

Statistics Worksheet 6

1. Which of the following can be considered as random variable?

d) All of the mentioned

2. Which of the following random variable that take on only a countable number of possibilities?

a) Discrete

3. Which of the following function is associated with a continuous random variable?

a) pdf

4. The expected value or _____ of a random variable is the center of its distribution.

c) mean

5. Which of the following of a random variable is not a measure of spread?

a) variance

6. The _____ of the Chi-squared distribution is twice the degrees of freedom.

d) none of the mentioned

7. The beta distribution is the default prior for parameters between _____

a) 0 and 10

8. Which of the following tool is used for constructing confidence intervals and calculating standard errors for difficult statistics?

b) bootstrap

9. Data that summarize all observations in a category are called _____ data.

b) summarized

10. What is the difference between a boxplot and histogram?

Histogram - A histogram is a type of bar chart that graphically displays the frequencies of a data set. Similar to a bar chart, a histogram plots the frequency, or raw count, on the Y-axis (vertical) and the variable being measured on the X-axis (horizontal).

The only difference between a histogram and a bar chart is that a histogram displays frequencies for a group of data, rather than an individual data point; therefore, no spaces are present between the bars. Typically, a histogram groups data into small chunks (four to eight values per bar on the horizontal axis), unless the range of data is so great that it is easier to identify general distribution trends with larger groupings.

Box plot - A box plot, also called a box-and-whisker plot, is a chart that graphically represents the five most important descriptive values for a data set. These values include the minimum value, the first quartile, the median, the third quartile, and the maximum value. When graphing this five-number summary, only the horizontal axis displays values. Within the quadrant, a vertical line is placed above each of the summary numbers. A box is drawn around the middle three lines (first quartile, median, and third quartile) and two lines are drawn from the box's edges to the two endpoints (minimum and maximum).

11. How to select metrics?

The key point is to choose metrics that clearly indicate where you are now in relation to your goals. Good metrics can be improved. Good metrics measure progress, which means there needs to be room for improvement. For example, reducing churn by 0.8% or increasing your activation rate by 3%.

12. How do you assess the statistical significance of an insight?

Steps in Testing for Statistical Significance

- State the Research Hypothesis
- State the Null Hypothesis
- Select a probability of error level (alpha level)
- Select and compute the test for statistical significance
- Interpret the results

13. Give examples of data that does not have a Gaussian distribution, nor log-normal.

- Allocation of wealth among individuals
- Values of oil reserves among oil fields (many small ones, a small number of large ones)

14. Give an example where the median is a better measure than the mean.

Mean (or average) and median are statistical terms that have a somewhat similar role in terms of understanding the central tendency of a set of statistical scores. While an average has traditionally been a popular measure of a mid-point in a sample, it has the disadvantage of being affected by any single value being too high or too low compared to the rest of the sample. This is why a median is sometimes taken as a better measure of a mid point.

15. What is the Likelihood ?

The term Likelihood refers to the process of determining the best data distribution given a specific situation in the data.

However, when you calculate the likelihood, you're attempting to determine whether the parameters in a model can be trusted based on the sample data you have observed.