How to describe data numerically: How can we present a set of observed values in an effective manner? What can we do with the dataset? What does it look like, graphically? What measures can I derive from the dataset? Data can be presented graphically such as bar graphs, box graphs, and histogram. Various statistics can be performed on dataset such as the computation of the central tendency such as mean, and dispersion such as variance, and standard deviation. The z-score is used to determine how far from the mean relative to standard deviation an observation value reside. We also learned how to detect, determine the outliers of a dataset. Outliers are often indication of insufficient accuracy data collection process. They also affect some of the central tendency values. For example, the mean can be greatly impacted by an outlier while the median is almost free of that dependency.

Discrete & Continuous probability distributions: A random variable is a variable that assumes unknown numerical value of an event occurs in an experiment. Discrete random variable can be of countable numerical value, while the continuous random variable can be of any value in a given intervals, therefore they are measurable. An event space of a given experiment is a set that consists of all possible events that it may occur. For example, the event space for a toss of a coin is head or tail. The event space for tossing an unbiased dice is 1, 2, 3, 4, 5, and 6. The probability is the ratio between the number of times a desirable event occurs and the number of times the experiment is repeated. For example, P (event is head for tossing a fair coin) = 1 / 2. However, P (2 heads event in tossing a fair coin twice) = 1 / 4. We know that the sum of all probabilities of an experiment is 1. A probability value is always between [0,1].

We compute the mean, or the expected value of a discrete random variable x by summing the product of probability of each event and the value x as follows, $\mu = \Sigma xp(x)$. The variance, $\sigma^2 = \Sigma(x- \mu)^2 p(x) = \Sigma x^2 p(x) - \mu^2$, can be easily computed. The standard deviation σ is the square root of the variance.

For the discrete probability distribution, one of the widely use probability distribution is the binomial. Its experiment contains only two possible events, true or false.

For the continuous probability distribution, we have the normal that is widely used. Normal distribution is commonly used because of the inference of population mean value, which, is based on the Central Limit Theorem. It also reflects many natural phenomena, height and weight, and manufacturing quality control. It has a bell shape distribution where both ends move closer to the x-coordinate. Uniform distribution is also significant.

Sampling distribution and Central Limit Theorem: Let's say we want to know on the average percentage of UB students like to eat in UB main cafeteria. How can we come up with such number? Well, let's say we would stand in front of the cafeteria and randomly ask 20 students who just had lunch there. How many times should we do it? When should we do these experiments? Obviously, if we were to get one sample of size 20, chance is that we get a lousy average for the entire population of UB students who eats from the UB cafeteria. So, let's say we would devise a plan to do in different hours of the day, different days of the weeks. Each time we ask 20 students. Let's say at the end of the experiment we have 40 samples of 20 responses each. The above process is called sampling technique. What comes out of this experiment is we would get 40 averages. From these averages, we would have an average of the sampling distribution. The final average is much more accurate.

The central limit theorem (CLT) states that if the size of the sample is large enough, the sampling distribution of sample means tends to approximate a normal distribution, regardless the distribution of the population. This profound and useful theorem is the corner stone of statistics field.

Estimation & Confidence Intervals: Since most of the time we cannot afford to sample the entire population, we can only experiment with a subset of the population. Well, then how do we know what the population statistics would be? For example, how do I know the average weight of an American adult of five feet nine? I would devise a plan to begin my experiment of a selected sample from the targeted population. From the statistic computed by using the sample values of the experiment, I would then estimate the population parameter. There are two types of estimations, one is point estimation and the other is interval estimation. Example above is of type point estimation. The confidence interval is a form of range estimate where we state that certain population parameter would be captured by the estimated range with a certain confidence level.

The point estimation is easy to understand as opposed to the confidence interval estimation. If we were to say 80% confidence interval of an average American adult weight is between 150 to 250 pounds, what does it really mean? It means that if we were to repeat our experiment m times of a fixed sample size n, at least 80m% we are sure this random range captures the population average weight. That is all it really means. It does not mean that we are 95% sure that the population's mean will fall within this range. The larger the sample size (i.e., having more information about the population), the less chance of the error we would get when we estimate. For example, if the sample is the population, then the error is zero. On the other extreme, if we estimate without any experiment, the error would be the worst! The smaller the sample size, the greater the error we would have when we estimate the population parameter.

