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?

ANS: d) All of the mentioned

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

ANS: a) Discrete

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

ANS: a) pdf

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

ANS: c) mean

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

ANS: a) variance

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

a) variance

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

ANS: c) 0 and 1

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

ANS: b) bootstrap

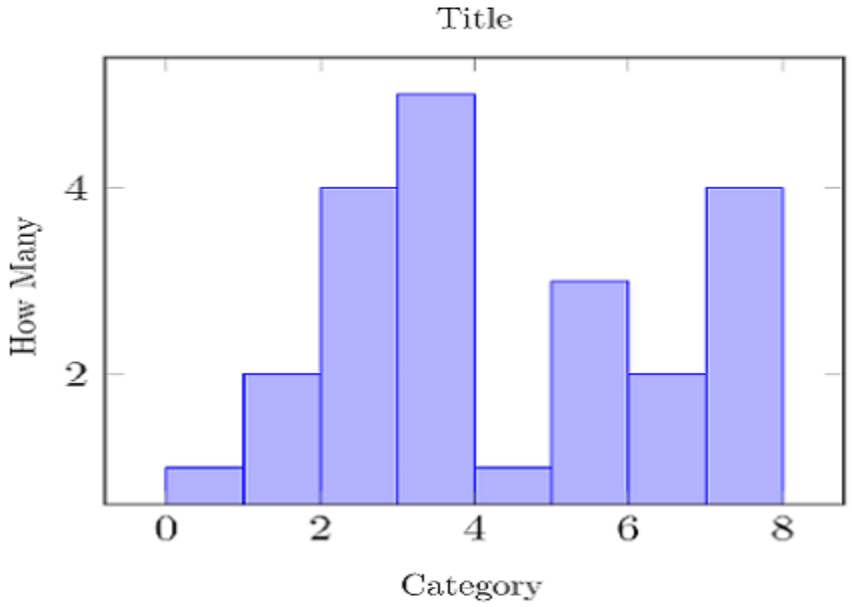
9. Data that summarize all observations in a category are called _____ data.

ANS: b) summarized

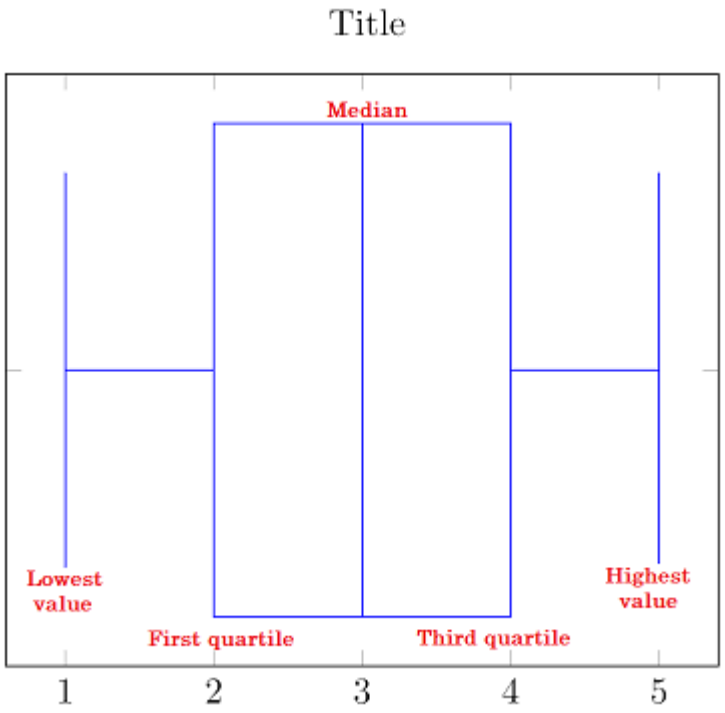
Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

ANS: Histogram: A histogram is a streamlined version of a dot plot, where, instead of dots, we display our information with bars. An example of a histogram is given below:



Box plot: A box plot shows the location of the lowest value, the first quartile (the cut-off mark for the bottom 25% of the data), the median (the middlemost value, or cut-off mark for 50% of the data), the third quartile (the cut-off mark for the top 25% of the data), and the highest value. An example of a box plot is given below:



11. How to select metrics?

ANS: KEY STEPS TO SELECTING EVALUATION METRICS

1. Classification. This algorithm will predict data type from defined data arrays. For example, it may respond with yes/no/not sure.
2. Regression. The algorithm will predict some values. For example, weather forecast for tomorrow.
3. Ranking. The model will predict an order of items.

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?

ANS: 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 unpooled 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 (i.e: is at least one of the predictor useful in predicting the response?)

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

ANS: Exponential distributions do not have a log-normal distribution or a Gaussian distribution. In fact, any type of data that is categorical will not have these distributions as well. Example: Duration of a phone car, time until the next earthquake, etc.

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.

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

ANS: Income is the classic example of when to use the median instead of the mean because its distribution tends to be skewed. The median indicates that half of all incomes fall below 27581, and half are above it. For these data, the mean overestimates where most household incomes fall.

Median is better in the sense that it is a robust statistic. Meaning that it is not influenced by outlier(s). However, when the data are symmetric, theretically, they are the same. Both measures the centre of population.

Median is the middle value in a rank-ordered sequence. Average is the sum of all observation values divided by the number of cases observed.

Medians are not affected by outliers, while averages can swing wildly due to extreme anomalies that are irrelevant to the norms.

The middle (median) remans the same middle value regardless of the size of the highest or the lowest case, which has great effects on the average.

- In a statistically random population sample, the median remains very close to the mode (the single most frequently encountered value), so the median is a superior measure of the norm. The average can bounce all over the place, based on the outliers and the sample distribution.
- If 10 kindergarten kids are visited in a room by one typical professional basketball player, the average height skyrockets, while the median height probably stays almost exactly the same.

15. What is the Likelihood?

ANS: It is the probability that each house y has the price as we observe given the distribution we assumed.
Likelihood, being the outcome of a likelihood function thus defined, describes the plausibility, under a certain statistical model (the null hypothesis in hypothesis testing), of a certain parameter value after observing a particular outcome.

As the earlier statement suggests, we call this term “likelihood” and the approach taken by Fin is that of maximising the likelihood (maximum likelihood). Likelihood refers to the probability of observing the data that has been observed assuming that the data came from a specific scenario