Statistics – 4

1. The central limit theorem states that the sampling distribution of the mean approaches a normal distribution as the size of the sample increases, regardless o the shape of the original population distribution.

   It is important because we can safely assume that the sampling distribution of the mean will be normal in most cases. And our machine learning models works best on normally distributed data. It will provide more certainty in estimates.

2. Sampling in statistics is the selection of a subset of individuals from within a statistical population to estimate characteristics of the whole population.

   I know 4 methods of sampling.

3. Type 1 error is when the null hypothesis is true and we rejected it. Type 2 error is when the null hypothesis is false and we failed to reject it.

4. A normal distribution is the proper term for a probability bell curve. In a normal distribution, mean = 0 and standard deviation = 1.
   There can be a normal distribution only on numerical continuous data and not on categorical data.

5. Covariance is an indicator of the extent to which two random variables are dependent on each other. A higher number denotes higher dependency.
   Whereas correlation is a statistical measure that indicates how strongly two variables are related.

6. Univariate statistics summarize only one variable at a time. Bivariate statistics compare two variables. Multivariate statistics compare more than two variables.

7. Sensitivity is the percentage of True Positives. $Sensitivity = \frac{TN}{TN+FP}$

8. A statistical hypothesis test is a method of statistical inference used to decide whether the data at hand sufficiently support a particular hypothesis.

   H0 is null hypothesis - decision always leads to status quo. Current status/assumption does not change.
   H1 is alternate hypothesis - decision always leads to opposite of H0.

   In a two-tailed test, the H0 is what currently stated to be true; and H1 is always what is being claimed whose value is different from H0. It is represented using 'not equal to' or '< or >' sign.

9. Quantitative data are measures of values or counts and are expressed as numbers. Whereas qualitative data describes qualities or characteristics.

10. Range is calculated by subtracting the lowest value from the highest value.
    Inter Quartile Range (IQR) = Q3 – Q1
    Where Q1 = First quartile
    Q3 = Third quartile

11. Bell curve distribution is same as normal distribution. Its mean is zero and standard deviation is 1.

12. IQR can be used to find outliers.

13. P-value is a number calculated from a statistical test that describes how likely you are to have found a particular set of observations if the null hypothesis were true. It helps us to decide whether to reject the null hypothesis or not.

14. $P_x = {}^nC_x p^x q^{n-x}$

$$where\ p + q = 1\ (maximum\ probability)$$

15. ANOVA (Analysis of Variance) is used to compare differences of means among more than 2 groups. It does this by looking at variation in the data and where that variation is found (hence the name). Specifically, ANOVA compares the amount of variation between groups with the amount of variation within groups.