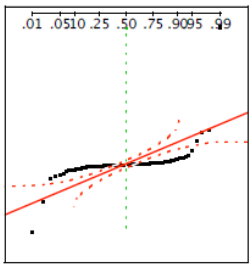
   III. Are skewed (i.e. not symmetric) ?
   ANS:-plot A and D

   IV. Have outliers on both sides of the center?
   ANS:-plot A

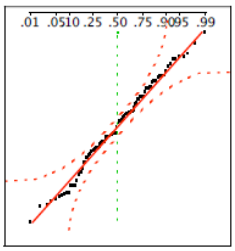
A



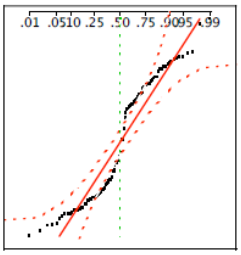
B



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:- FALSE

   (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 : -

Given : population mean = 50 , std = 40

between $45 and $55 :-

z1 = (55-50) / (40/sqrt 100) = 5 / 4

z1 = 1.25

z1 value = 0.8943

z2 = (45-50) / (40/sqrt 100) = -5 / 4

z2 = -1.25

z2 value = 0.1056

z value between 45 and 55 = z1 –z2 = 0.8943 - 0.1056 = 0.7887
So this is the probability when auditor will not investigate .
The probability when the auditor will investigate is 1 – 0.7887 = 0.2113 = 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:-

For 5 % , z will be -/+ 1.96

Z = 5* sqrt (n)/ 40

Sqrt (n) = 15.68

n = 245.86

It is nearly equal to the option D.

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.
      ANS :- FALSE, we do not know the sample size.
   B. The standard deviation of the mean of across several samples will be 120.
      ANS :- FALSE
   C. The mean score in any sample will be 720.
      ANS :- TRUE, but it can be less or more
   D. The average of the mean across several samples will be 720.
      ANS :- TRUE
   E. The standard deviation of the mean across several samples will be 0.60
      ANS :- TRUE

std / sqrt(n) = 120 / sqrt(40000)

= 120 / 200 = 0.6