# Basic Statistics (Module – 4 (Part – 2))

1. For each of the following statements, indicate whether it is True/False. If false, explain why.
   1. 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: The representation of the survey results should have a sample size. The sample size must be a fixed percentage of the total population size of the survey

* 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 refers to a list of an item which responds to the question and not the ones which do not respond to the questions

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

Ans:-

True: The larger conveys a more accurate impression of the population as larger surveys involve large sample size which reduces the chances of error.

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

Ans) p=x/n=225/9000=0.025

1. The parameter of interest

Ans) sample size,average,scale

1. The sampling frame

Ans)9000

1. The sample size

Ans) 225

1. The sampling design
2. Any potential sources of bias or other problems with the survey or sample

# 

Q3) Suppose we want to estimate the average weight of an adult male in Mexico. We draw a random sample of 2,000 men from a population of 3,000,000 men and weigh them. We find that the average person in our sample weighs 200 pounds, and the standard deviation of the sample is 30 pounds. Calculate 94%,98%,96% confidence interval?

Ans:-

## Part 1:-

**Make sure the sample is normal and is a simple random sample.** They are looking at average weight and we can assume normality from the setup, as the judgement for normality in a sample mean problem is the sample size. Is n**(the sample size) > 30**?   
2,000> 30, so it is normal

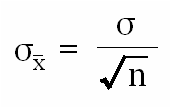
## Part 2:-

**Identify a sample statistic**. Since we are trying to estimate the mean weight in the population, we choose the mean weight in our sample (200) as the sample statistic.

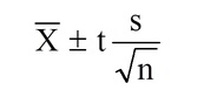
## Part 3:-

**Select a confidence level.** In this case, the confidence level is defined for us in the problem. We are working with a 94%,98%,96% confidence level

Standard Error Formula:



Margin of Error Formula:



## Part 4:

## **CONFIDENCE LEVEL = 94%**

**Find the standard error:**

=30/(square root of 2000) = 30/44.72 = 0.67

**Find critical value**.

* Compute alpha (α): α = 1 - (confidence level / 100)

= 1 - (94 / 100) = .06

* Find the critical probability (p\*): p\* = 1 - α/2 = 1 - 0.06/2 = 0.97
* Find the [degrees of freedom](http://stattrek.com/Help/Glossary.aspx?Target=Degrees%20of%20freedom) (df): df = n - 1 = 2000 - 1 = 1999
* The critical value is the t score having 1999 degrees of freedom and a [cumulative probability](http://stattrek.com/Help/Glossary.aspx?Target=Cumulative%20probability) equal to 0.97. From the [t Distribution Calculator](http://stattrek.com/Tables/T.aspx), we find that the critical value is 1.882.

**Compute margin of error (ME)**: ME = critical value \* standard error = 1.882 \* 0.67 = 1.26

Specify the confidence interval. The range of the confidence interval is defined by the sample statistic + margin of error. And the uncertainty is denoted by the confidence level. Therefore, this 94% confidence interval is 200 + 1.26.

## **CONFIDENCE LEVEL = 98%**

**Find the standard error:**

=30/(square root of 2000) = 30/44.72 = 0.67

**Find critical value**.

* Compute alpha (α): α = 1 - (confidence level / 100)

= 1 - (98 / 100) = .02

* Find the critical probability (p\*): p\* = 1 - α/2 = 1 - 0.02/2 = 0.99
* Find the [degrees of freedom](http://stattrek.com/Help/Glossary.aspx?Target=Degrees%20of%20freedom) (df): df = n - 1 = 2000 - 1 = 1999
* The critical value is the t score having 1999 degrees of freedom and a [cumulative probability](http://stattrek.com/Help/Glossary.aspx?Target=Cumulative%20probability) equal to 0.99. From the [t Distribution Calculator](http://stattrek.com/Tables/T.aspx), we find that the critical value is 2.328.

**Compute margin of error (ME)**: ME = critical value \* standard error = 2.328\* 0.67 = 1.56

Specify the confidence interval. The range of the confidence interval is defined by the sample statistic + margin of error. And the uncertainty is denoted by the confidence level. Therefore, this 98% confidence interval is 200 + 1.56.

