**Set+1\_Descriptive+statistics+Probability.**

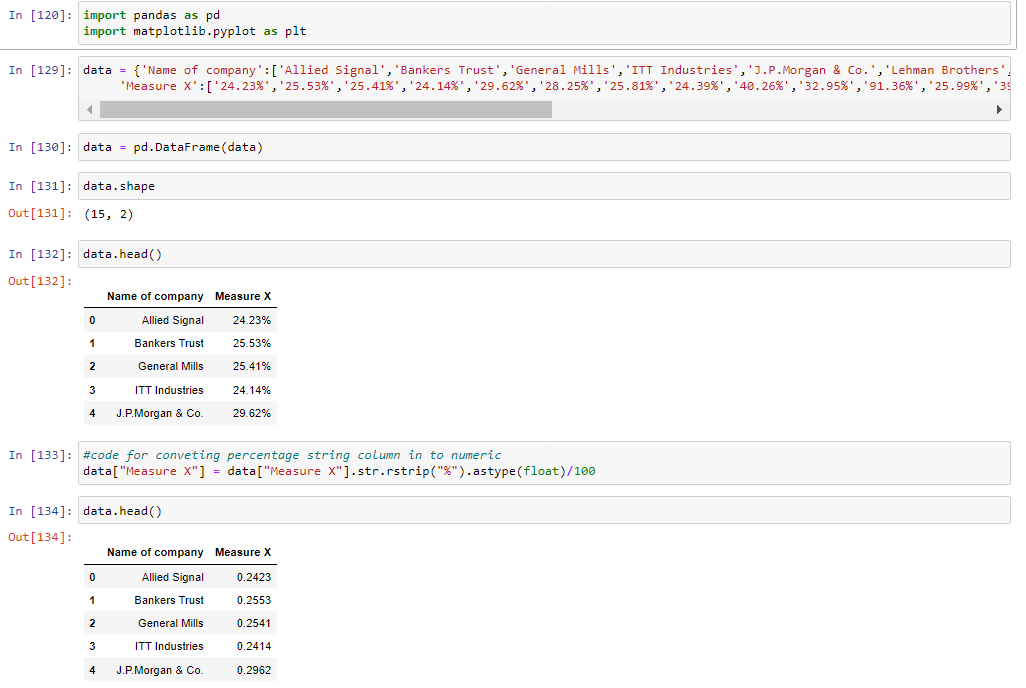
**Topics: Descriptive Statistics and Probability**

1. Look at the data given below. Plot the data, find the outliers and find out

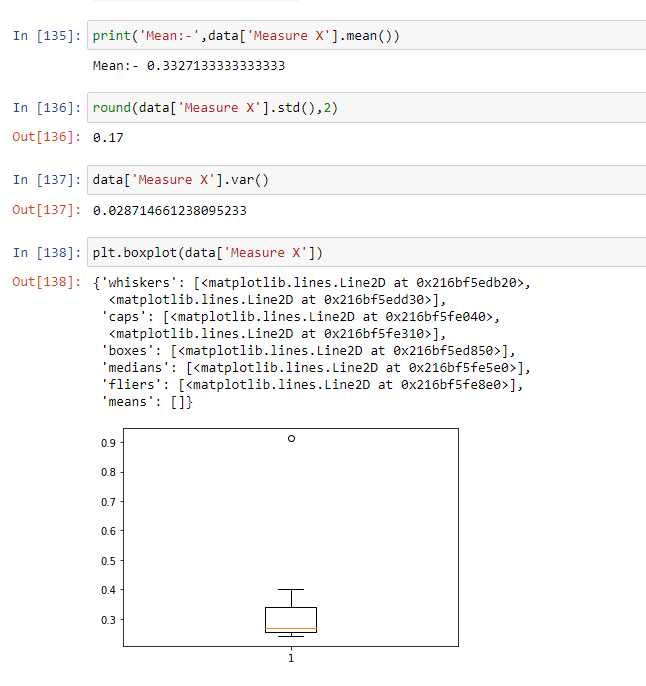
|  |  |
| --- | --- |
| **Name of company** | **Measure X** |
| Allied Signal | 24.23% |
| Bankers Trust | 25.53% |
| General Mills | 25.41% |
| ITT Industries | 24.14% |
| J.P.Morgan & Co. | 29.62% |
| Lehman Brothers | 28.25% |
| Marriott | 25.81% |
| MCI | 24.39% |
| Merrill Lynch | 40.26% |
| Microsoft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

Ans.

