Statistics Interview Questions

We believe that you have learned both theoritical and practical knowledge on Naive Bayes classification algorithm through your assignment.

So let's test your knowledge here. This will help you to be prepared for interviews too!

Best with Quest

1. What is the Central Limit Theorem and why is it important?

```
The central limit theorem states that if you have a population
with mean μ and standard deviation σ and take sufficiently large random
samples from the population with replacement, then the distribution of
the sample means will be approximately normally distributed.
This will hold true regardless of whether the source population is
normal or skewed, provided the sample size is sufficiently large
(usually n > 30).

Intuitively, if you take means of multiple samples, you can assume that
mean of all the samples combines will be equal to the mean of the
population. This is really important to understand as this allows us to
do significance tests like z-tests.
```

2. What is sampling? How many sampling methods do you know?

```
Sampling is the process of selecting a subset(a predetermined number of observations) from
a larger population. It’s a pretty common technique wherein, we run experiments and draw
conclusions about the population, without the need of having to study the entire
population.

To draw valid conclusions from your results, you have to carefully decide how you will
select a sample that is representative of the group as a whole. There are two types of
sampling methods:

1. Probability sampling involves random selection, allowing you to make strong statistical
inferences about the whole group.
2. Non-probability sampling involves non-random selection based on convenience or other
criteria, allowing you to easily collect data.
```

3. What is selection bias?

```
Selection bias occurs in an active sense when the sample data that is gathered and
prepared for modelling has characteristics that are not representative of the true, future
population of cases the model will see. That is, selection bias develops when a subset of
the data is systematically (i.e., non-randomly) excluded from the analysis.
```

So, the initial sample that was carefully planned no longer represents the broader population. This is why, for example, the US Government conducts a census at regular intervals, to provide government agencies with essential demographic information about the population at a given point in time. But that information becomes defunct, as do the economic models built upon it.

Continuing to use the outdated sample introduces bias into the data. However, selection bias can be mitigated with the help of various strategies. When the data sample is created, the sampling strategy should be documented, and any constraints of the procedure ought to be properly expressed. This documentation will highlight the probability of selection bias once the model is built and deployed.

4. What is an example of a data set with a non-Gaussian distribution?

Any distribution of money or value will be non--Gaussian. For example: distributions of income; distributions of house prices; distributions of bets placed on a sporting event. These distributions cannot have negative values and will usually have extended right hand tails.

5. What is the Binomial Probability Formula

Binomial distribution is a probability distribution for the number of successes in a sequence of Bernoulli trials (Weiss, 2015). Bernoulli trials is a series of repeated trials of an experiment with:
- only one of two possible outcomes, success (s) or failure (f)
- outcome on one trial is independent and would not affect the outcome on other trial
- probability of success remains the same from trial to trial

The binomial distribution is known as a discrete distribution as it represents the probability for a distinct “x” number of success in “n” number of trials.

6. What is the difference between “long” and “wide” format data?

In the wide-format, a subject’s repeated responses will be in a single row, and each response is in a
separate column. In the long-format, each row is a one-time point per subject. You can recognize data in
wide format by the fact that columns generally represent groups.

7. What do you understand by the term Normal Distribution?

Data is usually distributed in different ways with a bias to the left or to the right or it can all be jumbled up.
However, there are chances that data is distributed around a central value without any bias to the left or right
and reaches normal distribution in the form of a bell-shaped curve.

The random variables are distributed in the form of a symmetrical, bell-shaped curve.
Properties of Normal Distribution are as follows;
1. Unimodal -one mode
2. Symmetrical -left and right halves are mirror images
3. Bell-shaped -maximum height (mode) at the mean
4. Mean, Mode, and Median are all located in the center
5. Asymptotic

8. What is correlation and covariance in statistics?

Covariance and Correlation are two mathematical concepts; these two approaches are widely used in
statistics. Both Correlation and Covariance establish the relationship and also measure the dependency
between two random variables. Though the work is similar between these two in mathematical terms, they
are different from each other.

- Correlation: Correlation is considered or described as the best technique for measuring and also for estimating the
quantitative relationship between two variables. Correlation measures how strongly two variables are
related.
- Covariance: In covariance two items vary together and it's a measure that indicates the extent to which two
random variables change in cycle. It is a statistical term; it explains the systematic relation between a pair of
random variables, wherein changes in one variable reciprocal by a corresponding change in another
variable.

9. What is the difference between Point Estimates and Confidence Interval?

Point Estimation gives us a particular value as an estimate of a population parameter. Method of Moments
and Maximum Likelihood estimator methods are used to derive Point Estimators for population parameters.
A confidence interval gives us a range of values which is likely to contain the population parameter. The
confidence interval is generally preferred, as it tells us how likely this interval is to contain the population
parameter. This likeliness or probability is called Confidence Level or Confidence coefficient and represented
by 1 – alpha, where alpha is the level of significance.

