**Module 1: Descriptive Statistics & Regression Analysis**

Find a real-world business example with data set (or example from a case study) and comment on the way hypothesis testing can be applied to answer a relevant business question. Identify the question(s) in your post and provide the data set at an attachment as a reference.

**Hypothesis Testing:**

It is a critical part and uses a systematic approach to assess models through observations

**Business Problem/ Question:**

On an average, will customers (especially students) pay more than $ 5 for Netflix?

**Steps:**

1. **Null Hypothesis**: It is denoted by **Ho**, which is the hypothesis that sample observations result purely from chance. It’s a commonly accepted fact. It shows no significant difference between two or more groups.

* The average value customers are willing to pay for Netflix is smaller than or equal to $5
* **Ho:** µ ≤ 5

1. **Alternative Hypothesis:** It is denoted by **Ha**, which is the hypothesis that sample observations are influenced by some non-random cause. This is the opposite of the **Ho**, supporting a statistically significant result.

* The average value customers are willing to pay for the product is greater than $5
* **Ha:** µ > 5

1. **Significance Level:** It is denoted as **α**. It defines the probability that the null hypothesis will be rejected. There is a relationship between confidence intervals and hypothesis testing. When a 95% confidence interval is made, all values in the interval are considered reasonable values for the parameter being estimated. If the value of the parameter specified by the **Ho** is contained in the 95% interval then the Ho can’t be rejected at the 0.05 level. If the value specified by the **Ho** is not in the interval then it can be rejected at the 0.05 level

* Let’s use a Z-test, we want to be confident of our result, so let us pick as **α = 5%**
* The Ho is µ ≤ 5, so our rejection point will be z0.05 =1.645

1. **Conclusion:** If your p-value meets your significance level requirements, then your **Ha** may be valid and you may reject the **Ho.**

* Given our calculated z is greater than z0.05 =1.645, we have strong evidence to reject the Ho at a 5% significance level. We are then in the likelihood of Ha, that the average value customers are willing to pay for the subscription is greater than $5

**Dataset Reference:** <http://dss.princeton.edu/training/students.xls>

**References:**

[1] Charles Zaiontz, Null and Alternative Hypothesis retrieved from <https://www.real-statistics.com/hypothesis-testing/null-hypothesis/>

[2] Hypothesis Testing retrieved from <https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/>

**Module 2: Inferential Statistics – Hypothesis Testing**

Find a real-world business example with data set (or example from a case study). Load the data set into R and use the descriptive statistics and graphical techniques we have been developing in class to describe the measures of central tendency and measures of variability. What are the unique qualities of the data set? How would you describe the implications of these qualities within the context of the business? Do you notice anything unusual? Provide the data set as an attachment for reference.

**Business Problem/ Question:**

Do Colleges consider only SAT scores for admissions?

**Unique Qualities of the Dataset:**

Age variable is right skewed and SAT score distribution has 2 humps. In real life, we don’t get a standardized distribution and have multiple humps in the dataset

**Steps:**

1. **Null Hypothesis**: It is denoted by **Ho**, which is the hypothesis that sample observations result purely from chance.

* The average SAT score of students is 1750.
* **Ho:** µ = 1750

1. **Alternative Hypothesis:** It is denoted by **Ha**, which is the hypothesis that sample observations are influenced by some non-random cause. This is the opposite of the Ho, supporting a statistically significant result.

* The average SAT score of students is not 1750
* **Ha**: µ != 1750

1. **Significance Level:** It is denoted as α. It defines the probability that the Ho will be rejected. There is a relationship between confidence intervals and hypothesis testing. Assuming Ho to be true.
2. **P-Value:** The p-value obtained is less than the significance level 𝛼 . We therefore reject the Ho and accept the Ha (the negation).
3. **Conclusion**: We can therefore arrive at the following conclusion from this analysis as the average SAT score of students is not 1750.

**Business Implications:**

Students entering this university does not have an average SAT score of 1750. Instead, they have wide range of SAT scores. There are other parameters that college decides for accepting students in the college which is not specified in the dataset

**Unusual Parameters:**

The dataset should have provided other details like undergrad score and overall profile information of the students.

**Dataset Reference:** <http://dss.princeton.edu/training/students.xls>

**Reference:**

[1] Charles Zaiontz, Null and Alternative Hypothesis retrieved from <https://www.real-statistics.com/hypothesis-testing/null-hypothesis/>

**Module 3: Regularization Practices**

Present a real-world example or from a journal paper which entails regularization technique and describe your findings based on the course materials with as much detail as possible. In your response to classmates, comment on an alternative business question that can be answered using the data they provided.

**Business Problem:**

Unemployment is a critical issue for the governments in any country, let’s go through the available regularization techniques for predicting the employment rates.