1) Plot the data:-



2) Find the outliers and find out:-



As we can see from above plot,

There is one outlier approximate to 0.10

1. Answer the following three questions based on the box-plot above.



Answer the following three questions based on the box-plot above.

1. What is inter-quartile range of this dataset? (please approximate the numbers) In one line, explain what this value implies.

Ans. The inter-quartile range of this dataset if from 5 - 12, And in this range maximum

(50%) amount of data points are there.

1. What can we say about the skewness of this dataset?

Ans. The skewness of this dataset is Right Skewed.

1. 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. As we can see in above boxplot stats from 0 and end to 20. And 25 is outlier for above box plot. So if the data point with the value 25 is actually 2.5 will not be considered an outlier.

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



1. Where would the mode of this dataset lie?

Ans. The mode always occurs at the highest point of the peak. So, Mode lies between 4 and 7 for this dataset.

1. Comment on the skewness of the dataset.

Ans. The skewness of this dataset is Right Skewed.

1. 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. Histogram shows us frequency distribution from that we can get to known about how many times each data point of the dataset is occurring.

Boxplot shows us the quantile distribution in which as we can see that the maximum (50%) amount of data points lies between 5 and 12.

The outlier in boxplot is easily identified shown in question 2 boxplot diagram ( 25 is outlier).

The outlier in histogram is easy to spot. The points which are very far away from other data points (as we can see in above hist. plot 25 data point which is lying very far away from other data point so, 25 is outlier).

1. 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.

Given :- One in 200 long-distance telephone calls is misdirected.

One in 200 long-distance telephone calls is misdirected.

Therefore, probability of call misdirecting  p = 1/200

Probability of call not Misdirecting = 1 - 1/200 = 199/200

Number of Calls = 5

P(x) = (nCx)(p^x)(q^(n-x))………………………..………*Formula*

n = 5

p = 1/200

q = 199/200

So, At least one in five attempted telephone calls reaches the wrong number = 1 - none of the call reaches the wrong number

= 1 - P(0)

= 1 - (5C0)((1/200)^0)((199/200)^(5-0))

= 1 - (199/200)^5

= 0.02475

Therefore, probability that at least one in five attempted telephone calls

reaches the wrong number = 0.02475.

1. 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 |

1. What is the most likely monetary outcome of the business venture?

Ans. As we can see from above table the most likely monetary outcome of the business venture is 2000 as it has **maximum Probability 0.3**

1. Is the venture likely to be successful? Explain

Ans. Venture is successful if x is Positive.

Hence from above table x is 1000 , 2000 or 3000

So, P(1000) + P(2000) + P(3000) = 0.2 + 0.3 + 0.1 = 0.6

As, P(x > 0) = 0.6, implies there is a 60% chance that **venture is likely to be successful.**

1. What is the long-term average earning of business ventures of this kind? Explain

Ans. From above table weighted average = x\*P(x) = 800.

Therefore the average expected earnings over a long period of time would be 800, this also includes all losses and gains over the period of time.

1. What is the good measure of the risk involved in a venture of this kind? Compute this measure.

Ans. P(x = - 2000) + P(x = - 1000) = 0.2.

So the risk associated with this venture is 20%.

**Set+2\_Normal+Distribution+Functions+of+random+variables**

**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?

Ans.

