Basic statistics Assignment-02

Balaji N
nbalaji743@gmail.com

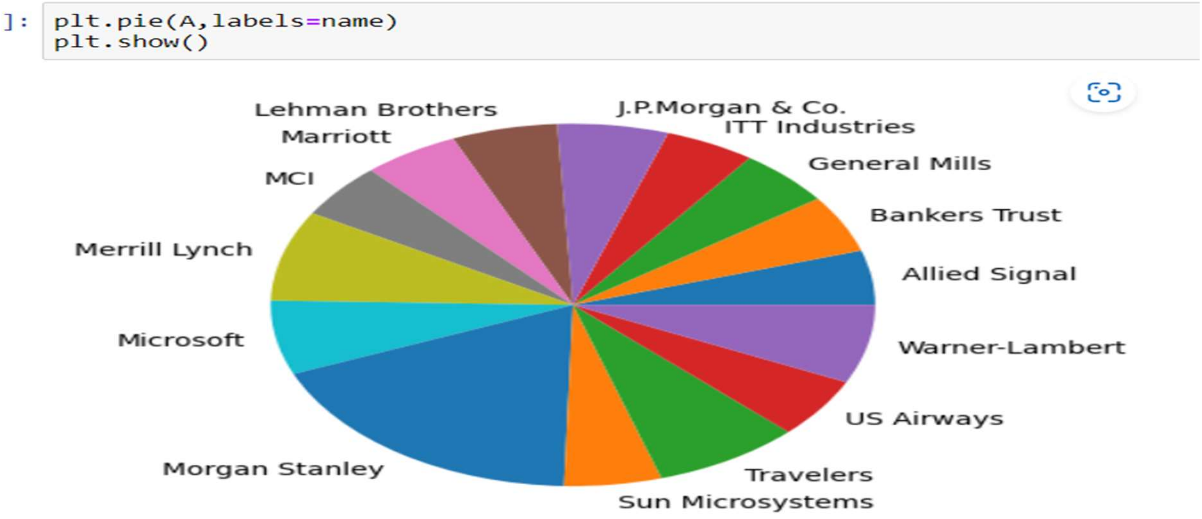
SET-01 Topics: Descriptive Statistics and Probability

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:

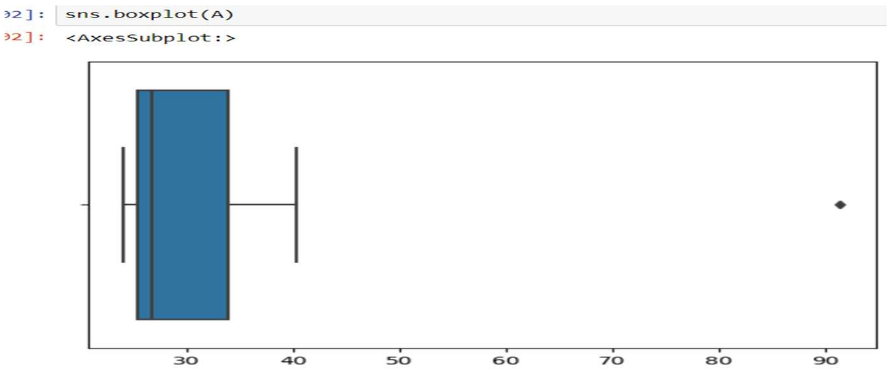
- **Plot of the data**



```
]: plt.pie(A,labels=name)
   plt.show()
```

Basic statistics Assignment-02

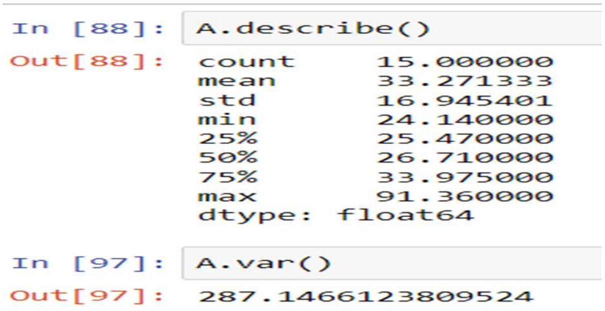
Balaji N
nbalaji743@gmail.com



```
2]: sns.boxplot(A)
2]: <AxesSubplot:>
```

