Hypothesis Testing 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 so important? Can you give an example to denote the working of the central limit theorem? ¶

```
The central limit theorem is very powerful — it states that the distribution of sample
means approximates a normal distribution.
To give an example, you would take a sample from a data set and calculate the mean of that
sample. Once repeated multiple times, you would plot all your means and their frequencies
onto a graph and see that a bell curve, also known as a normal distribution, has been
created. The mean of this distribution will closely resemble that of the original data.
The central limit theorem is important because it is used in hypothesis testing and also
to calculate confidence intervals.




Let’s consider the population of men who have normally distributed weights, with a mean of
60 kg and a standard deviation of 10 kg, and the probability needs to be found out.

If one single man is selected, the weight is greater than 65 kg, but if 40 men are
selected, then the mean weight is far more than 65 kg.

The solution to this can be as shown below:

Z = (x – μ) / ? = (65 – 60) / 10 = 0.5

For a normal distribution P(Z > 0.5) = 0.409
Z = (65 - 60) / 5 = 1
P(Z > 1) = 0.090
```

2. What general conditions must be satisfied for the central limit theorem to hold? What are some of the properties of a normal distribution?

```
- The data must be sampled randomly
- The sample values must be independent of each other
- The sample size must be sufficiently large, generally it should be greater or equal than
30


A normal distribution, regardless of its size, will have a bell-shaped curve that is
symmetric along the axes.

Following are some of the important properties:

- Unimodal: It has only one mode.
```

- Symmetrical: Left and right halves of the curve are mirrored.
- Central tendency: The mean, median, and mode are at the midpoint.

3. What is the meaning of degrees of freedom (DF) in statistics?

Degrees of freedom or DF is used to define the number of options at hand when performing an analysis. It is mostly used with t-distribution and not with the z-distribution.

If there is an increase in DF, the t-distribution will reach closer to the normal distribution. If DF > 30, this means that the t-distribution at hand is having all of the characteristics of a normal distribution.

4. Briefly explain the procedure to measure the length of all sharks in the world.

Following steps can be used to determine the length of sharks:

1. Define the confidence level (usually around 95%)
2. Use sample sharks to measure
3. Calculate the mean and standard deviation of the lengths
4. Determine t-statistics values
5. Determine the confidence interval in which the mean length lies

5. What is an alternative hypothesis? How do the standard error and the margin of error relate?

The alternative hypothesis (denoted by H1) is the statement that must be true if the null hypothesis is false. That is, it is a statement used to contradict the null hypothesis. It is the opposing point of view that gets proven right when the null hypothesis is proven wrong.

The standard error and the margin of error are quite closely related to each other. In fact, the margin of error is calculated using the standard error. As the standard error increases, the margin of error also increases.

6. What is one sample t-test?

This T-test is a statistical hypothesis test in which we check if the mean of the sample data is statistically or significantly different from the population’s mean.

7. What is the difference between the Ist quartile, the IInd quartile, and the IIIrd quartile?

Quartiles are used to describe the distribution of data by splitting data into three equal portions, and the boundary or edge of these portions are called quartiles.

That is,

- The lower quartile (Q1) is the 25th percentile.
- The middle quartile (Q2), also called the median, is the 50th percentile.
- The upper quartile (Q3) is the 75th percentile.

8. What is the relationship between mean and median in a normal distribution? What are the examples of symmetric distribution?

In a normal distribution, the mean is equal to the median. To know if the distribution of a dataset is normal, we can just check the dataset’s mean and median.

Symmetric distribution means that the data on the left side of the median is the same as the one present on the right side of the median.

There are many examples of symmetric distribution, but the following three are the most widely used ones:

- Uniform distribution
- Binomial distribution
- Normal distribution

9. Where is inferential statistics used? What is the difference between descriptive and inferential statistics?

Inferential statistics is used for several purposes, such as research, in which we wish to draw conclusions about a population using some sample data. This is performed in a variety of fields, ranging from government operations to quality control and quality assurance teams in multinational corporations.

Descriptive statistics: Descriptive statistics is used to summarize from a sample set of data like the standard deviation or the mean.

Inferential statistics: Inferential statistics is used to draw conclusions from the test data that are subjected to random variations.

10. What are left-skewed and right-skewed distributions? 37. What is the difference between one tail and two tail hypothesis testing?

A left-skewed distribution is one where the left tail is longer than that of the right tail. Here, it is important to note that the mean < median < mode.

Similarly, a right-skewed distribution is one where the right tail is longer than the left one. But, here the mean > median > mode.

1. 2-tail test: Critical region is on both sides of the distribution
H0: x = μ
H1: x <> μ
2. 1-tail test: Critical region is on one side of the distribution
H1: x <= μ
H1: x > μ

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

You would perform hypothesis testing to determine statistical significance. First, you would state the null hypothesis and alternative hypothesis. Second, you would calculate the p-value, the probability of obtaining the observed results of a test assuming that the null hypothesis is true. Last, you would set the level of the significance (alpha) and if the p-value is less than the alpha, you would reject the null — in other words, the result is statistically significant.

12. Explain selection bias (with regard to a dataset, not variable selection). Why is it important? How can data management procedures such as missing data handling make it worse?