Given:-

*μ* = 45 min

*σ* = 8 min

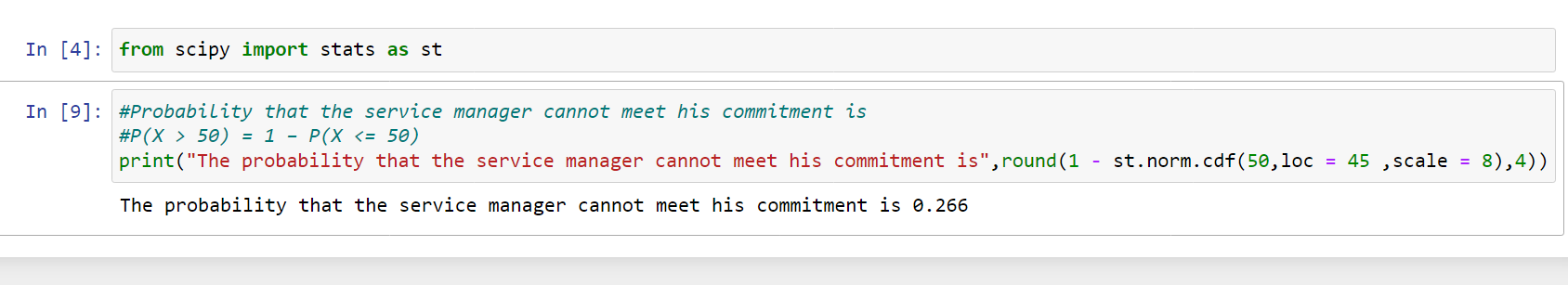
The service manager plans to have work begin on the transmission of a customer’s car 10 mins after the car is dropped off. But, the customer is told that the car will be ready within 1 hrs (60 mins) from drop-of………….*condition 1*

Therefore, the time left to complete work is 60 – 10 = 50 mins.

According to given condition 1,

Probability that the service manager cannot meet his commitment is

P(X > 50) = 1 – P(X <= 50) == > 0.266



1. 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.
2. More employees at the processing center are older than 44 than between 38 and 44.
3. A training program for employees under the age of 30 at the center would be expected to attract about 36 employees.

Ans.