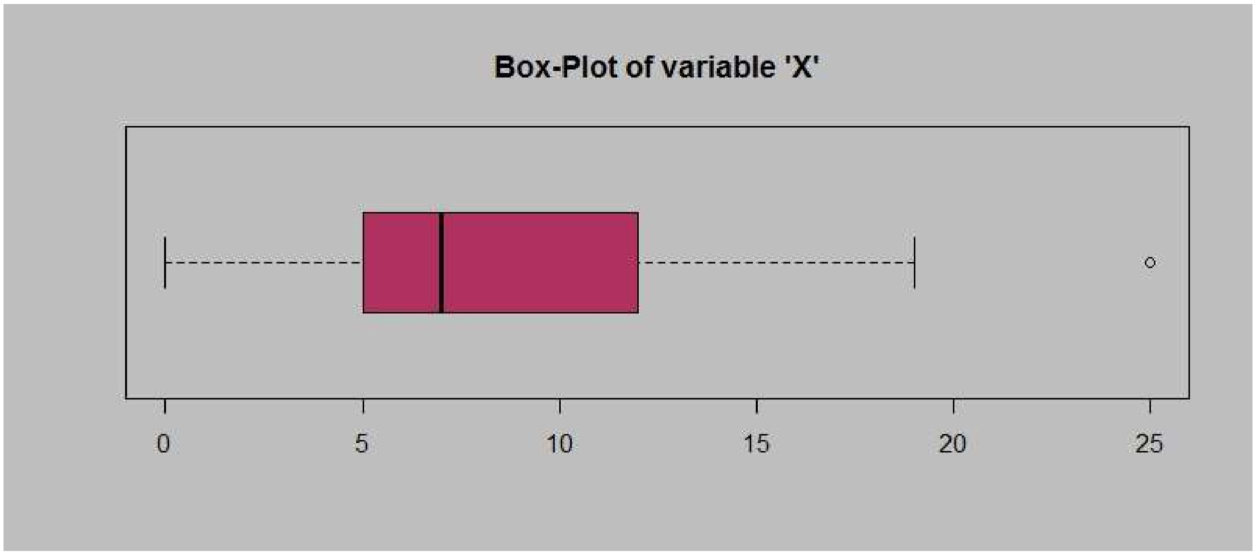
- Outlier is **Morgan Stanley 91.36%**

- Mean= **33.27**
- Standard Deviation= **16.94**
- Variance= **287.146**



```
In [88]:  A.describe()
Out[88]:  count      15.000000
          mean       33.271333
          std        16.945401
          min        24.140000
          25%        25.470000
          50%        26.710000
          75%        33.975000
          max        91.360000
          dtype: float64

In [97]:  A.var()
Out[97]:  287.1466123809524
```

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.

- IQR = Upper quartile-Lower Quartile=12-5=7

- 50% of the data lies in this IQR , and 1.5 IQR from the upper and lower quartile denotes upper and lower extreme points.

