1. How is the statistical significance of an insight assessed?
   1. Hypothesis testing is used to find out the statistical significance of the insight. To elaborate, the h1 and h0 are stated, and the p-value is calculated. After calculating the p-value, the null hypothesis is assumed true, and the values are determined. To fine-tune the result, the alpha value, which denotes the significance, is tweaked. If the p-value turns out to be less than the alpha, then the null hypothesis is rejected. This ensures that the result obtained is statistically significant
2. What is the central limit theorem?
   1. The central limit theorem states that if you have a population with mean µ and standard deviation σ and take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be approximately normally distributed.
3. How to explain confidence interval to a non-technical audience?
   1. Confidence interval are a range of values with a lower and an upper bound such that if you were to sample the parameter of interest (sample mean) a large number of times, the 95% confidence interval would contain the true value of this parameter 95% of the time.
4. What is exploratory data analysis
   1. ETA is the process of performing investigation on data to understand the data better
   2. Includes determine pattern, spot abnormalities, test hypotheses, and check if the assumptions are right.
5. What is the probability of throwing two fair dice when the sum is 5 and 8?
   1. There are 4 ways of rolling a 5 (1+4, 4+1, 2+3, 3+2): P(Getting a 5) = 4/36 = 1/9
   2. There are 7 ways of rolling an 8 (1+7, 7+1, 2+6, 6+2, 3+5, 5+3, 4+4): P(Getting an 8) = 7/36 = 0.194
6. State the case where the median is a better measure when compared to the mean
   1. In the case where there are a lot of outliers that can positively or negatively skew data, the median is preferred.
7. What is the meaning of the five-number summary in statistics?
   1. Low extreme (Min)
   2. First quartile (Q1)
   3. Median
   4. Upper quartile (Q3)
   5. High extreme (Max)
8. What are population and sample in inferential statistics and how are they different?
   1. A population is a large volume of observations. The sample is a small portion of that population. We calculate the statistics using the sample and using these sample statistics, we make conclusions about the population.
9. What is skewness?
   1. Skewness measures the lack of symmetry in a data distribution.
10. What is correlation?
    1. Correlation is used to test relations between quantitative variables and categorical variables.
11. What is the difference between descriptive and inferential statistics?
    1. Descriptive statistics: Descriptive statistics is used to summarize a sample set of data like the standard deviation or the mean.
    2. Inferential statistics: Inferential statistics is used to draw conclusions from the test data that are subjected to random variations.
12. What are types of sampling in statistics?
    1. Simple random: Pure random division
    2. Cluster: Population divided into clusters
    3. Stratified: Data divided into unique groups
    4. Systematical: Picks up every ‘n’ member in the data
13. What is covariance and correlation?
    1. Covariance indicates the direction of the linear measure of relationship
14. The standard normal curve has a total area to be under one, and it is symmetric around zero true or false?
    1. True
15. What is Bessel’s Correction?
    1. Bessel’s correction refers to the ‘n-1’ found in several formula, including sample variance and sample standard deviation formulas. It causes the standard deviation to be less biased, thereby, providing more accurate result.
16. What us the relationship between the confidence level and the significance level in statistics?
    1. The significance level is the probability of rejecting the ho when ho is true.
    2. Confidence level = 1 - significance level
17. What is Person’s Correlation Coefficient
    1. PCC is the test statistics that measures the statistical relationship, or association, between two continuous variables.
    2. Variables to be used for PCC must be either in ratio or in an interval
18. What is the benefit of using boxplots?
    1. Box plots allow us to provide a graphical representation of the 5-number summary and can also be used to compare groups of histograms.
19. Briefly explain the procedure to measure the length of all sharks in the worlds.
    1. Define the confidence level (usually around 95%)
    2. Use sample sharks to measure
    3. Calculate the mean and standard deviation of the length
    4. Determine t-statistics values
    5. Determine the confidence interval in which the mean length lies
20. How does the width of the confidence interval change with length?
    1. The confidence level increases, the width also increases.
21. What is the law of large numbers in statistics?
    1. It states that the increase in the number of trials performed will cause a positive proportional increase in the average of the results becoming the expected value.
22. What is the meaning of degrees of freedom (DF) in the statistics?
23. What is the meaning of sensitivity in statistics?
    1. Is used to determine the accuracy of classifier.
    2. Sensitivity = True Positive/ (True Positive + False negative)
24. How to explain p-value to non-statistician
    1. The p-value is a measurement of the probability of obtaining the results in the data assuming that null hypothesis is true
    2. A p-value of 0.05 means that if we repeat the experiment many times, we won’t be wrong more than 5% of the time
25. Explain the statistical background behind power
    1. Power is the probability of rejecting the null hypothesis when it is false. The higher the power, the higher the probability of detecting a genuine effect
26. Hypothesis Testing explain to a non-technical (explained by an example may be better)
    1. We want to test a drug can help people lose weight
       1. Null hypothesis:
          1. The people take the drug do not lose weight
       2. Alternative Hypothesis:
          1. The people take the drug do lose weight
    2. We collect weight loss data for sample of 10 people who take the drug for over one month
       1. WeightLoss Sample Mean = 2 kg
       2. Sample Standard Deviation = 1kg
    3. Does it prove that the drug does reduce weight? We are not sure the drug whether take the effect, so an experiment should be set up to test it.
    4. A good question to ask ourselves is- Assuming that null hypothesis is true, what is the probability of observing a sample mean of 2 kg or more extreme than 2 kg?
    5. The probability of observing a sample mean of 2 kg or more extreme than 2kg given there is no effect of the drug is the p-value
    6. Then, we want to compare the p-value to a threshold as the significance level(α).
    7. Why don’t we say accept the null hypothesis
       1. The best intuitive example of this is using trial courts. In a trial court, the null hypothesis is that the accused is not guilty. If we are not able to disprove the null hypotheses the judge does not say that the accused has not committed the crime. The judge only says that based on the given evidence, we are not able to convict the accused.