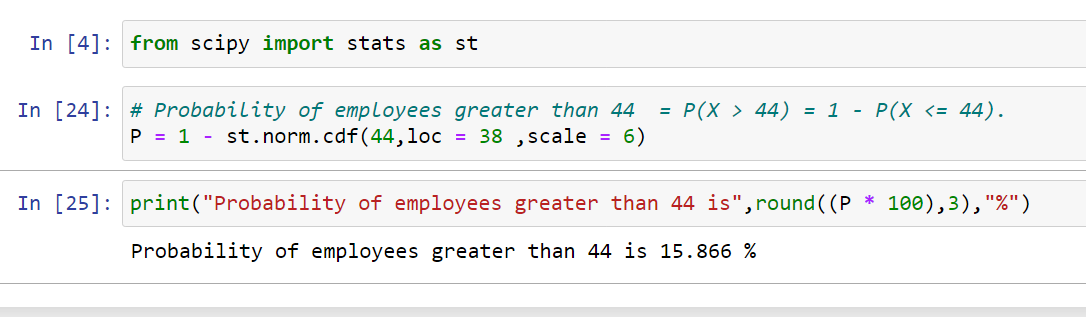
Given:-

*μ* = 38

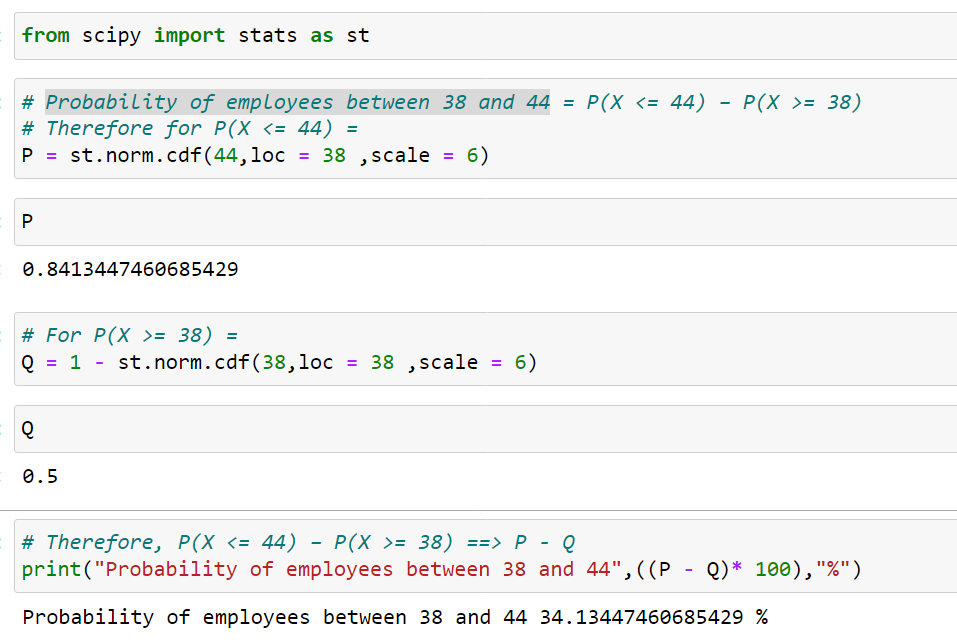
*σ* =6

According to given (a) condition, More employees at the processing center are older than 44 than between 38 and 44.

Probability of employees greater than 44 = P(X > 44).



Probability of employees between 38 and 44 = P(X <= 44) – P(X >= 38)

Here,

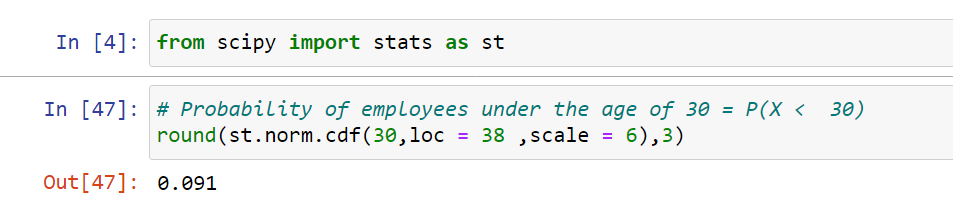
Probability of employees greater than 44 is 15.866 %

Probability of employees between 38 and 44 is 34.13447460685429 %

So, the statement (a) More employees at the processing center are older than 44 than between 38 and 44 is **TRUE.**

According to given (b) condition, A training program for employees under the age of 30 at the center would be expected to attract about 36 employees.

Probability of employees under the age of 30 = P(X < 30)



400 employees are there………..…………………..…………………*given*

And Probability of employees under the age of 30 = P(X < 30) = 0.091

So, the number of employees with probability of employees under the age of 30 = 0.091\*400 = 36.48 ~ 36

Therefore, the statement (b) Training program for employees under the age of 30 at the center would be expected to attract about 36 employees is also **TRUE.**

1. If *X1* ~ *N*(μ, σ2) and *X*2 ~ *N*(μ, σ2) 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.

We know that,

If, X ∼ N(µ1, σ1^2 )

Y ∼ N(µ2, σ2^2 ) are two independent random variables

Then,

X + Y ∼ N(µ1 + µ2, σ1^2 + σ2^2 )

X − Y ∼ N(µ1 − µ2, σ1^2 + σ2^2 )

Similarly,

Z = aX + bY , where X and Y are as defined above, i.e Z is linear combination of X and Y, then Z ∼ N(aµ1 + bµ2, a^2σ1^2 + b^2σ2^2 ).

Therefore in the question

2X1 ~ N(2 µ,4 σ^2)

And X1+X2 ~ N(µ + µ, σ^2 + σ^2 ) ~ N(2µ, 2σ^2 ) 2X1-(X1+X2)

= N( 4µ,6 σ^2).

1. Let X ~ N(100, 202). 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.
2. 90.5, 105.9
3. 80.2, 119.8
4. 22, 78
5. 48.5, 151.5
6. 90.1, 109.9

Ans.