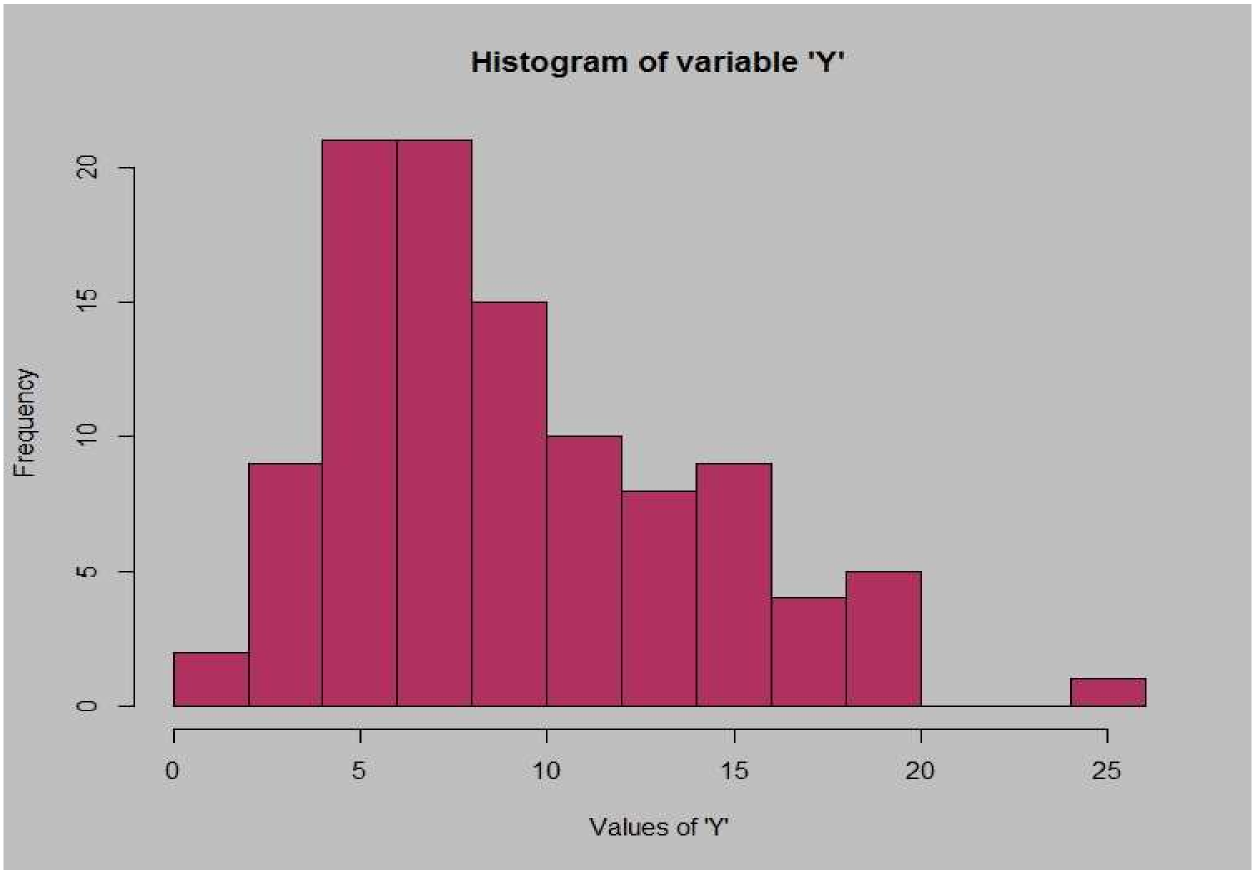
(ii) What can we say about the skewness of this dataset?

- This given dataset is Right Skewed as the size of the plot towards the right side of the mean is more

(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?

- This change in value will affect the mean in a way that it reduces , and the median also reduces

3.



Answer the following three questions based on the histogram above.

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

This dataset is multimodal, mode is at 4 to 6 values of Y

(ii) Comment on the skewness of the dataset.

This is Right Skewed

(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.

- We can find the outlier and the kind of skewness caused in the dataset from both histogram and boxplot.
- Also the histogram gives the frequency of the distribution of values of Y and the box plot provides the IQR and the Whiskers.

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:
The total no long distance calls = 200
Probability of call being misdirected= 1/200

Sample calls = 5
Probability of at least one in five attempted calls reaches wrong number = 5*(1/200) = 0.025 = **2.5%**

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?