Estimate population mean: From sample statistic one can estimates the mean of the population. Size of the sample is significant because if it's less than 30, student t-statistic will be used, given the population is normal. Otherwise, for large sample size, one would use z statistic to perform the estimate of the population mean by applying the Central Limit Theorem.

Estimate the difference between two population means: We wanted to know the difference between the Canadian and the American average grades of business statistics classes. Here we have two samples average and we want to know the difference of the two population means. If the sample size is large, z statistic would be used, otherwise, given each population is normal then student t-statistic is used instead. If both populations have almost the same variance, we pool the sample variances.

Estimate the difference between two dependent population means: The matched pairs: Here the two populations have equal size. For example, we want to study the effectiveness of Tylenol in reducing the headache. We want to know the difference in the average number of hours takes to begin reduce the headache. We conduct a pair wise experiment by selecting a set of 20 people from the population. Each person with a headache takes a Tylenol. We then measure the effectiveness of it. Again, we would use z or student t statistic to perform the estimations.

Estimate a population variance: We perform an experiment by randomly select the coffee cans from the coffee maker factory. We measure the variation in weight from the sample. We want to know the variance of the entire population of coffee cans this company produces. We would use new statistic call the chi-square statistic.

Test Hypothesis: Hypothesis is testing a statement (a claim) about a numerical value of a population's parameter. We are studying two types of hypothesis; one is the null (always in + form) and the other is the alternative. The null hypothesis, H0, is to state the status-quo, statement that researcher wants to prove wrong. The alternative hypothesis, Ha, is a statement the researcher wants to use all their evidence to support. For example, people on average live to be around 77 years of age, but if the new drug that we developed were to be used, they will live up to 87 years of age. What would be the null hypothesis? It would be that the first statement, that on average human would live to around 77 years of age. Our alternative hypothesis would be that on average, human would live more than 77 years of age.

The testing hypothesis procedure is to gather enough evidence to either reject or not to reject the null hypothesis. There are two types of testing hypothesis; one is two-tailed test and the other is one-tailed test. In two-tail test the alternative hypothesis uses not equal sign (≠) while the one-tailed test, the alternative hypothesis uses greater than or less than operations (> or <). Our goal is to either reject or not to reject the null hypothesis with provided evidence. When we make such decision based on sample(s), we are facing with reducing the chance of making the wrong decision. The following is the chart shows possible outcomes in the decision making process:

GIVEN

	H_0 is true	H_0 is false
Not Reject H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision.

As you can see, you can make the type I error if you were to reject H_0 while it is true. The type II error is made when you do not reject the null hypothesis while it is false. The level of significance, α, is the probability of one would make type I error for a hypothesis test. In order to make decision on a null hypothesis, we would need the test statistic provided from the sample dataset. The test statistic is calculated based on the observation values, and the type of test hypothesis used. Based on the user's requirement about the level of significance we can use z, t, chi-square or F table (or these table required the normality condition) to determine the rejection regions. If the test statistic falls in the rejection region, we then should reject the null hypothesis. Otherwise, we would accept it. The rejection region is the set of possible values of the test statistic for which the null hypothesis will be rejected. The rejection region is confined by the critical value at the boundary. These critical values are often looked up from relevant statistical table.

Test Hypothesis – Applications: We now apply the test hypothesis: Population mean: McDonald claim there quarter hamburger on average is a quarter pounds. Wile customers believe it is much less. The FDA wishes to test the claim. This example requires the following null hypothesis.

H_0: $\mu = 1/4$ Pounds

H_a: $\mu < 1/4$ Pounds

For a large sample size, say 30 or more, we would use the z statistic to determine the rejection region before make decision. Otherwise, for the smaller sample size, use the student t statistic. Of course, we would have one-tailed as well as two-tailed test. Assumption is also that the relative frequency distribution of the population where the sample is drawn is approximately normal, which can be satisfied by measuring the pack containing 40 frozen hamburgers.

