Statistics

Question 1: a) True

Question 2: a) Central Limit Theorem

Question 3: b) Modeling bounded count data

Question 4: d) All of the mentioned

Question 5: c) Poisson

Question 6: b) False

Question 7: b) Hypothesis

Question 8: a) 0

Question 9: c) Outliers cannot conform to the regression relationship

Question 10:

According to me , the normal distribution is an arrangement of a data set in which most values cluster in the middle of the range and the rest becomes smaller symmetrically toward either end.

Height is one simple example of something that follows a normal distribution pattern: Most people are of average height, the numbers of people that are taller and shorter than average are fairly equal and a very small (and still roughly equivalent) number of people are either extremely tall or extremely short.

A graphical representation of a normal distribution is sometimes called a bell curve because of its flared shape. The precise shape can vary according to the distribution of the population but the peak is always in the middle and the curve is always symmetrical. In a normal distribution, the mean,median,mode are all the same.

Question 11:

I would do it only if the learning algorithm didn't handle missing values correctly.

The problem is that you're assuming your attributes are independent of one another -- in other words P(missing value | other attributes) = P (missing value). If that were the case in general, then Naive Bayes would perform very well on your data and you wouldn't need a fancy learning algorithm.

For example, consider a simple learning task: learning XOR. The truth table -- just to remind you is:

A	**B**	X
0	0	0
0	1	1
1	0	1
1	1	1

And now let's say you randomly inject "missing values" into column B randomly and have some repetitions, e.g.

0	0	0
0	?	0
0	1	1
1	0	1
1	1	0
1	1	0
1	?	1

So now we replace it with the median, so now we have:

0	0	0
0	1	0
0	1	1
1	0	1
1	1	0
1	1	0
1	1	1

And that's essentially unlearnable for any classifier, because it has -- for example, no way to get a 100% accurate solution because it's getting contradictory info: in one case 0,1 gives 0 and in one case it gives 1. Also, in two cases 1,1 gives 0 and one case 1,1 gives 1. Sometimes the learner will get it right and in others it will get it wrong, depending on the bias of the classifier. The bolding indicates the now-contradictory data.

But if you model the missing values correctly, you can still learn the correct solution easily, since the learner will pick up that it can substitute for the missing values correctly and get the correct results.

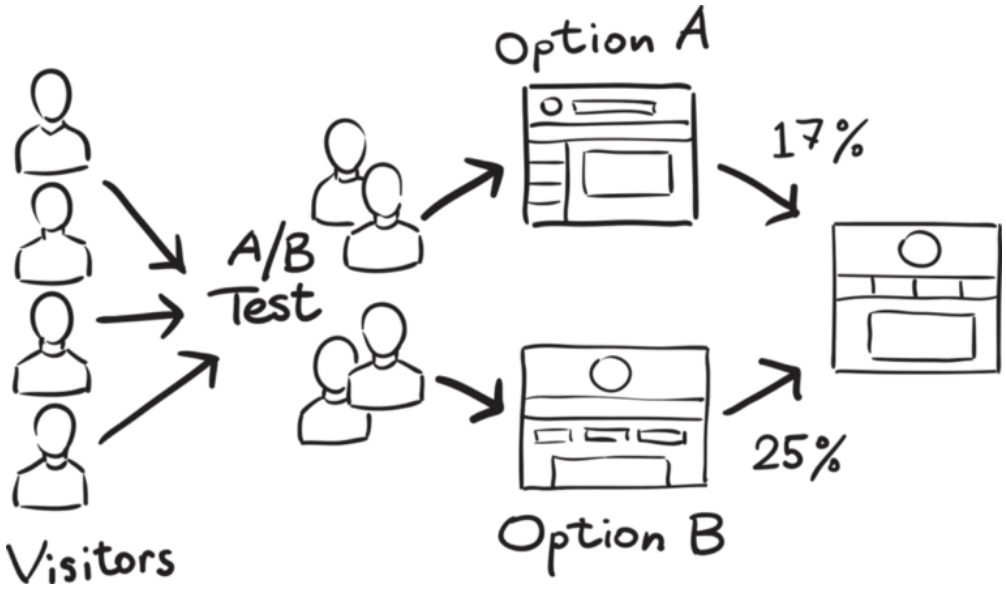
Decision trees, for example, would eat the above example for breakfast intelligently handling the missing values and assuming that they are whatever the **local** median rather than the global one.

Question 12:

A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

For instance, let’s say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.

In the above scenario, you may divide the products into two parts – A and B. Here A will remain unchanged while you make significant changes in B’s packaging. Now, on the basis of the response from customer groups who used A and B respectively, you try to decide which is performing better.



It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. The **population** refers to all the customers buying your product, while the **sample** refers to the number of customers that participated in the test.

Question 13:

The process of replacing null values in a data collection with the data's mean is known as mean imputation.

Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does.

Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

Question 14:

Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things:

(1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable?

(2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable?

These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula $y = c + b*x$, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

Question 15:

The Branches of Statistics

(1) Descriptive Statistics:

- The branch of statistics that focuses on collecting, summarizing, and presenting a set of data.

- The average age of citizens who voted for the winning candidate in the last presidential election, the average length of all books about statistics, the variation in the weight of 100 boxes of cereal selected from a factory's production line.

(2) Inferential Statistics

- The branch of statistics that analyzes sample data to draw conclusions about a population.
- The average age of citizens who voted for the winning candidate in the last presidential election, the average length of all books about statistics, the variation in the weight of 100 boxes of cereal selected from a factory's production line.