The value of x with highest probability= 2000

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

Probability of the venture to fail = 0.1+0.1 = 0.2

Probability of the venture to neither fail nor be successful = 0.2

Probability of the venture to be successful= 1-(0.2+0.2) = 0.6=60%

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

Expected mean = $\sum X*p(X)$

= [(-2000*0.1)+(-1000*0.1)+(0)+(1000*0.2)+(2000*0.3)+(3000*0.1)]

Average earnings = 800

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

Risk involved in a venture = $E(X^2)-[E(X)]^2$ { E(X2)= X2*P(X) }

$$Var = 2800000 - 800^2 = 2160000$$

Standard Deviation = √var=1470

Since the deviation from the expected value is more high **the risk involved is also very high.**

Basic statistics Assignment-02

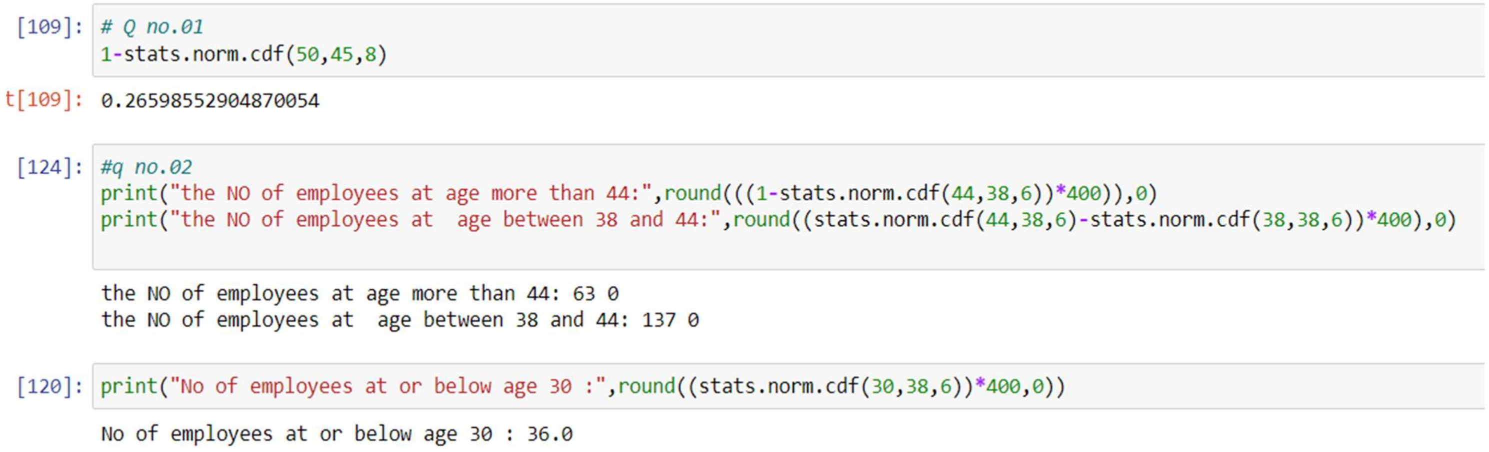
Balaji N
nbalaji743@gmail.com

SET-02 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
   Since the service manager plans that servicing begins after 10 mins ,
   hence the available time for servicing = 60-10 = 50 minutes
   By using distribution function **1-stats.norm.cdf(50,45,8) = 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.



```
[109]: # Q no.01
       1-stats.norm.cdf(50,45,8)

t[109]: 0.26598552904870054
```

```
[124]: #q no.02
       print("the NO of employees at age more than 44:",round(((1-stats.norm.cdf(44,38,6))*400)),0)
       print("the NO of employees at  age between 38 and 44:",round((stats.norm.cdf(44,38,6)-stats.norm.cdf(38,38,6))*400),0)

       the NO of employees at age more than 44: 63 0
       the NO of employees at  age between 38 and 44: 137 0
```

```
[120]: print("No of employees at or below age 30 :",round((stats.norm.cdf(30,38,6))*400,0))

       No of employees at or below age 30 : 36.0
```

   A. More employees at the processing center are older than 44 than between 38 and 44.
   The statement is false
   Because, no of employees older than 44 is = 63
   No of employees between 38 and 44 is = 137
   B. A training program for employees under the age of 30 at the center would be expected to attract about 36 employees.