Two population means: Is there a difference in productivity of those who work at the office and those who telecommute daily? If so, what null hypothesis would it be? The comparison between two means is defined in the following null hypothesis:

H_0: $(\mu_1 - \mu_2) = 0$ there is no difference between the two groups.

H_a: $(\mu_1 - \mu_2) \neq 0$ there is difference in productivity level between the two groups.

The one-tailed and two-tailed tests can be used. Two different distributions are used for different sizes of the samples. For the large sample size, 30 or more, z distribution is used while student t distribution is used for small sample size of less than 30 (given both populations are normal. The assumption that the variances of the two samples are same, leads to pooling the variances.

Two population means of matched pairs: Similar to other test hypothesis, small size sample uses student t statistic while the large sample uses z statistic, on differences of the pairs.

Population variance: Consumers have been complaining that the amount of coffee in a 1lb coffee can seem to vary widely (i.e, low quality). The coffee can producer did a study and found that was true. The company is determines to fix by replacing new equipment. They expect the variance should decrease dramatically (i.e., improve the quality). The null hypothesis for this case is as follows:

H_0: $\sigma^2 = \sigma_0^2$ There is no difference in the variance of the amount of coffee. No change has occur.

H_a: $\sigma^2 < \sigma_0^2$ There variance is much smaller now.

Here we have one-tailed (or could have) the two-tailed tests. Chi-square statistic will be used to decide whether or not to reject the null hypothesis.

Ratio of two population variances: An automobile production manager wants to know compare the variance in the length of time each automobile is taken to produce from two separate factories. Each factory produces the same model of car. The production manager wants to know if the variances from two factories are same. The null hypothesis is as follows:

H_0: $\sigma^2 / \sigma_0^2 = 1$. The variances between two factories are similar.

H_a: $\sigma^2 / \sigma_0^2 \neq 1$. The variances between two factories are NOT same.

In this test hypothesis we uses F-statistic to decide whether or not to reject the null hypothesis.

The Analysis of Variance ANOVA: We started out learned how to compute mean, variance of one our two populations. The mechanic for computation of such statistics is useful. ANOVA is used to compare several populations means. We will use null hypothesis as a way to compare the say three populations' means as in the following:
H_0: $\mu_1 = \mu_2 = \mu_3$ - The means are equal.
H_a: at least one pair of means is not equal.
We would need to compute the sum of square (SS) within each sample, SSw, and the total SS. SSb is the difference between these two. These statistics that assisted in ANOVA analysis are shown as follows. As usual, we determine the rejection region based on the probability. F statistics must be computed as the ratio between MSb and MSw. If the F statistic falls well inside the rejection region, the null hypothesis will be rejected. The F_α can be looked up with F-table using the numerator DF (Degree of freedom) or denominator DF. The looked up value will be come the boundary value that determine the rejection region. We should note the different between the F-distribution and t or chi-square distribution is that we have 2 parameters describing the F. Student t and chi-square distributions have only one parameter.
F distribution can be used only when the population for each sample is normally distributed with almost identical variance. Also, observations must be independent in all samples. For this hypothesis, we have the following decision rule:
If $F \leq F\alpha$, do not reject H_0
If $F > F\alpha$, reject H_0

Source	DF	SS	MS	F
Treatment/between	k – 1	?	SS/DF	MSb/MSw
Error/Within	n - k		SS/DF	
Total	n - 1	?		

