Assignment -02

# #Set+1

2.



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.
2. What can we say about the skewness of this dataset?
3. 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. The interquartile range (IQR) measures the spread of the middle part of your data set.Inter quartile range(IQR)=75quartile -25quartile =Q3-Q1=13-5=8
2. The skewness of the our data set is the positively SKEWED.
3. If data point 25 turns to the 2.5 we may get the NO SKEW DATA because of no outliers

3.



Answer the following three questions based on the histogram above.

1. Where would the mode of this dataset lie?
2. Comment on the skewness of the dataset.
3. 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. Mode in the above dataset lie approximately at near to the 5
2. Frequency highly lied between the 5-10 that’s why it is the positively skewed data
3. We can say that question 2&3 are complement to each other because both data is positively skewed and both the data have the one outlier that is the 25
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: one in 200 long-distance telephone calls is misdirected

Then probability=p=1/200

Not misdirected=q=1-1/200=199/200

For five misdirected n=5

P(x) = ⁿCₓ pˣ qⁿ⁻ˣ P(x) = (nCx)(p^x)(q^n-x)

# nCr = n! / r! \* (n - r)!

P(1) = (5C1)(1/200)^1(199/200)^5-1 P(1)=0.0245037

Conclusion : there is significantly 2% chance that at least one in the five calls Is misdirected.

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: (i): Here the highest probability for 2000

1. Is the venture likely to be successful? Explain

Ans: (ii): Yes, because the total earnings of the venture is positive in value i.e 800 and highest probability of earning is 2000.

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

Ans:

(iii)averageearnings=x\*p(x)

averageearnings=

0.1\*(2000) + 0.1\*(1000)+0.2(0)+0.2\*(1000)+0.3(2000)+0.1(3000)

=800

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

Ans: (iv): stats.norm.ppf(0.80)= 0.8416212335729143

# #set+2

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?
2. 0.3875
3. 0.2676
4. 0.5
5. 0.6987

**Ans.** B=(0.2676)

By using Z score value is 60-55/8=0.625 and 1-stats.norm.cdf (0.625).

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.

**Ans.** False, because people between 38-44 and more than 44 at age group are 137 and 63 out of 400 respectively.

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

**Ans.**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**. Normal distribution. Parameters are same.

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.** D-48.5, 151.5

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.

Ans. Range: 99.008-980.992

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

Ans. 170 millions

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

Ans. 170 millions

# #set+3

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.

Answer: FALSE

Reason: A sample size of 30 is considered large enough, but that may or may not be adequate.

1. The sampling frame is a list of every item that appears in a survey sample, including those that did not respond to questions.

Answer: TRUE

Reason: The population is generic and the sampling frame is a specific list of all items in the population. Hence the sampling frame includes those that did not respond to questions

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

Answer: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 ----- p(x/n)=225/9000=0.025
3. The parameter of interest-------Average,scale,size
4. The sampling frame---------9000
5. The sample size-----------225
6. The sampling design------0-10
7. Any potential sources of bias or other problems with the survey or sample --- Bias

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

REASON: It doesn't provide a definitive answer about whether fewer than half of all moviegoers purchase concessions

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

Ans:TRUE

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

Ans: B.1/2

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?

**Ans**. p =4.6%

n=2000

t- Statistics=0.821

z score critical value= -1.96 (left tail skewed )

Hence we conclude that t>z. so Mozilla has a less than 5% shares.

1. 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.**  Mozilla has less than 5% shares because 4.6% is mean sample of shares.

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 : INCORRECT

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

Ans : INCORRECT

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 : INCORRECT

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

Ans : INCORRECT

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. 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.

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**. A-600

n=number of employers, Assume ^ P=0.5,q^ =0.5 Margin of Error=0.04 95% confidence interval, the critical value Z= 1.96 ME = Z \*  
n= √p^ q^ ; n= 2 1.96 ∗0.5∗0.5 0.9604

n= 600

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: C-848.

Necessary Sample Size = (Z-score)² \* StdDev\*(1-StdDev) / (margin of error)²

Z= 2.576

0.04 = 2.326 \*

n=√0.5∗0.5   
n=2.3262∗0.5∗0.5 1.3525 = = 845.35

# #set+4

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



i. Answer is C.

Nearly Normal Distribution: In a normal quantile plot, if the data points

closely follow a straight line without any significant deviations or bends, it

suggests that the data is nearly normally distributed.

ii. Answer is B.

Bimodal Distribution: A bimodal distribution will have two distinct peaks or

modes in the plot, indicating that the data has two different groups or sub-

populations.

iii. Answer is A, C and D.

Skewed Distribution: A skewed distribution will have a longer tail on one

side of the plot, suggesting that the data is not symmetric around the center.

iv. Answer is A.

Outliers: Outliers are data points that significantly deviate from the overall

pattern in the plot. If there are outliers on both sides of the center, it indicates

that the data has outliers in both the lower and upper tails.

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.

False : A sampling distribution is a probability distribution of a statistic obtained from a large number of samples drawn from a specific population. In our case the samples contain 25 packages and the larger number of samples contain of each such 25 packages taken into different samples (25+25+25+25…and so on). The mean for one these samples is 22lbs and standard deviation of 5lbs which means each individual package is having a weight varying between + or – 5lbs with respect to mean(22lbs). Hence it is invalid to take a weight of individual packages and confirm that it follows normal distribution before using a normal model for the sampling distribution. The Sample Central Limit Theorem states that the sampling distribution of the samples mean approaches normal distribution as the sample size is large enough.

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

True : As SE(Standard Error) = sample standard deviation / Square root of (number of sample) SE = 5 / (25)^1/2 SE = 1

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:D

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:D

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:

A. The standard deviation of the scores within any sample will be 120. This statement is unlikely to be true. The standard deviation within each sample is determined by the variability of scores present in that specific sample. Since the scores are distributed between 650 and 790 with a long and thin tail towards the higher end, the standard deviation within any sample may vary depending on the composition of that sample. Therefore, the standard deviation within a

sample is not fixed at 120.

B. The standard deviation of the mean across several samples will be 120. This statement is also unlikely to be true. The standard deviation of the mean (standard error) across several samples can be estimated using the formula: SE = SD / √(sample size). However, the sample size is not specified in the question. Without the sample size, it is not possible to calculate the standard deviation of the mean across several samples accurately.

C. The mean score in any sample will be 720. This statement is possible and likely to be true. The population average GMAT score is given as 720. When randomly selecting samples from the population, each sample should have an average (mean) that is centered around the population mean. Therefore, it is likely that the mean score in any sample will be close to 720.

D. The average of the mean across several samples will be 720. This statement is likely to be true. If we calculate the mean of each sample and then average those means across several samples, the resulting value should be close to the population mean. Since each sample is randomly selected from the population, the average of the means across several samples is expected to be around 720.

E. The standard deviation of the mean across several samples will be 0.60. This statement is unlikely to be true. The standard deviation of the mean (standard error) across several samples is determined by the standard deviation of the population and the sample size. In this case, the standard deviation of the population is given as 120, and the sample size is not specified. Without knowing the sample size, it is not possible to calculate the exact standard deviation of the mean across several samples. Based on the analysis above, it is likely that statement C (&quot;The mean score in any sample will be 720&quot;) and statement D (&quot;The average of the mean across several samples will be 720&quot;) are true for randomly chosen samples of aspirants.