Selection bias is the phenomenon of selecting individuals, groups or data for analysis in such a way that proper randomization is not achieved, ultimately resulting in a sample that is not representative of the population.
Understanding and identifying selection bias is important because it can significantly skew results and provide false insights about a particular population group.

Types of selection bias include:
- sampling bias: a biased sample caused by non-random sampling
time interval: selecting a specific time frame that supports the desired conclusion. e.g. conducting a sales analysis near Christmas.
- exposure: includes clinical susceptibility bias, protopathic bias, indication bias. Read more here.
- data: includes cherry-picking, suppressing evidence, and the fallacy of incomplete evidence.
- attrition: attrition bias is similar to survivorship bias, where only those that ‘survived’ a long process are included in an analysis, or failure bias, where those that ‘failed’ are only included
- observer selection: related to the Anthropic principle, which is a philosophical consideration that any data we collect about the universe is filtered by the fact that, in order for it to be observable, it must be compatible with the conscious and sapient life that observes it. [3]

Handling missing data can make selection bias worse because different methods impact the data in different ways. For example, if you replace null values with the mean of the data, you adding bias in the sense that you’re assuming that the data is not as spread out as it might actually be.

13. Is mean imputation of missing data acceptable practice? Why or why not? Give an example where the median is a better measure than the mean

Mean imputation is the practice of replacing null values in a data set with the mean of the data.
Mean imputation is generally bad practice because it doesn’t take into account feature correlation. For example, imagine we have a table showing age and fitness score and imagine that an eighty-year-old has a missing fitness score. If we took the average fitness score from an age range of 15 to 80, then the eighty-year-old will appear to have a much higher fitness score that he actually should.
Second, mean imputation reduces the variance of the data and increases bias in our data. This leads to a less accurate model and a narrower confidence interval due to a smaller variance.

median is a better measure than the mean when there are a number of outliers that positively or negatively skew the data.

14. When you sample, what bias are you inflicting? How do you control for biases? What is the empirical rule?

Potential biases include the following:
1. Sampling bias: a biased sample caused by non-random sampling
2. Under coverage bias: sampling too few observations
3. Survivorship bias: error of overlooking observations that did not make it past a form of selection process.

There are many things that you can do to control and minimize bias. Two common things include randomization, where participants are assigned by chance, and random sampling, sampling in which each member has an equal probability of being chosen.

The empirical rule states that if a dataset is normally distributed, 68% of the data will fall within one standard deviation, 95% of the data will fall within two standard deviations, and 99.7% of the data will fall within 3 standard deviations.

15. When should you use a t-test vs a z-test?

A Z-test is a hypothesis test with a normal distribution that uses a z-statistic. A z-test is used when you know the population variance or if you don’t know the population variance but have a large sample size.
A T-test is a hypothesis test with a t-distribution that uses a t-statistic. You would use a t-test when you don’t know the population variance and have a small sample size.

16. How would you describe what a ‘p-value’ is to a non-technical person?

The best way to describe the p-value in simple terms is with an example. In practice, if the p-value is less than the alpha, say of 0.05, then we’re saying that there’s a probability of less than 5% that the result could have happened by chance. Similarly, a p-value of 0.05 is the same as saying “5% of the time, we would see this by chance.”

17. What does it mean if a model is heteroscedastic? what about homoscedastic?

A model is heteroscedastic when the variance in errors is not consistent. Conversely, a model is homoscedastic when the variances in errors is consistent.

18. Is it better to have too many false positives, or too many false negatives? Explain.

It depends on the question as well as on the domain for which we are trying to solve the question.

In medical testing, false negatives may provide a falsely reassuring message to patients and physicians that disease is absent, when it is actually present. This sometimes leads to inappropriate or inadequate treatment of both the patient and their disease. So, it is desired to have too many false positive.

For spam filtering, a false positive occurs when spam filtering or spam blocking techniques wrongly classify a legitimate email message as spam and, as a result, interferes with its delivery. While most anti-spam tactics can block or filter a high percentage of unwanted emails, doing so without creating significant false-positive results is a much more demanding task. So, we prefer too many false negatives over many false positives.

19. What does it mean by bell curve distribution and Gaussian distribution? How to convert normal distribution to standard normal distribution?

Normal distribution is called bell curve distribution / Gaussian distribution
It is called bell curve because it has the shape of a bell
It is called Gaussian distribution as it is named after Carl Gauss

Standardized normal distribution has mean = 0 and standard deviation = 1

```
Standardized normal distribution has mean = 0 and standard deviation = 1

To convert normal distribution to standard normal distribution we can use the formula

X (standardized) = (x-μ) / σ
```

20. What is the difference between population parameters and sample statistics? What do you think of the tail (one tail or two tail) if H0 is equal to one value only? What is the critical value in one tail or two-tail test? Why is the t-value same for 90% two tail and 95% one tail test?

```
1. Population parameters are:
- Mean = μ
- Standard deviation = σ

2. Sample statistics are:
- Mean = x (bar)
- Standard deviation = s

Tail (one tail or two tail) if H0 is equal to one value only is a two-tail test.

- Critical value in 1-tail = alpha
- Critical value in 2-tail = alpha / 2


t-value same for 90% two tail and 95% one tail test because_:-

P-value of 1-tail = P-value of 2-tail / 2

It is because in two tail there are 2 critical regions
```

In [ ]:

```
----------------------

#  Now Rest with this Quest :)


----------------------
```