The statement is true
{ the python statement is shown above }

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:

$2\ X1 = N(2\mu, 4\ \sigma 2)$

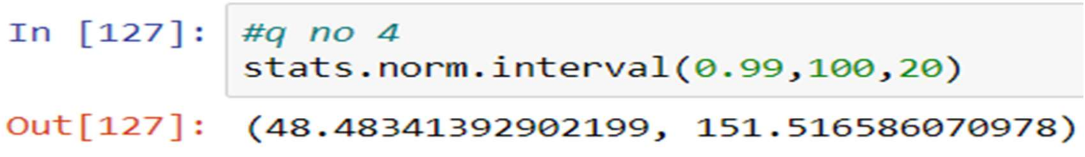
$X1 + X2 = N(\mu 1 + \mu 2, \sigma 12 + \sigma 22) = N(2\mu, 2\ \sigma 2)$

$[2X1-(X1+X2)] = \mathbf{N(0, 6\ \sigma 2)}$

- The mean of 2 X1 and X1 + X2 are same = 2 μ
- The variance of 2 X1 is twice as the variance of X1 + X2

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



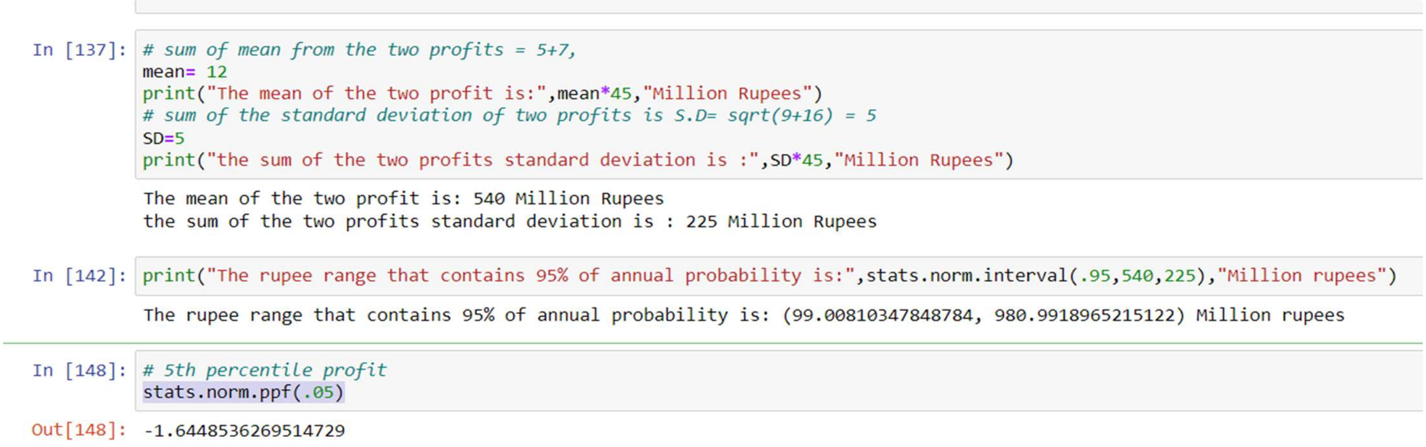
```
In [127]: #q no 4
          stats.norm.interval(0.99,100,20)

Out[127]: (48.48341392902199, 151.516586070978)
```

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

Basic statistics Assignment-02

Balaji N
nbalaji743@gmail.com

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



```
In [137]: # sum of mean from the two profits = 5+7,
          mean= 12
          print("The mean of the two profit is:",mean*45,"Million Rupees")
          # sum of the standard deviation of two profits is S.D= sqrt(9+16) = 5
          SD=5
          print("the sum of the two profits standard deviation is :",SD*45,"Million Rupees")

          The mean of the two profit is: 540 Million Rupees
          the sum of the two profits standard deviation is : 225 Million Rupees

In [142]: print("The rupee range that contains 95% of annual probability is:",stats.norm.interval(.95,540,225),"Million rupees")

          The rupee range that contains 95% of annual probability is: (99.00810347848784, 980.9918965215122) Million rupees

In [148]: # 5th percentile profit
          stats.norm.ppf(.05)

Out[148]: -1.6448536269514729
```