## **CONFIDENCE LEVEL = 96%**

**Find the standard error:**

=30/(square root of 2000) = 30/44.72 = 0.67

**Find critical value**.

* Compute alpha (α): α = 1 - (confidence level / 100)

= 1 - (96 / 100) = .04

* Find the critical probability (p\*): p\* = 1 - α/2 = 1 - 0.04/2 = 0.98
* Find the [degrees of freedom](http://stattrek.com/Help/Glossary.aspx?Target=Degrees%20of%20freedom) (df): df = n - 1 = 2000 - 1 = 1999
* The critical value is the t score having 1999 degrees of freedom and a [cumulative probability](http://stattrek.com/Help/Glossary.aspx?Target=Cumulative%20probability) equal to 0.98. From the [t Distribution Calculator](http://stattrek.com/Tables/T.aspx), we find that the critical value is 2.055.

**Compute margin of error (ME)**: ME = critical value \* standard error = 2.055\* 0.67 = 1.37

Specify the confidence interval. The range of the confidence interval is defined by the sample statistic + margin of error. And the uncertainty is denoted by the confidence level. Therefore, this 96% confidence interval is 200 + 1.37.

1. What are the chances that

*X*  ** ?

1. ¼
2. ½
3. ¾
4. 1

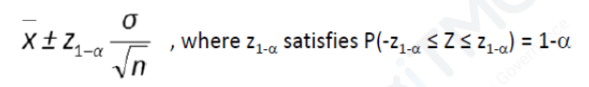
Ans:-

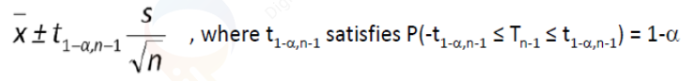
B. ½ ((( in standared normal distribution half of the samples are higher than population mean and another half is lower than population mean or population mean we treating such as mean of the all the sample means)))

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.
3. 95% of shipments are between 205 and 295 books.
4. The procedure that produced this interval generates ranges that hold the population mean for 95% of samples.
5. If we get another sample, then we can be 95% sure that the mean of this second sample is between 205 and 295.
6. 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?
8. The z-interval is shorter
9. The t-interval is shorter
10. Both are equal
11. We cannot say

Ans:-

Equation for confidence interval for *z*-interval & *t*-interval





From the equation we can see that the length of the interval is determined by the margin of error (**margin of error (ME)**: ME = critical value \* standard) and the value of ME depends on critical value of of z & t . As critical value increases correspondingly ME and thereby interval size would also increase. And 95 % confidence interval for mean is shorter for z-interval because t-critical is greater than z-critical value.

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)?

A. 600

B. 400

C. 550

D. 1000

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?

A. 1000

B. 757

C. 848

D. 543

1. Examine the following normal Quantile plots carefully. Which of these plots indicates that the data?
2. Are nearly normal?
3. Have a bimodal distribution? (One way to recognize a bimodal shape is a “gap” in the

spacing of adjacent data values.)

1. Are skewed (i.e. not symmetric)?
2. Have outliers on both sides of the center?



Ans:-

1. C

2. D ( gap present )

3. A ( severe right skewness)

4. A

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:-

TRUE. In this case, at least 30 sample packages must be selected and weighed everyday. Based on the central limit theorem, the sampling distribution of the sample mean approach normal distribution as the sample size become bigger (over 30)

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

Ans:-

TRUE. Standard error equal to standard deviation divided by square root of sample size = 5/sqrt(25) =1

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:-

The SEM is sd/sqrt(n)=120/sqrt(40000)=0.6

A) SD will not be 120 of scores in any one sample, especially since we don't know the sample size.

B) SD of mean across several samples will also not be 120. It will be less; indeed, probably about 0.6

C) The mean score in any sample will be 720. Maybe, but no reason it couldn't be less or more.

D) The average of the mean across several samples will be 720. This is certainly possible, but it requires the mean of all samples that sample size, which would be the case

E) The SEM will be 0.60. This is likely, given the sample size, which even with a lot of skewness will tend towards normality given the sample size. I would use this in calculations. The mean would have an expected value of 720, but in calculations, the SEM is 0.6.

**Hints:**

1. Business Problem
   1. Objective
   2. Constraints (if any)
2. For each assignment the solution should be submitted in the below format
3. Research and Perform all possible steps for obtaining solution
4. For Basic Statistics explanation of the solutions should be documented in black and white along with the codes.

One must follow these guidelines as well:

* 1. Be thorough with the concepts of Probability, Central Limit Theorem and Perform the calculation stepwise
  2. For True/False Questions, explanation is must.

R & Python code for Univariate Analysis (histogram, box plot, bar plots etc.) for data distribution to be attached Ans:-

1. All the codes (executable programs) should execute without errors