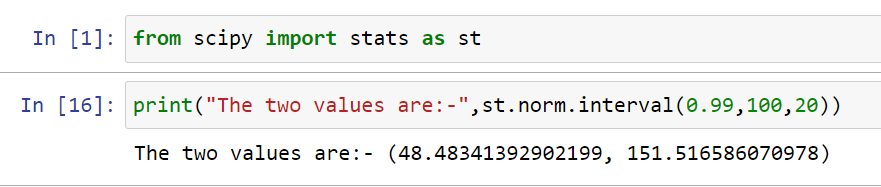
The two values, *a* and *b* symmetric about the mean, such that the probability of the random variable taking a value between them is 0.99………….…..*Given*

Also,

Mean - 100,

St Deviation - 20…………………………………………………………. *Given*

Therefore,



1. Consider a company that has two different divisions. The annual profits from the two divisions are independent and have distributions Profit1 ~ N(5, 32) and Profit2 ~ N(7, 42) 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
2. Specify a Rupee range (centered on the mean) such that it contains 95% probability for the annual profit of the company.
3. Specify the 5th percentile of profit (in Rupees) for the company
4. Which of the two divisions has a larger probability of making a loss in a given year?

Ans.

Given.

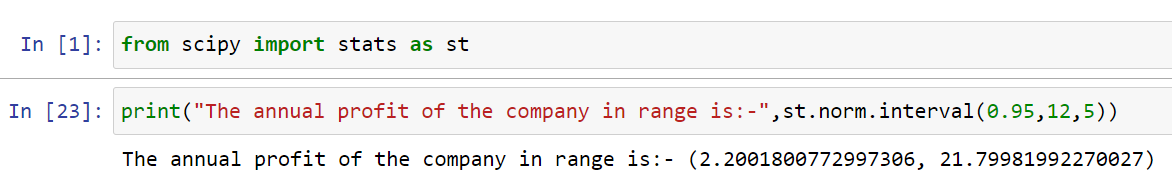
Profit1 ~ N(5, 32)

Profit2 ~ N(7, 42)

Thus company’s profit :-

P ~ N(5+7 , 3^2 + 4^2) = N(12 , 25) = N(12,5^2)

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



$1 = Rs. 45…………………………………………………………. *Given*

So, (2.2 \* 45 = 99 , 21.8 \* 45 = 981) is rupee range such that it contains 95% probability for the annual profit of the company.

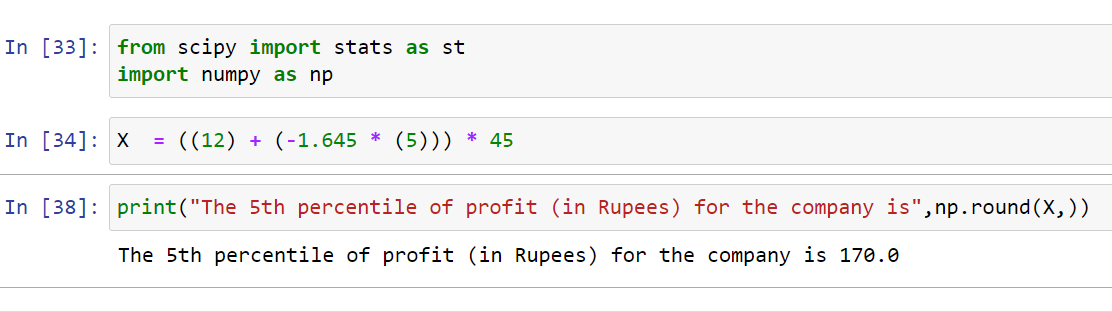
B) Specify the 5th percentile of profit (in Rupees) for the company.

To compute 5th Percentile,

The formula is X=μ + Zσ

Where in from z table - - > 5 percentile = -1.645.

