Worksheet 4 Statistics answers

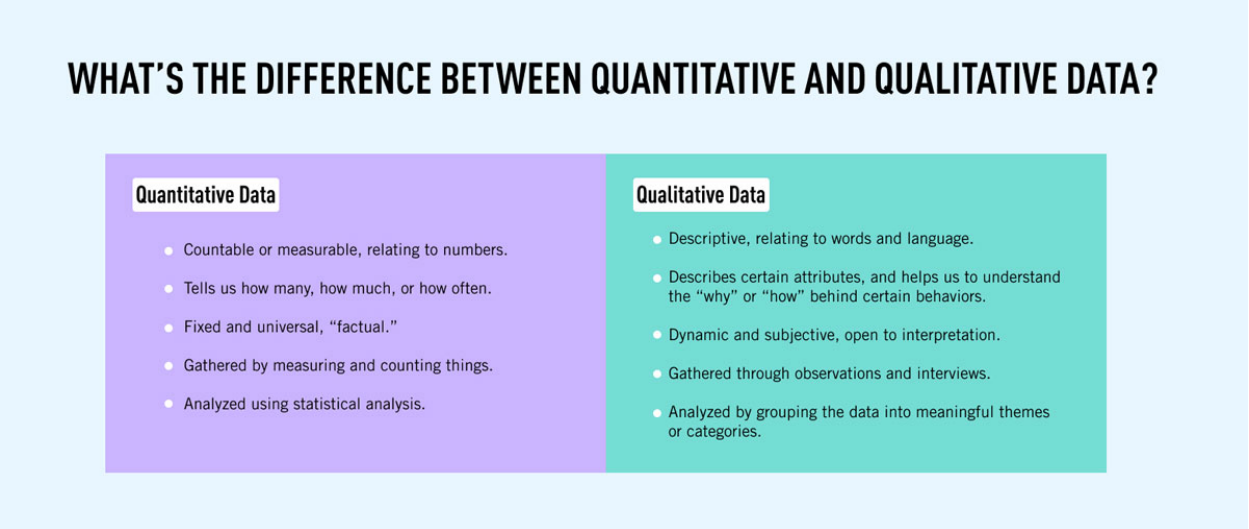
1. Central limit theorem states that when the samples have a finite variance then the samples would get distributed normally and the mean of the whole sample can be considered to be the mean of the whole population. It is very important because it tells us exactly how much sample size is supposed to be increased in order to decrease the error which will then benefit us in various other statistical analyses.
2. Sampling is basically a group from a larger data that will be used for further analysis. Few of the sampling methods are, Random sampling, Clustering sampling , Systematic sampling etc.
3. Type 1 error or False positive actually occurs when a null hypothesis is rejected even if it is true for the complete population whereas Type 2 error or False negative occurs when a null hypothesis is not rejected even if it is false.
4. Normal distribution is nothing but a symmetrical plot of data around the mean value.
5. Correlation is a statistical measure as to how much the two variables are related to each other. Covariance tells us to what extent the two variables are related to each other. Higher value of covariance means higher the dependency.
6.

Univariate	Bivariate	Multivariate
Univariate statistics summarize only one variable at a time.	Bivariate statistics compare two variables.	Multivariate statistics compare more than two variables.
It does not deal with causes or relationships	It deals with causes and relationships and the analysis is done	It also deals with causes and relationships and the analysis is done
It does not contain any dependent variable.	It contain only one dependent variable.	It is similar to bivariate but contains more than one dependent variable.
The purpose of univariate analysis is to describe	The purpose of univariate analysis is to explain.	The purposes of multivariate data analysis is to study the relationships among the P attributes, classify the n collected samples into homogeneous groups, and make inferences about the underlying populations from the sample.
The example of a univariate data can be height.	Example of bivariate data can be temperature and ice cream sales in summer season.	Example of this type of data is suppose an advertiser wants to compare the popularity of four advertisements on a website, then their click rates could be measured for both men and women and relationships between variables can then be examined.

7. Sensitivity is the percentage of true positives. We can calculate the sensitivity by Dividing the true positives with a summation of true positives and false negatives.

$$\text{Sensitivity} = \frac{\text{True positive}}{\text{True positive + false negative}}$$

8. Hypothesis testing is a type of statistical analysis technique that puts our assumption regarding the population to the test. It is basically used to calculate the relationship between the two statistical variables. H0 and H1 are two statistical hypotheses one of which will be proved to be wrong, example- the defendant is guilty or the defendant is innocent. In the two tailed test it will be considered that the mean is equal to a variable x and then it will be calculated how significantly the mean varies from this value of x.
9. Quantitative data is measurably related to numbers whereas Qualitative data is more about words, language etc. Quantitative data tells us about the value, how much, how many and Qualitative data tells us about the why of the data.



10. Range is the Highest value - Lowest value whereas to find the IQR to find IQR we will need the value of the median of the lower and upper half of the data. After finding this Q3 and Q1 the difference of these two data will give us the Interquartile Range.
11. A bell curve is a very important part of statistics. It is a type of graph in which a certain value from a group of data is chosen to depict a curve in which we have a central normal value and a few extreme values to give this graph symmetry.

12. We can use interquartile range to make a fencing for our data and only take values that are extremely important. This way we can nullify the outliers using interquartile range.
13. P value or probability value tells how likely it was for the data to occur in the null hypothesis.
14.

Formula

$$P_x = \binom{n}{x} p^x q^{n-x}$$

P = binomial probability

x = number of times for a specific outcome within n trials

$\binom{n}{x}$ = number of combinations

p = probability of success on a single trial

q = probability of failure on a single trial

n = number of trials

15. ANOVA stands for analysis of variance and it is used to determine the difference between the means of more than two groups.ANOVA is very useful because it helps us to analyze multiple groups. It makes sure that overall testing of all the groups is taking place without any issues and also tells us about the type 1 error rate.