10. What is the goal of A/B Testing?

It is a hypothesis testing for a randomized experiment with two variables A and B.
The goal of A/B Testing is to identify any changes to the web page to maximize or increase the outcome of
interest. A/B testing is a fantastic method for figuring out the best online promotional and marketing strategies
for your business. It can be used to test everything from website copy to sales emails to search ads

An example of this could be identifying the click-through rate for a banner ad.

11. What is p-value?

When you perform a hypothesis test in statistics, a p-value can help you determine the strength of your

results. p-value is a number between 0 and 1. Based on the value it will denote the strength of the results.
The claim which is on trial is called the Null Hypothesis.
Low p-value (≤ 0.05) indicates strength against the null hypothesis which means we can reject the null
Hypothesis. High p-value (≥ 0.05) indicates strength for the null hypothesis which means we can accept the
null Hypothesis p-value of 0.05 indicates the Hypothesis could go either way. To put it in another way,
High P values: your data are likely with a true null. Low P values: your data are unlikely with a true null.

12. What do you understand by statistical power of sensitivity and how do you calculate it?

Sensitivity is commonly used to validate the accuracy of a classifier (Logistic, SVM, Random Forest etc.).
Sensitivity is nothing but "Predicted True events/ Total events". True events here are the events which were
true and model also predicted them as true.
Calculation of seasonality is pretty straightforward.

Seasonality = (True Positives) / (Positives in Actual Dependent Variable)

13. Why Is Re-sampling Done?

Resampling is done in any of these cases:
• Estimating the accuracy of sample statistics by using subsets of accessible data or drawing randomly
with replacement from a set of data points
• Substituting labels on data points when performing significance tests
• Validating models by using random subsets (bootstrapping, cross-validation

14. What Is the Law of Large Numbers?

It is a theorem that describes the result of performing the same experiment a large number of times. This
theorem forms the basis of frequency-style thinking. It says that the sample means, the sample variance
and the sample standard deviation converge to what they are trying to estimate.

15. What Are Confounding Variables?

In statistics, a confounder is a variable that influences both the dependent variable and independent variable.
For example, if you are researching whether a lack of exercise leads to weight gain,
- lack of exercise = independent variable
- weight gain = dependent variable.
A confounding variable here would be any other variable that affects both of these variables, such as the age
of the subject.

16. What Are the Types of Biases That Can Occur During Sampling?

```
- Selection bias
• Under coverage bias
• Survivorship bias

- Survivorship Bias
It is the logical error of focusing aspects that support surviving some process and
casually overlooking those
that did not work because of their lack of prominence. This can lead to wrong conclusions
in numerous
different means.
- selection Bias
Selection bias occurs when the sample obtained is not representative of the population
intended to be
analysed.
```

17. How to handle skewness?]What are the mathematical properties of skewed data? How can you fix it?

```
Skewness is when the distribution of data which is concentrated
in 1 side(left or right) more, or it is sort of collected on one side.
data can be left-skewed or right-skewed.

Right skewed is when data is concentrated on the left, and the tail is
on right. This is also known as positive skewed. Left is -ve skewed
when the tail is on left.
Skewness is caused by the presence of Outliers in the data.
In mathematical terms, the right skewness is when the mean is greater
than the median.
We can remove Skewness by omitting Outliers or taking Log of the
features.
```

18. What is Mean, Median, Mode, and Range?

```
The "mean" is the "average" you're used to, where you add up all the numbers and then
divide by the number of numbers. The "median" is the "middle" value in the list of
numbers. To find the median, your numbers have to be listed in numerical order from
smallest to largest, so you may have to rewrite your list before you can find the median.
The "mode" is the value that occurs most often. If no number in the list is repeated, then
there is no mode for the list.

The "range" of a list a numbers is just the difference between the largest and smallest
values.
```

19. Difference between normal and gaussian distribution?

```
None. Gauss was one of the people who first gave a derivation for the distribution. It was
called “normal” because it was regarded as a norm or standard. It was later called
“Gaussian” in honour of Gauss and because some people think that “normal” exaggerates its
importance.
```

20. Difference between standard scalar and normal scalar?

Normalization vs. standardization is an eternal question among machine learning newcomers. Let me elaborate on the answer in this section.

Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.
Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.
However, at the end of the day, the choice of using normalization or standardization will depend on your problem and the machine learning algorithm you are using. There is no hard and fast rule to tell you when to normalize or standardize your data. You can always start by fitting your model to raw, normalized and standardized data and compare the performance for best results.

Now Rest with this Quest :)