STATISTICS WORKSHEET – 1

SUBJECTIVE TYPE QUESTION ANSWER

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

   Ans. Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graphical form, the normal distribution appears as a "BELL CURVE". The normal distribution is the proper term for a probability bell curve. In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3. Normal distributions are symmetrical, but not all symmetrical distributions are normal. Many naturally-occurring phenomena tend to approximate the normal distribution. In finance, most pricing distributions are not, however, perfectly normal. The normal distribution model is important in statistics and is key to the Central Limit Theorem (CLT). This theory states that averages calculated from independent, identically distributed random variables have approximately normal distributions, regardless of the type of distribution from which the variables are sampled (provided it has finite variance). The normal distribution has several key features and properties that define it. First,
   its mean (average), median (midpoint), and mode (most frequent observation) are all equal to one another. Moreover, these values all represent the peak, or highest point, of the distribution. The distribution then falls symmetrically around the mean, the width of which is defined by the standard deviation.

2. How do you handle missing data? What imputation technique do you recommend?

   Ans. Missing data can be deal with various ways. The most common reaction is to ignore it. Choosing to make no decision, on the other hand, indicates that your statistical programme will make the decision for you. Your application will remove things in a listwise sequence most of the time. Depending on why and how much data is gone, listwise deletion may or may not be a good idea.
   Another common strategy among those who pay attention is imputation. Imputation is the process of substituting an estimate for missing values and analysing the entire data set as if the imputed values were the true observed values. Here are some methods:

   ### Mean imputation

   Calculate the mean of the observed values for that variable for all non-missing people. It has the advantage of maintaining the same mean and sample size, but it also has a slew of drawbacks.

   Substitution Assume the value from a new person who was not included in the sample. To put it another way, pick a new subject and employ their worth instead.

   ### Regression imputation

The result of regressing the missing variable on other factors to get a predicted value. As a result, instead of utilising the mean, you're relying on the anticipated value, which is influenced by other factors. This keeps the associations between the variables in the imputation model, but not the variability around the anticipated values.

3. What is A/B testing?

Ans. A/B testing involves comparing two different approaches to solving a problem – approach A and approach B – and working out which is better based on data. A/B testing is also known as randomized testing, two-sample testing, split testing, and bucket testing.

There are two key aspects to an A/B test. The first is that people must be randomly assigned to one of two groups, with one group shown the A approach and the other group, the B approach. The second aspect is that statistical testing is used work out whether A, B, or neither is better.

A/B testing requires:

The comparison of two treatments or approaches. You can do more than two, but then you are doing *multivariate testing*.

A clear hypothesis about the difference between the two treatments. At the end of the A/B test you want to be able to say not only which option won but also understand why it won. This usually means that you should design an A/B test with treatments that are pretty similar, making it relatively easy to isolate which part of the treatment led to conversion. For example, you may just test the color of a button.

You also need a relatively large number of people to expose to each of the treatments. How many people depends on how big the difference is likely to be between the effects of the treatments. If testing minor differences in wording or colors, you usually need thousands of people. If you expect big differences, such as 20% differences between the A and the B, then samples with as few as 50 in each group can be OK.

A metric for comparing the two treatments. In the example above the metric was the click- through rate. There are many other possible metrics, such as including average dollars spent, and percentage that signed up.

A *randomization* mechanism for assigning people to each of the treatments. *Random* means that software, rather than people, used a random number generator to allocate people to each group.

No ability for human intervention to ruin the test. For example, A/B testing involving salespeople testing different scripts often fails because if a person does not like their sales script, they tend to modify it or make a half-hearted attempt.

Statistical testing software to work out the strength of evidence in favor of A versus B. The smaller the sample size, the more important it is to use statistical testing software, because as a rule, most people over-interpret differences, thinking that A or B has tested better, when there is no evidence one way or another.

4. Is mean imputation of missing data acceptable practice?

Ans. Imputation is the process of replacing missing values with substituted data. It is done as a preprocessing step. 3. NORMAL IMPUTATION In our example data, we have an f1 feature that has missing values. We can

replace the missing values with the below methods depending on the data type of feature f1. Although imputing missing values by using the mean is a popular imputation technique, there are serious problems with mean imputation. The variance of a mean-imputed variable is always biased downward from the variance of the un-imputed variable. It is acceptable when the missing value proportion is not large enough.

But, when the missing values are large enough and you impute them with the mean, the standard errors will be lesser than what they actually would have been.

Small standard errors can lead to small p-values and this can create problems for us,

because some variables will start appearing significant, which are ideally not significant.

5. What is linear regression in statistics?

Ans. Linear regression is basically a statistical modeling technique which used to show the relationship between one dependent variable and one or more independent variable. It is one of the most common types of predictive analysis. This type of distribution forms in a line hence this is called linear regression.

Linear regression is used to predict the relationship between two variables by applying a linear equation to observed data. There are two types of variable, one variable is called an independent variable, and the other is a dependent variable.

Linear regression is used to predict the relationship between two variables by applying a linear equation to observed data. There are two types of variable, one variable is called an independent variable, and the other is a dependent variable. Linear regression is commonly used for predictive analysis.

6. What are the various branches of statistics?

Ans. The two main branches of statistics are descriptive statistics and inferential statistics. Both of these are employed in scientific analysis of data and both are equally important for the student of statistics.

Descriptive Statistics

Descriptive statistics deals with the presentation and collection of data. This is usually the first part of a statistical analysis. It is usually not as simple as it sounds, and the statistician needs to be aware of designing experiments, choosing the right focus group and avoid biases that are so easy to creep into the experiment.

Different areas of study require different kinds of analysis using descriptive statistics. For example, a physicist studying turbulence in the laboratory needs the average quantities that vary over small intervals of time. The nature of this problem requires that physical quantities be averaged from a host of data collected through the experiment.

Inferential Statistics

Inferential statistics, as the name suggests, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics.

Most prediction of the future and generalizations about a population by studying a smaller sample come under the purview of inferential statistics. Most social sciences experiments deal with studying a small sample population that helps

determine how the population in general behaves. By designing the right experiment, the researcher is able to draw conclusion relevant to his study. While drawing conclusions, one needs to be very careful so as not to draw the wrong or biased conclusions. Even though this appears like a science, there are ways in which one can manipulate studies and results through various means. For example, data dredging is increasingly becoming a problem as computers hold loads of information and it is easy, either intentionally or unintentionally, to use the wrong inferential methods.

Both descriptive and inferential statistics go hand in hand and one cannot exist without the other. Good scientific methodology needs to be followed in both these steps of statistical analysis and both these branches of statistics are equally important for a researcher.