STATISTICS WORKSHEET- 6

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
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
3. Which of the following function is associated with a continuous random variable?
   a) pdf
   b) pmv
   c) pmf
   d) all of the mentioned
4. The expected value or_____of a random variable is the center of its distribution.
   a) mode
   b) median
   c) mean
   d) bayesian inference
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
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
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
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

9. Data that summarize all observations in a category are called_____data.
   a) frequency
   b) summarized
   c) raw
   d) none of the mentioned

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

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

Ans :- Histograms are a special kind of bar graph that shows a bar for a range of data values instead of a single value. A box plot is a data display that draws a box over a number line to show the interquartile range of the data. The 'whiskers' of a box plot show the least and greatest values in the data set.

Histograms give a good sense of the distribution of a variable. Box plots attempt to do the same thing however, don't give as good of a picture of the distribution of this variable.

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.

Histograms are better in displaying the distribution of data, you can use a box plot to tell if the distribution is symmetric or skewed.

11. How to select metrics?

Ans :- Prioritize objectives, examine which metric consistently predicts their achievement, and identify which activities influence predictors, in that order. And continuously re-evaluate this process to keep up with the times. three key steps to choosing good outcome metrics (sometimes known as KPIs). Although there is some overlap among the three, we can look at them as distinct steps.

All outcomes need to be precisely defined and clearly prioritized. Often, when an organization implements a new initiative, it will state myriad outcomes that it hopes will occur. For example, a new employee wellness program ideally will improve employee mental health, reduce healthcare costs, and increase productivity, which may even boost customer satisfaction and NPS. Such a broad list of outcomes, however, makes it hard to gauge whether the program is actually effective. For example, what if the new wellness program reduces healthcare costs by 1%, but leaves mental health unchanged? What if healthcare costs remain unchanged but productivity increases? Identifying too many potential outcomes with no clear prioritization or criteria will make it impossible to know whether an initiative merits its cost.

When an outcome changes, is it clearly good or bad? This seems straightforward, but in fact things can get quite tricky. Using the wellness and mental health example, having employees access more resources sounds like a good outcome. But there could be a hidden story here. For example, what if the underlying reason utilization went up is that the program caused employees to be even more stretched and stressed out? Clearly in this case, more use of mental health resources is not a desirable outcome. There are underlying workplace issues that also need to be addressed.

Be aware of unintended consequences. A positive score in one area may not tell the full story. For instance, a retailer tests a new product that is well received in the market. A narrowly defined outcome metric might prompt the decision to roll out the product more broadly. However, a more comprehensive metric would look at sales more broadly, perhaps finding indicators that the new product too greatly cannibalizes another offering. In that case, the retailer would be better served by rethinking the new product roll-out. The bottom line: metrics cannot be viewed through a narrow lens; it takes a broad view.

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

Ans:- To assess statistical significance, you would use hypothesis testing. The null hypothesis and alternate hypothesis would be stated first. Second, you'd calculate the p-value, which is the likelihood of getting the test's observed findings if the null hypothesis is true. Finally, you would select the threshold of significance (alpha) and reject the null hypothesis if the p-value is smaller than the alpha — in other words, the result is statistically significant.

13. Give examples of data that doesnot have a Gaussian distribution, nor log-normal.

Ans :- Many random variables have distributions that are asymptotically Gaussian but may be significantly non-Gaussian for small numbers. For example the Poisson Distribution, which describes (among other things) the number of unlikely events occurring after providing a sufficient opportunity for a few events to occur. It is pretty non-Gaussian unless the mean number of events is very large. The mathematical form of the distribution is still Poisson, but a histogram of the number of events after many trials with a large average number of events eventually looks fairly Gaussian.

For me, the best examples come from my field of research (astrophysical data analysis). For example, something that comes up all the time is that we detect stars in astronomical images and solve for their celestial coordinates. My current project uses images about 1.5 degrees on a side and typically detects 60 to 80 thousand stars per image, with the number well modeled as a Poisson Distribution, assuming that the image is not of a star cluster surrounded by mostly empty space. That's about 8 or 9 stars per square arcminute. If we cut out "postage stamps"from the image that are half an arcminute per side, then the mean number of detected stars in them is about 2. If we do that for (say) 1000 postage stamps and make a histogram of the number of detected stars in them, it will not look very Gaussian, but as we increase the size of the postage stamps, it becomes asymptotically Gaussian.

What generally never becomes Gaussian, however, is the Uniform Distribution. 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.

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.

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.

Ans - The two most widely used measures of the "center" of the data are the mean (average) and the median. To calculate the mean weight of
50
people, add the
50
weights together and divide by
50
. To find the median weight of the
50
people, order the data and find the number that splits the data into two equal parts. The median is generally a better measure of the center when there are extreme values or outliers because it is not affected by the precise numerical values of the outliers.

For example... 20 students got selected in a placement... with a salary of 3lac 4lac 4.5lac and so on....one of the student started his start up and assigned himself as a CEO of his company with a salary of 50 lac...if we have to find the performance of the class according to salary ....we will go with two method either mean or median...

When using mean method the ans will be incorrect because of the 1 outlier whose salary is 50lac... In this situation if we use median method we will firt sort the data either in ascending or descending order....the outlier whose salary is 50 lac with be in the last number....in median the middle number having salary of 5lac or 6 lac will be the answer...

15. What is the Likelihood?

Ans :- The likelihood is the probability that a particular outcome is observed when the true value of the parameter is , equivalent to the probability mass on ; it is not a probability density over the parameter . The likelihood, , should not be confused with , which is the posterior probability of given the data .


FLIP ROBO