**DSCI-5180 Midterm Project.**

**Introduction to Business Decision Making Process**

**Professor-Fouad Hassan Mirzaei**

**Student: Saadana Shyam Sunder**

**Student ID-11724719**

The objective of this project is to conduct a comprehensive statistical analysis of a dataset comprising information about the top 1000 YouTubers. The dataset includes various variables, such as "Rank," "Username," "Categories," "Subscribers," "Country," "Visits," "Likes," "Comments,”.

This Data set was obtained from **Kaggle website**  Dataset link – <https://www.kaggle.com/datasets/mabelhsu/api-clean-top-1000-youtubers-statistics>**.**

**Module 1 Normal distribution**

To pinpoint YouTubers who are exceptionally successful in terms of their subscriber count, enabling us to recognize the top performers in the dataset, we are finding z-score

Considering the dataset of 1000 youtubers are normally distributed

find What is the probability for number of subscribers greater than 20million ?

**Solution:**

**A screenshot of a table

Description automatically generated**

To find the probability that the average number of subscribers greater than 20 million is 54percent(0.547159)

Formula in excel = NORM.DIST(x,mean,standard deviation, culminative)

X=20million

mean =21998700

culminative =true

Standard deviation **=**16868481

P(x>20000000)=1-p(x<20000000)=1-0.452841=0.547159

**The probability that the average number of subscribers greater than 20 million is 54percent(0.547159)**

**Module 2 Estimation**

we aim to determine the 95% confidence interval for the average number of subscribers. This interval will help us understand the range within which we can reasonably expect the true population average of subscribers to fall, with a high level of confidence (95% confidence level). This information is vital for making statistically sound decisions, evaluating the performance of YouTubers, and gaining insights into the overall subscription trends within this dataset.

What is the 95% confidence interval for the average number of 'subscribers' received by the top 1000 YouTubers?

**Solution**

Since this is the sample of 1000 youtubers among all the youtubers we use t table

Xbar= 21998700

sample standard deviation= 16868481

n= 1000

s of x-bar=533428.21

t(alpha/2)= 1.962



**We are 95 percent confidence that average number of subscribers is between lower limit *20951930* and upper limit *23045470***

**Module 3 Hypothesis Testing**

We are conducting this analysis to investigate whether there exists a significant difference in the average number of 'comments' received by YouTubers across different categories. Our analysis is structured around the following hypotheses:

Null Hypothesis (H0): The average number of comments for YouTubers across different categories is equal to 2400(H0: μ = 2400). Alternative Hypothesis (H1): The average number of comments for YouTubers across different categories is greater than 2400 (H1: μ ≠ 2400).

**Solution**

- H0: μ=2400 (Null Hypothesis)  
- Ha: μ > 2400(Alternative Hypothesis Right Tailed Test)  
- No. of Samples taken n=217  
- Level of test for α=0.05,  
- Critical value of t statistic for Right-side alternative is z=1.917  
- Standard Deviation σ = 8893.766  
- Mean x-bar = 2408.018  
- Reject Null Hypothesis H0 if the test statistic is less than the level of significance (0.05)  
- t-value= 1.917, P value: 0.49  
Since P-value is more than α (0.49>0.05)



**We cannot Reject the Null Hypothesis H0, as there is no sufficient evidence at the α=0.05 level of significance that the average number of comments for YouTubers across different categories is greater than 2400**

**Module 4 multiple linear regression analysis**

"What is the individual impact of subscribers , likes, and comments on the number of views/visits gained by YouTubers, and how do these variables collectively contribute to the growth of views/visits ?"

Null Hypothesis (H0):The number of subscribers, likes, and comments collectively does not significantly influence the number of views received by YouTubers.

Alternative Hypothesis (H1):At least one of the independent variables (subscribers, likes, comments) has a significant impact on the number of views received by YouTubers.

* We can able to Predict the number of visits with 59700000subscribers, 156500 likes, and 4200 comments using the multiple regression equation

This question prompts an investigation into the relationship between the independent variables (subscribers, likes, and comments) and the dependent variable (number of views/visits), aiming to understand how changes in each independent variable affect the number of views/visits gained by YouTubers. The goal is to quantify and assess the relative influence or contribution of subscribers, likes, and comments on the growth of views/visits in the YouTube channel**.**(we are analyzing number of views/visits because youtubers can determine their advertisement sponsors with this value)

**Solution:**

Dependent Variable: "number of views/visits is the dependent variable for this analysis.

Independent Variables: The independent variables are "Number of subscribers", "Number of likes", "number of comments” of youtuber across different category

Multiple Regression Model: The multiple regression model can be formulated as: number of views/visits= β0 + β1 \* subscribers+ β2 \* likes+ β3 \* comments.

Below is the Regression Statistics Summary for the data:



Using the above summary, we calculate the below:

Regression model equation:

**Estimated number of visits= (-428620) + (0.00911535) \* subscribers+ (21.6189123) \* likes + (72.033931) \* comments**

**Regression Analysis Interpretation**:

* + Intercept (β0): The intercept of -428,620 represents the expected number of views/visits when the independent variables (subscribers, likes, comments) are all zero. In practical terms, this value might not hold much meaning since having zero subscribers, likes, or comments is unlikely for a YouTube video.
  + Coefficients (β1, β2, β3):
    - For every unit increase in subscribers, the number of views/visits is expected to increase by 0.0091, holding other variables constant.
    - Similarly, for every additional like, there's an expected increase of 21.62 views/visits, keeping other variables constant.
    - Lastly, each comment might contribute to an increase of approximately 72.03 views/visits, holding other variables constant.
  + Overall Model Significance: The overall model, considering all three predictors (subscribers, likes, comments), appears statistically significant (p-value much smaller than 0.05). This means that at least one of these independent variables has a significant impact on the number of views/visits.

**Correlation Analysis:**

* + Checking the correlations between variables can provide insights into how they relate to each other.
    - There's a positive correlation between the number of likes and the number of views, indicating that as likes increase, views tend to increase as well.
    - Similarly, comments might also show a positive correlation with views, suggesting a relationship where more comments correspond to more views.
    - The correlation between subscribers and views might not necessarily be as strong as likes or comments, as subscriber count doesn't always directly translate to immediate views. However, there might still be a positive correlation.

Test Statistic = 2517.96902

P-Value = 5.28E-166

The p value (5.28E-166) is smaller than the alpha 0.05, we can infer that there is substantial evidence, at the 0.05 level of the significance, which supports the assertion that the linear relationship between the dependent and independent variables is statistically significant.(i.e) we can reject the null hypothesis and infer that At least one of the independent variables (subscribers, likes, comments) has a significant impact on the number of views received by YouTubers.

* Predicting the number of visits with 59700000subscribers, 156500 likes, and 4200 comments using the multiple regression equation

Substituting the values in the equation:

Estimated number of visits/views= (-428620) + (0.00911535) \* subscribers+ (21.6189123) \* likes + (72.033931) \* comments

= -428620+59700000\*0.00911535+21.6189\*156500+72.03339\*4200

=3801464.48

**Therefore, the estimated number of visits with subscribers, likes, and comments using the multiple regression equation is 3801464.48 visits**