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

In normal distribution, $z = \frac{x-\mu}{\sigma}$

$$x = \sigma z + \mu$$

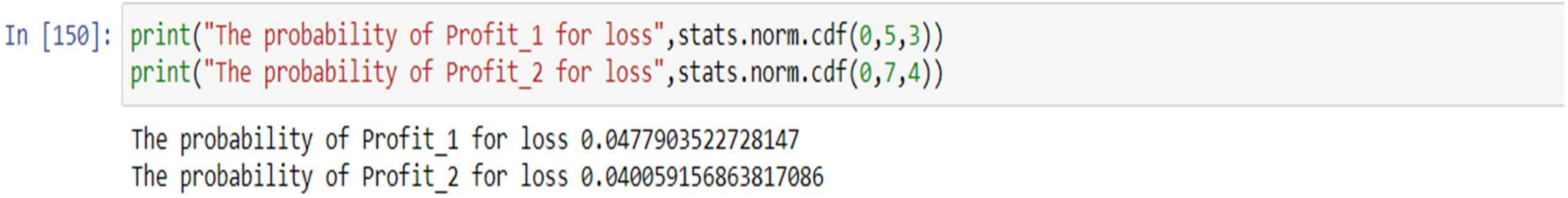
For the 5th percentile(stats.norm.ppf(.05)) , z= -1.64 , μ=540, σ=225 million rupees

X=540+(-1.64*225)

= 171 million rupees

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

For making loss , X<0



```
In [150]: print("The probability of Profit_1 for loss",stats.norm.cdf(0,5,3))
          print("The probability of Profit_2 for loss",stats.norm.cdf(0,7,4))

          The probability of Profit_1 for loss 0.0477903522728147
          The probability of Profit_2 for loss 0.040059156863817086
```

Therefore the profit_1 has the larger probability of making a loss in a given year

Basic statistics Assignment-02

Balaji N
nbalaji743@gmail.com

SET-03 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.

   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.

   False

   The sampling frame is a list of every item in a population that appears in a survey sample , but those did not respond to the questions will not be considered under sampling frame**, sampling frame includes only those responds to 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

   The readers of the PC ,

   Population Mean (μ) = $\sum X / N = 225/9000 = 0.025$

   B. The parameter of interest

   Sample size, population mean, rating scale

C. The sampling frame

Sampling frame is the 9000 readers

D. The sample size

The sample size is the 225 readers

E. The sampling design

The sample design is of voluntary response which is 9000 readers(only those who used the product would voluntarily respond)

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

No , because of the limited information and voluntary response of the sample design there would no potential source of bias

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.

True

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

False (Confidence intervals can be used with distributions that aren't normal**)**

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

   A. ¼

   B. ½

   C. ¾

   D. 1

There is 50% of chance that the sample mean is greater the population mean in the normal distribution

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?

   Let us take the

   Null Hypothesis , H0: P ≥ 5% {meaning Mozilla has more than or equal to 5 percent share of the market}

   Alternative Hypothesis , H0: P< 5% {meaning Mozilla has less than 5 percent share of the market}

   Test statistics = (0.046 – 0.05)/√ [0.05(1-0.05)]/2000

   = - 0.833

   The critical value of z for 5% significance level = -1.96

   Test statistics > Z (-1.96)

   (-0.833)

   The null hypothesis is true , hence mozilla has more than or equal to 5% market share.

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?

Answer: It is claimed by WebSideStory that their sample contains all the internet users using daily. Thus it means 4.6 percent share of market shows the entire population.

With this we can conclude 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.

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.