**Regularization:**

* It is a technique used to avoid the overfitting problem. The idea behind regularization is that models that overfit the data are complex models that have for example too many parameters
* Overfitting can also be controlled by increasing the size of the training dataset. Define a loss or cost function that describes how well the model fits the data. The goal is to find the model that minimizes this loss function
* The idea is to penalize this loss function by adding a complexity term that would give a bigger loss for more complex models. It significantly reduces the variance of the model, without a substantial increase in its bias

**Regularization Techniques:**

* There are few techniques, namely Ridge Regression (L2 Norm), Lasso Regression (L1 Norm), Elastic Net Regression, Dropout, Data Augmentation, Early Stopping, and Ensembling
* In L1, the sum of the absolute values of the weights is imposed as a penalty while in L2, the sum of the squared values of weights is imposed as a penalty. They both differ in the way they assign a penalty to the coefficients
* L1 is usually preferred when we are interested in fitting a linear model with fewer variables. It is also useful when considering a categorical variable with many levels
* L2 is useful when there are a large number of variables with relatively smaller data samples, like in the case of genomic data
* Elastic-net is a compromise between the L1 and L2 regularisation that attempts to shrink and do a sparse selection simultaneously. Lambda is the regularization parameter whose value is optimized for better results

**References:**

[1] Megha Mishra (May 26, 2018), REGULARIZATION: An important concept in Machine Learning, <https://towardsdatascience.com/regularization-an-important-concept-in-machine-learning-5891628907ea>

[2] Saurabh Singh (Oct 8, 2019), Regularisation Techniques in Machine Learning and Deep Learning, <https://medium.com/analytics-vidhya/regularisation-techniques-in-machine-learning-and-deep-learning-8102312e1ef3>

**Module 4: Data Mining**

Identify a problem from your experience that you think would be amenable to data mining. Address the following in your initial post

* Describe the type of data.
* Describe the type of questions you attempt to answer using Data Mining.
* Describe the type of data mining technique (classification, clustering, etc.) and why do you think would be relevant.

**Data:**

Instagram connecting you to certain ads, news, and pages that you open. So, ultimately our feed is stacked and maintained with these interests which doesn’t trigger my interest all the time. There might be many reasons and recommendations which run in the background to make the User stick to the content and the App for more time.

**Questions:**

Data is the next big oil, so the more meaningful information a company can gather that gives it knowledge. This is how a company can establish a competitive advantage which separates it from competition. We can also see the trends and use that for certain business advantages. Some of the major issues in data mining include scalability, privacy provisions, heterogenous, data ownership, complex data, and data quality

**Data Mining Technique:**

Clustering which divides data into similar patterns. If someone links to one page on Instagram, it is bunched into pages of parallel content or connection. Regression is numerical and usually used to measure future data and make certain predictions

**References:**

[1] Larry Alton (December 22, 2017) The 7 Most Important Data Mining Techniques was retrieved from <https://www.datasciencecentral.com/profiles/blogs/the-7-most-important-data-mining-techniques>

[2] Learntek (February 8, 2019) Data Mining Examples and Techniques was retrieved from <https://www.learntek.org/blog/data-mining-examples-and-techniques/>

**Module 5: Time Series Analysis**

Present a real-world example or from a journal paper that entails a time series technique and describe your findings based on the course material. Include either one of the following patterns: trend patterns, cyclical pattern, seasonal patterns or autocorrelation effect in the data sets.

**Time Series Models:**

These are very useful models when we have serially correlated data. Most of the business work on time series data to analyze stock prices, weather, heart rate monitoring, sales number for the next year, web traffic, competition position, and much more.

**Example:**

For example, let’s take our monthly expenses as a time-series, something that can be measured over time. The data should be stationary to make a prediction by building a time series model. The dataset should have properties like mean and variance are constant.

**Seasonal Patterns:**

Seasonal differencing is a method that can be applied by subtracting the previous data point in the same season. In our example, we’re dealing with monthly data, so each year will correspond to a season containing 12 months. So, calculating the seasonal difference for the month of January of any given year means subtracting by current value by the value of January of the previous year.

**ARIMA:**

Once we analyze the data, then we can do modeling. There are 2 models that we will consider for this case, they are Auto Regression (AR) model and also the Moving Average (MA) model. Before we use these models, we have to identify which model that perfectly fits the data based on its autocorrelation and which lag suits the model.

**References:**

[1] Selva Prabhakaran, Time Series Analysis was retrieved from http://r-statistics.co/Time-Series-Analysis-With-R.html

[2] Tavish Srivastava (Dec 2015) A Complete Tutorial on Time Series Modeling in R was retrieved from https://www.analyticsvidhya.com/blog/2015/12/complete-tutorial-time-series-modeling/