X = (12) + (-1.645 \* (5))



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

Divisions One,

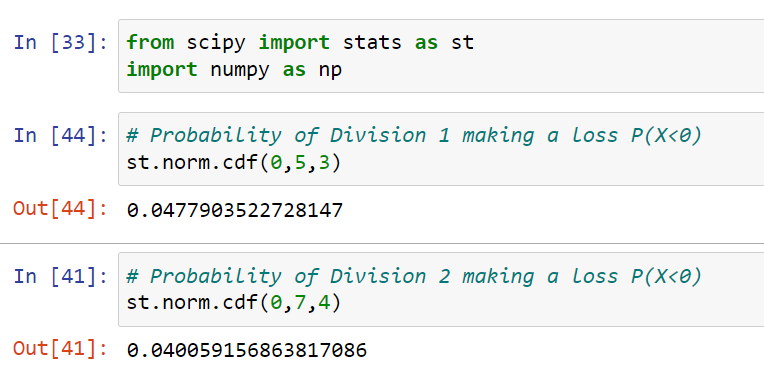
Mean - 5,

St Deviation - 3.

Divisions Two,

Mean - 7,

St Deviation - 4.



**Set - 3**

**Topics: Confidence Intervals**

1. For each of the following statements, indicate whether it is True/False. If false, explain why.
2. 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**.

1. 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 item that including those things which respond to questions.

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

Ans. **True.**

1. *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:
2. The population 🡪9000.
3. The parameter of interest 🡪7.5.
4. The sampling frame 🡪 All the readers of PC Magazine.
5. The sample size 🡪 225
6. The sampling design 🡪 Voluntary voting
7. Any potential sources of bias or other problems with

the survey or sample 🡪 People only who are extremely happy or

unhappy might be included in the sample which

makes the output unreasonable.

1. For each of the following statements, indicate whether it is True/False. If false, explain why.
2. 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.**

1. 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.**

Because we can’t get into such conclusion with the provided data.

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

Ans. **False.**

It can’t be applied for any kind of confidence interval.

1. What are the chances that ?
2. ¼
3. ½
4. ¾
5. 1

Ans. **D -** (Mean of a sample means is equal to population mean).

1. 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.
2. If the sample were based on 2,000 users, could Microsoft conclude that Mozilla has a less than 5% share of the market?
3. 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.

Let p = population proportion share of the market by Mozilla.

1) If the sample were based on 2,000 users, could Microsoft conclude that Mozilla has a less than 5% share of the market.

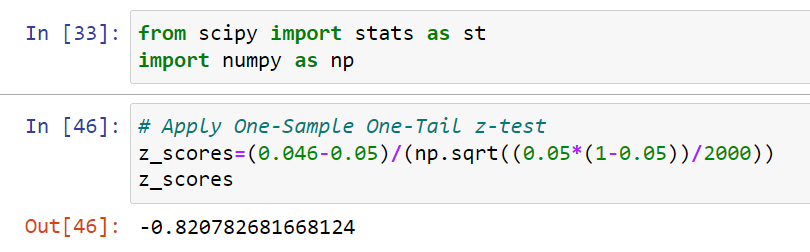
So, Null Hypothesis p >= 5% {means that Mozilla has more than or equal to 5% share of the market}

Alternate Hypothesis p < 5% {means that Mozilla has a less than 5% share of the market}

Given:-

Sample proportion of the share of the market grabbed by Mozilla in 2004 = 4.6%

n = sample of users = 2,000

Since the value of our test statistics is more than the critical value of z.

So we fail to reject our null hypothesi*s*.

Therefore, we conclude that Mozilla has more than or equal to 5% share of the market.

2) 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.

Given:-

WebSideStory claims that its sample includes all the daily Internet users.

So, this means that the 4.6% share of the market represents the whole population.

Therefore, we can conclude that Mozilla has a less than 5% share of the market.

1. 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?
2. All shipments are between 205 and 295 books.

Ans. Not correct.

The 95% confidence interval for the size of shipment was books. All shipments are not between 205 and 295 books. Because due to 95% confidence interval for the size of shipment not all shipments.

1. 95% of shipments are between 205 and 295 books.

Ans. Correct.

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

Ans. Correct.

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

Ans. Not correct.

The interval does not describe the mean of another sample.

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

Ans. Not correct.

