WORKSHEET - 6 STATISTICS

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

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

a) The outcome from the roll of a die

b) The outcome of flip of a coin

c) The outcome of exam

d) All of the mentioned

Answer: d) All of the mentioned

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

a) Discrete

b) Non Discrete

c) Continuous

d) All of the mentioned

Answer: a) Discrete

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

b) pmv

c) pmf

d) all of the mentioned

Answer: a) pdf

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

b) median

c) mean

d) bayesian inference

Answer: c) mean

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

b) standard deviation

c) empirical mean

d) all of the mentioned

Answer: c) empirical mean

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

b) standard deviation

c) mode

d) none of the mentioned

Answer: a) variance

7. The beta distribution is the default prior for parameters between _____ a) 0 and 10

b) 1 and 2

c) 0 and 1

d) None of the mentioned

Answer: c) 0 and 1

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

a) baggyer

b) bootstrap

c) jacknife

d) none of the mentioned

Answer: b) bootstrap

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

a) frequency

b) summarized

c) raw

d) none of the mentioned

Answer: b) summarized

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

Answer: Histograms are preferred to determine the underlying probability distribution of a data. Box plots on the other hand are more useful when comparing between several data sets. They are less detailed than histograms and take up less space.

Although histograms are better in displaying the distribution of data, you can use a box plot to tell if the distribution is symmetric or skewed. In a symmetric distribution, the mean and median are nearly the same, and the two whiskers has almost the same length.

Histograms and box plots are very similar in that they both help to visualize and describe numeric data. Although histograms are better in determining the underlying distribution of the data, box plots allow you to compare multiple data sets better than histograms as they are less detailed and take up less space.

11. How to select metrics?

Answer: Metrics are measures of quantitative assessment commonly used for assessing, comparing, and tracking performance or production. Generally, a group of metrics will typically be used to build a dashboard that management or analysts review on a regular basis to maintain performance assessments, opinions, and business strategies.

Choosing Metrics

Every business executive, analyst, portfolio manager, and the project manager has a range of data sources available to them for building and structuring their own metric analysis. This can potentially make it difficult to choose the best metrics needed for important assessments and evaluations. Generally, managers seek to build a dashboard of what has come to be known as key performance indicators (KPIs).

In order to establish a useful metric, a manager must first assess its goals. From there, it is important to find the best outputs that measure the activities related to these goals. A final step is also setting goals and targets for KPI metrics that are integrated with business decisions.

Academics and corporate researchers have defined many industry metrics and methods that can help shape the building of KPIs and other metric dashboards. An entire decision analysis method called applied information economics was developed by Douglas Hubbard for analyzing metrics in a variety of business applications. Other popular decision analysis methods include cost-benefit analysis, forecasting, and Monte Carlo simulation.

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

Answer: Statistical significance can be accessed using hypothesis testing:

- Stating a null hypothesis which is usually the opposite of what we wish to test (classifiers A and B perform equivalently, Treatment A is equal of treatment B)
- Then, we choose a suitable statistical test and statistics used to reject the null hypothesis
- Also, we choose a critical region for the statistics to lie in that is extreme enough for the null hypothesis to be rejected (p-value)
- We calculate the observed test statistics from the data and check whether it lies in the critical region Common tests:
- One sample Z test ○ Two-sample Z test ○ One sample t-test ○ paired t-test ○ Two sample pooled equal variances t-test ○ Two sample un-pooled unequal variances t-test and unequal sample sizes (Welch's t-test) ○ Chi-squared test for variances ○ Chi-squared test for goodness of fit ○ ANOVA (for instance: are the two regression models equals? F-test) ○ Regression F-test

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

Answer: Some of the examples are as follows:

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

- Life table is example of exponential distribution
- Wind speed is Weibull distribution
- Surgery patient's stay in hospital is gamma distribution
- Another example is the location of the centers of raindrop ripples on a pond; they are not uniformly spaced in (say) the east-west direction, but they are uniformly distributed.
- A histogram of the stars' right ascensions or declinations (the azimuthal and elevation angles used in astronomy) looks a lot like a step function, i.e., flat within the image boundaries. The positions are not uniformly spaced, but they are distributed in the same way as a uniformly distributed random variable for any size postage stamp, including the entire image.
- The simplest example is the distribution of numbers that show up on the top of a fair die after a large number of throws. Each number from 1 to 6 will occur with approximately equal frequency. Increasing the number of throws will not tend to produce a bell-shaped histogram, in fact the fractional occurrence will approach a constant 1/6 over the possible numbers.

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

Answer: Mean is simply another term for Average. It takes all of the numbers in the dataset, adds them together, and divides them by the total number of entries. Median, on the other hand, is the 50% point in the data, regardless of the rest of the data. For example, if you have the following data:

1, 1, 1, 1, 1, 1, 2, 2, 4

The median is just “1” since that is the middle number in the dataset, while the mean average is 1.56. For a lot of analysis, the mean is very useful. Indeed, if you’re trying to understand data that falls under a normal curve, the mean can tell you a lot of information, because it helps remove some statistical noise from the data and gives you an overall average score for the group.

But the mean is far too often overused, because when it comes to collecting data, it’s not uncommon to find that there are extreme scores that may be altering the final results of your analysis.

Example of When Median is More Useful

Let’s say you run a customer satisfaction survey with a sample of 9 and rate their overall satisfaction scores on a scale of 1 to 10. You get an average of 5.22. You know that in general, you tend to retain customers with a score over 3, so you’re satisfied, because this indicates that you’re still above where you want to be. But then, suddenly, you lose 6 of those 9 customers. You go back to look at your data, and you find these scores:

1, 3, 3, 3, 3, 5, 9, 10, 10

The median of this group is a 3, indicating that at least half of your customers or more were unhappy. The scores became lopsided because of the unexpected 10’s, and you missed out on an important part of your data – the midpoint that indicated that as many as half of your customers or more were dissatisfied with your company.

Median can play a major role in things like income level research as well, because a few millionaires may make it look like the socio-economic status of your sample is higher than it really is.

Whenever a graph falls on a normal distribution, using the mean is a good choice. But if your data has extreme scores (such as the difference between a millionaire and someone making 30,000 a year), you will need to look at median, because you'll find a much more representative number for your sample.

15. What is the Likelihood?

Answer: In statistics, the likelihood function (often simply called the likelihood) measures the goodness of fit of a statistical model to a sample of data for given values of the unknown parameters. It is formed from the joint probability distribution of the sample, but viewed and used as a function of the parameters only, thus treating the random variables as fixed at the observed values. The likelihood function describes a hypersurface whose peak, if it exists, represents the combination of model parameter values that maximize the probability of drawing the sample obtained. The procedure for obtaining these arguments of the maximum of the likelihood function is known as maximum likelihood estimation, which for computational convenience is usually done using the natural logarithm of the likelihood, known as the log-likelihood function. Additionally, the shape and curvature of the likelihood surface represent information about the stability of the estimates, which is why the likelihood function is often plotted as part of a statistical analysis.