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

The 95% interval of for z-score is (-1.96, 1.96)

The 95% interval of for t-score is (-2.58, 2.58) and also the z-score is always shorter than the t score

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

Solution:

Margin of Error = Z-Score × (S ÷ √n)

The Z-score at 95% confidence interval= 1.96 (stats.norm.ppf(0.975)

(Margin of Error)2 = (Z-Score)2× (S2 ÷ n)

(Margin of Error)2 = (Z-Score)2× (P(1-P) ÷ n)

n = [(Z-Score)2× P(1-P)] / (Margin of Error)2

= 1.962 * 0.5(1-0.5) / (0.04)2

N = 600

600 employers are randomly chosen in order to guarantee a margin of error is not more than 4%

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

Margin of Error = Z-Score × (S ÷ √n)

The Z-score at 98% confidence interval= 2.33 (stats.norm.ppf(0.99)

(Margin of Error)2 = (Z-Score)2× (S2 ÷ n)

(Margin of Error)2 = (Z-Score)2× (P(1-P) ÷ n)

n = [(Z-Score)2× P(1-P)] / (Margin of Error)^2

= 2.33^2 * 0.5(1-0.5) / (0.04)^2

N = 848

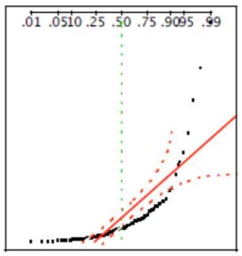
848 employers are randomly chosen in order to guarantee a margin of error is not more than 4%s

---

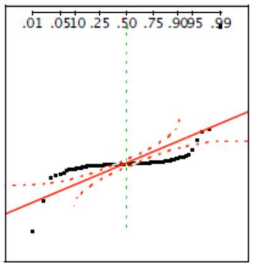
CBA: Practice Problem Set 4
Topics: Sampling Distributions and Central Limit Theorem

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.) the bimodal distribution has two peaks **Ans: B, D**
   - III. Are skewed (i.e. not symmetric)? **Ans: A, B, D**
   - IV. Have outliers on both sides of the center? **Ans: A**

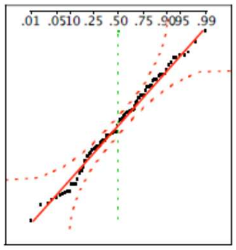
A



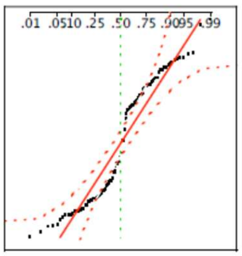
B



C



D



Basic statistics Assignment-02

Balaji N
nbalaji743@gmail.com

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 $\mu = 22$ lbs. and $\sigma = 5$ lbs.