The interval is not corresponding to 95% C.I.

1. Which is shorter: a 95% *z*-interval or a 95% *t*-interval for *μ* if we know that σ = s?
2. The z-interval is shorter
3. The t-interval is shorter
4. Both are equal
5. We cannot say

Ans. The z-interval is shorter

Because the critical value of t is greater than the critical value of z.

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.

1. 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)?
2. 600
3. 400
4. 550
5. 1000

Ans :

Consider, p = 0.5

Z-scores for 95% confidence interval = 1.960

N = ((1.960/0.4) ^2 \*0.5\* (1-0.5))\*100

N = 600.25

N ≈ 600.25

Hence , A is correct.

1. Suppose we want the above margin of error to be based on a 98% confidence level. What sample size (minimum) must we now use?
2. 1000
3. 757
4. 848
5. 543

Ans. Consider, p = 0.5

Z-scores for 98% confidence interval = 2.326

N = ((2.326/0.4) ^2 \*0.5\* (1-0.5))\*100

N = 848.36

N ≈ 848

Hence , C is correct.

**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 …
2. Are nearly normal?

Ans. C

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

Ans. D

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

Ans. A

1. Have outliers on both sides of the center?

Ans. B



1. 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.

1. 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.**

Because the individual values do not need to be normally distributed as

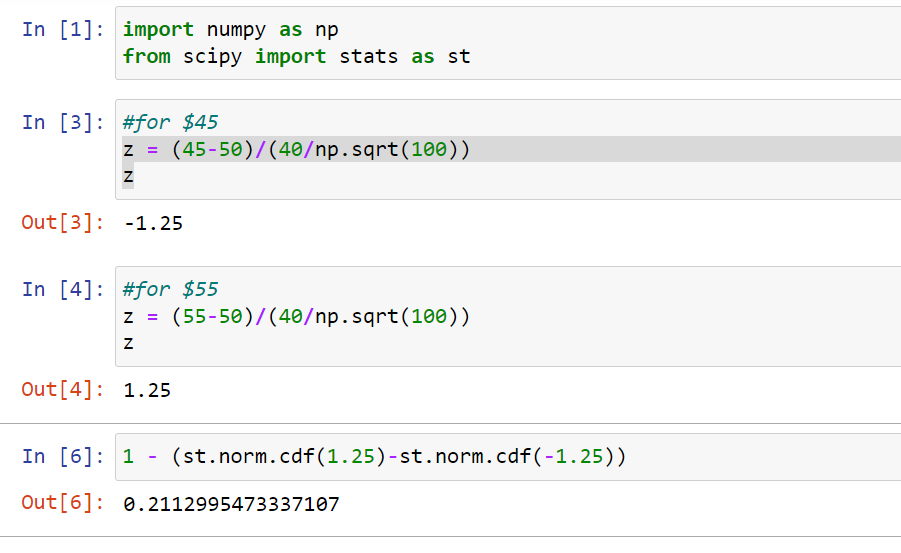
long as the sample size condition is satisfied.

1. The standard error of the daily average SE() = 1.

Ans. **True**

1. 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?
2. 1.25%
3. 2.5%
4. 10.55%
5. 21.1%
6. 50%

Ans.



Therefore, 0.211 = 21%

Option D is correct.

1. 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.
2. 144
3. 150
4. 196
5. 250
6. Not enough information

Ans. For 5%,

Z is +/- 1.96

So, 1.96 = (5)\*sqrt(n)/40

Sqrt(n) = 15.68

n = 245.86.

Option D is correct.

1. 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?
2. The standard deviation of the scores within any sample will be 120.
3. The standard deviation of the mean of across several samples will be 120.
4. The mean score in any sample will be 720.
5. The average of the mean across several samples will be 720.
6. The standard deviation of the mean across several samples will be 0.60

Ans. Option D is correct because data is distributed with kurtosis mean. Hence, More information at the center and Lesser information at the tail. This mean there is higher chance that average of mean of randomly chosen will be 720 that fall in between 650 and 790 at the center.