The linear regression model: In basic algebra, we've learned a linear function that has the following form: y = a + bx. This equation is very basic in mathematic that describes the linear relationship between x and y. So, if a set of x and y, (x, y) obeys the above equations, then all the points are lined up in an x-y coordinate.
How would I know if two sets of numbers are linearly related? For example, how would I know if the height and weight of a people are related? In order to determine if two variables are related, we often defined them in a linear relationship. First, we would experiment and then collect data. We draw data on a scatter graph to eyeball it to see if it somehow related. This linear relationship is specified as:
$y = \beta_0 + \beta_1 x$
X and y are two variables that we think are linearly related. The problem is we don't know what β_0 and β_1 like. So, we use our statistic knowledge to estimate these values. β_0 is called the y intercept and β_1 is called the slope. We call x the independent variable or the predictor. It is the basis for the estimation. We call y the dependent variable, the estimated, or the predicted. Because if we know they are related, we can predict y based on x. The y-intercept means if x is zero i.e., the line intercept with y at $\beta_{0.}$ Then slope β_1 tells the directional relationship between x and y. The slope is positive when x increases then, y increases. The slope is negative when x increases then y decreases.
We want to devise an experiment to collect information about x and y so that we can figure out their relationship.

We will employ the five most important activities of statistics as follows: Collect data, Organize data, Present data, Analyze data, and Interpret data. We now draft the scatter graph and eyeball it.

We will use the method of least squares to estimates the linear parameters β_0 and β_1. We needs the following summation using observed data from our sample that we performed the experiment on; Σx, Σx^2, Σy, Σy^2, and Σxy. Well, these can be done very easily.

We then use the following formulas to estimates the parameters; SSxy, SSxx, SSyy. These three statistics are vital in estimating the parameters and inferences. Once we got these figured out, we can use them to estimate the slope parameter β_1 as the ratio of SSxy to SSxx. The y-intercept β_0 can be computed using the two means, x_{bar}, y_{bar} and the slope along with the regression line equation: $\beta_0 = y_{bar} - \beta_1 x_{bar.}$

So, we now have a complete equation for the straight-line statistical model. This is needed for the least squares line, or least squares prediction equation. That means would select a line that has the least error method. The sum of squares error (SSE) can be defined as follows: $\Sigma(y – y_{hat})^2$. We then can compute the variability around the line. This variability determines how good the data fit a straight line . The variance s^2 is SSE divided by degree of freedom, n-2. The Mean Square Error, or MSE is the square root of the variance.

How can we determine if x is related to y? We use the testing hypothesis as following for determining whether the linear model we use is good enough for predicting y from x:

H_0: $\beta_1 = 0$ (no linear relationship)

H_a: $\beta_1 \neq 0$ or H_a: $\beta_1 < 0$ or H_a: $\beta_1 > 0$

The H0 hypothesis states that there are no linear relationship between x and y. The slope in this case is zero. We can also construct a $(1 – \alpha)$ 100% confidence interval (CI) for the slope of β_1.

We introduced two brand new concepts in this area, namely, testing hypothesis and confidence interval. We would like to know overall how strong of a linear relationship between x and y. To compute the relationship strength, we use the coefficient of correlation of a sample, R. It can be computed as follows: SSxy / SQRT(SSyy * SSxx). The value of R is in [-1, 1]. If R = -1, they are perfectly related linearly. All the observed values are line up on the linear model. That would be rare or unusual. However, the slope β_1 is negative. If r = 1, the relationship is also perfect, however the slope is positive. If r = 0, there is no linear relationship whatsoever between x and y. We can use sample R to determine if population tests of linear correlation, H_0 can be rejected. From this we can say that the sign of R is the same as the sign of the slope.

With the population coefficient correlation is ρ, we can use the testing hypothesis to test the strength of linear relationship between x and y.

H_0: $\rho = 0$. There is no relationship between x and y.

H_a: $\rho \neq 0$ or H_a: $\rho < 0$ or H_a: $\rho > 0$. This is for two-tailed and one-tailed test hypothesis.

The next important statistic for the linear regression analysis is the coefficient of determination, R^2. Its range is [0, 1]. This statistic is used to determine how strong x is linearly related to y. One way to think of R^2 is the ratio between unexplained variation and total variation of the distribution. R^2 is preferred over R when regression analysis occurred since it provides the percentage of variation in y-values that can be explained because of variation in x-values. Therefore it provides a measure of association between the two variables. Once the usefulness of the model is determined, we can use the model to predict values of y based on x.

Estimation would involve sampling errors that we must be aware of. We can also determine the confidence interval for the mean value y, the prediction interval for the given x value.