**Statistics**

1. What is the Central Limit Theorem?

The CLT is a statistical theory that states that - if you take a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from that population will be roughly equal to the population mean. The Central Limit Theorem is important for statistics because it allows us to safely assume that the sampling distribution of the mean will be normal in most cases. This means that we can take advantage of statistical techniques that assume a normal distribution, as we will see in the next section

1. Sampling is a process in statistical analysis where researchers take a predetermined number of observations from a larger population. The method of sampling depends on the type of analysis being performed, but it may include simple random sampling or systematic sampling. Types of Sampling

*There are five types of sampling: Random, Systematic, Convenience, Cluster, and Stratified.*

*Random sampling* is analogous to putting everyone's name into a hat and drawing out several names. Each element in the population has an equal chance of occurring. While this is the preferred way of sampling, it is often difficult to do. It requires that a complete list of every element in the population be obtained. Computer generated lists are often used with random sampling. You can generate random numbers using the TI82 calculator.

*Systematic sampling* is easier to do than random sampling. In systematic sampling, the list of elements is "counted off". That is, every kth element is taken. This is like lining everyone up and numbering off "1,2,3,4; 1,2,3,4; etc". When done numbering, all people numbered 4 would be used.

*Convenience sampling* is very easy to do, but it's probably the worst technique to use. In convenience sampling, readily available data is used. That is, the first people the surveyor runs into.

*Cluster sampling* is accomplished by dividing the population into groups -- usually geographically. These groups are called clusters or blocks. The clusters are randomly selected, and each element in the selected clusters are used.

*Stratified sampling* also divides the population into groups called strata. However, this time it is by some characteristic, not geographically. For instance, the population might be separated into males and females. A sample is taken from each of these strata using either random, systematic, or convenience sampling.

1. A type I error (false-positive) occurs if an investigator rejects a null hypothesis that is true in the population; a type II error (false-negative) occurs if the investigator fails to reject a null hypothesis that is false in the population.
2. normal distribution is **a type of continuous probability distribution in which most data points cluster toward the middle of the range, while the rest taper off symmetrically toward extreme**. The middle of the range is also known as the mean of the distribution.
3. **Covariance is an indicator of the extent to which 2 random variables are dependent on each other**. A higher number denotes higher dependency. Correlation is a statistical measure that indicates how strongly two variables are related
4. Univariate statistics summarise only one variable at a time.

Bivariate statistics compare two variables.

Multivariate statistics compare more than two variables.

1. Sensitivity analysis **determines how different values of an independent variable affect a particular dependent variable under a given set of assumptions**. In other words, sensitivity analyses study how various sources of uncertainty in a mathematical model contribute to the model's overall uncertainty.

The sensitivity is calculated by **dividing the percentage change in output by the percentage change in input**.

1. Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter. The methodology employed by the analyst depends on the nature of the data used and the reason for the analysis. Null hypothesis (H0): The null hypothesis here is what currently stated to be true about the population. In our case it will be the average height of students in the batch is 100. Alternate hypothesis (H1): The alternate hypothesis is always what is being claimed.
2. **Quantitative**data are measures of values or counts and are expressed as numbers.  
   Quantitative data are data about numeric variables (e.g., how many; how much;   
   **Qualitative**data are measures of 'types' and may be represented by a name, symbol  
   Qualitative data are data about categorical variables (e.g. what type).
3. The interquartile range is a measure of where the “middle fifty” is in a data set. Where a range is a measure of where the beginning and end are in a set, an interquartile range is a measure of where the bulk of the values lie. That’s why it’s preferred over many other measures of spread when reporting things like school performance or SAT scores.

The interquartile range formula is the first quartile subtracted from the third quartile:

IQR = Q3 – Q1.

1. A bell curve is a common type of distribution for a variable, also known as the normal distribution. The term "bell curve" originates from the fact that the graph used to depict a normal distribution consists of a symmetrical bell-shaped curve.

The highest point on the curve, or the top of the bell, represents the most probable event in a series of data (its mean, mode and *median* in this case), while all other possible occurrences are symmetrically distributed around the mean, creating a downward-sloping curve on each side of the peak. The width of the bell curve is described by its standard deviation.

1. This method is helpful if you have a few values on the extreme ends of your dataset, but you aren’t sure whether any of them might count as outliers.

**Interquartile range method**

Sort your data from low to high

Identify the first quartile (Q1), the median, and the third quartile (Q3).

Calculate your IQR = Q3 – Q1

Calculate your upper fence = Q3 + (1.5 \* IQR)

Calculate your lower fence = Q1 – (1.5 \* IQR)

Use your fences to highlight any outliers, all values that fall outside your fences.

Your outliers are any values greater than your upper fence or less than your lower fence.

1. The ***p*value**, or probability value, tells you how likely it is that your data could have occurred under the null hypothesis. It does this by calculating the likelihood of your **test statistic**, which is the number calculated by a statistical test using your data.

The *p*value tells you how often you would expect to see a test statistic as extreme or more extreme than the one calculated by your statistical test if the null hypothesis of that test was true. The *p*value gets smaller as the test statistic calculated from your data gets further away from the range of test statistics predicted by the null hypothesis.

The *p*value is a proportion: if your *p*value is 0.05, that means that 5% of the time you would see a test statistic at least as extreme as the one you found if the null hypothesis was true.

1. The binomial distribution formula is for any random variable X, given by.

P(x:n,p) = nCxx px(1-p)n-x **Or** P(x:n,p) = nCx px (q)n-x

where,

n = the number of experiments

x = 0, 1, 2, 3, 4, …

p = Probability of success in a single experiment

q = Probability of failure in a single experiment (= 1 – p)

The binomial distribution formula is also written in the form of n-Bernoulli trials, where nCx = n!/x!(n-x)!. Hence, P(x:n,p) = n!/[x!(n-x)!].px.(q)n-x

1. Analysis of variance, or ANOVA, is a statistical method that separates observed variance data into different components to use for additional tests.

A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables.

If no true variance exists between the groups, the ANOVA's F-ratio should equal close to 1.