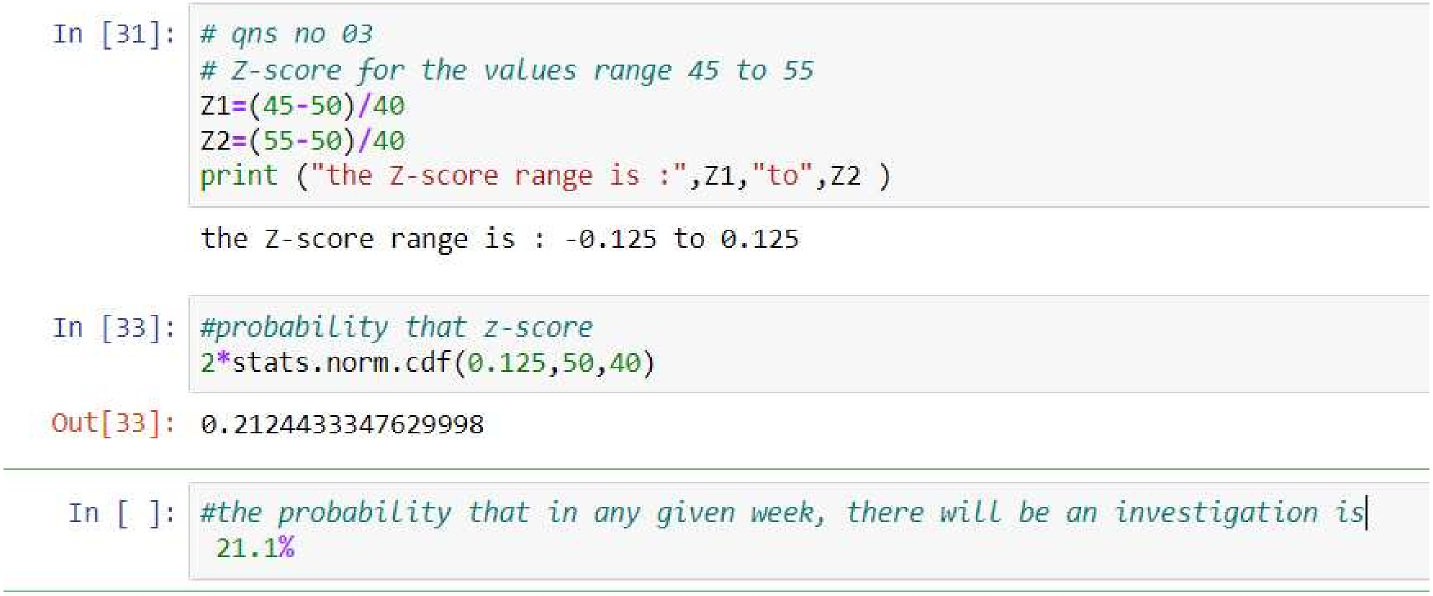
   (i) Before using a normal model for the sampling distribution of the average package weights, the manager must confirm that weights of individual packages are normally distributed.
   Answer: True
   The distribution of the sample is a normal model if the individual packages are normally distributed and also the distribution is normal if the sample is fairly large (Central limit theorem)

   (ii) The standard error of the daily average $SE(\bar{x}) = 1$. **Answer: True**
   Standard error = σ/√n
   = 5/√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%

set - 04

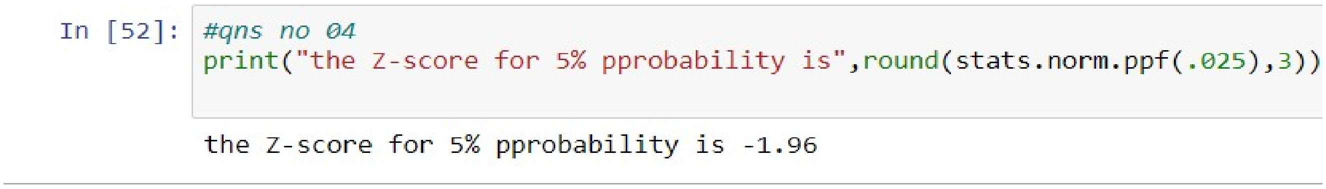


```
In [31]: # qns no 03
         # Z-score for the values range 45 to 55
         Z1=(45-50)/40
         Z2=(55-50)/40
         print ("the Z-score range is :",Z1,"to",Z2 )

         the Z-score range is : -0.125 to 0.125

In [33]: #probability that z-score
         2*stats.norm.cdf(0.125,50,40)

Out[33]: 0.2124433347629998

In [ ]: #the probability that in any given week, there will be an investigation is
          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:



```
In [52]: #qns no 04
         print("the Z-score for 5% pprobability is",round(stats.norm.ppf(.025),3))

         the Z-score for 5% pprobability is -1.96
```

$$\boldsymbol{Z} = \frac{\boldsymbol{x} - \mu}{\frac{\sigma}{\sqrt{\boldsymbol{n}}}}$$

$$-\mathbf{1.96} = \frac{\mathbf{5}}{\mathbf{40}/\sqrt{\boldsymbol{n}}}$$

√n=1.9*40/5=15.7
N= 247 ≈ **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. **False**
B. The standard deviation of the mean of across several samples will be 120. **False**
C. The mean score in any sample will be 720. **True**
D. The average of the mean across several samples will be 720. **True**
E. The standard deviation of the mean across several samples will be 0.60. **True**
SEM= 120/√40000=0.60