Name - Niranjan Nevase
Batch - 5th June Data Science

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.** 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.

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

   III. Larger surveys convey a more accurate impression of the population than smaller surveys.
      - **True**. The larger conveys a more accurate impression of the population as larger surveys involve large sample size which reduces the chances of error.

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 , p = 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)

Name - Niranjan Nevase
Batch - 5th June Data Science

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.
        - **False**. The 95% confidence interval for the number of moviegoers who purchase concessions is 30% to 45%. This means that if we were to repeat the experiment many times, 95% of the time, the true proportion of moviegoers who purchase concessions would fall between 30% and 45%. It does not imply that fewer than half of all moviegoers purchase concessions

    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

    - B. ½
    - 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?

Name - Niranjan Nevase
Batch - 5^{th} June Data Science

To determine if Microsoft can conclude that Mozilla has a less than 5% share of the market based on the sample size of 2,000 users, we can use statistical analysis. In this case, you would typically perform a hypothesis test to determine the likelihood of Mozilla's market share being less than 5%.

The null hypothesis (H0) in this case would be that Mozilla's market share is equal to or greater than 5%, and the alternative hypothesis (Ha) would be that Mozilla's market share is less than 5%.

The sample indicates that Mozilla has a 4.6% share of the market in the sample, which can be considered the sample mean ($\hat{p}$). To test the hypothesis, you can use a z-test for proportions.
Calculating this gives:
$Z \approx -0.183$

-0.183 is much greater (in absolute value) than the critical z-value for a one-tailed test at $\alpha = 0.05$, so it would fail to reject the null hypothesis.

Therefore, based on this sample size of 2,000 users, **Microsoft cannot conclude that Mozilla has a less than 5% share of the market. The sample does not provide enough evidence to support that conclusion.**

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?
- If WebSideStory claims that its sample includes all daily Internet users, this would imply that the sample is a census rather than a sample. In a census, you have data from the entire population, not just a sample. In this case, if WebSideStory has data from all daily Internet users, there's no statistical sampling involved.
- If WebSideStory's claim is accurate and they have data from all daily Internet users, then the 4.6% figure they reported would represent the actual market share of Mozilla among all daily Internet users. In this case, Microsoft can conclude that Mozilla has a 4.6% share of the market based on the information provided by WebSideStory. There would be no need for statistical hypothesis testing because the data represents the entire population, not a sample.

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.

- Ans – **Correct statement is statement C.**
  A 95% confidence interval is an interval estimate of a population parameter. In this case, it's the mean size of the shipments. The interpretation is that if we were to take many samples and build a confidence interval from each sample, then 95% of those intervals would contain the true population mean. This is what option C correctly states.

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**
- This is because the t-distribution has fatter tails than the standard normal distribution, which means that the t-interval will be wider than the z-interval for the same level of confidence and sample size. When σ = s, we can use the standard normal distribution to construct a z-interval, which will be shorter than the t-interval .
- 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

Name - Niranjan Nevase
Batch - 5th June Data Science

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

- **Ans –**
  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 x 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%