جامعة الأخوين

ⵜⴰⵙⴷⴰⵡⵉⵜ ⵏ ⵓⵅⴰⵡⵉⵏ

AL AKHAWAYN UNIVERSITY

**Fall 2022**

**CSC 5341**

**Paper Report**

# Graduate Admission: A Statistical Approach


| Full Name | ID |
|---|---|
| **Yassir Benabdallah** | 76865 |


**Supervised By:**
**Dr. Amine AMAR**

# Table of Contents

# I. Introduction

The discussion over students' academic performance is always a topic that is under debate. All students care about maintaining a good academic performance because it is considered as a major factor in getting accepted by the most prestigious universities for post-graduate studies or getting hired by great companies. Also, as students, we cannot deny that having good grades in all the subjects is a key factor in achieving success. However, there are many other variables that help in determining students' future other than their GPA.

This paper aims to help students in shortlisting their choices when applying for a university or a college to complete their graduate studies. Therefore, we will try to discover to what extent the academic performance affects the chances of students getting admission by a university of their choice. Not only that, but we will also try to identify other factors that increase or decrease the chances for getting accepted.

# II. Dataset Description and Methodology

The dataset of this project consists of 500 examples that represent different students. Each one has their own Graduate Record Examination (GRE) score, Test of English as a Foreign Language (TOEFEL) score, university rating, Standard Operating Procedure (SOP), Letter of Recommendation (LOR) score, Cumulative Grade Point Average (CGPA) and Research.

The described dataset is analyzed through multiple statistical methods, which are Hypothesis Testing I-sample, Hypothesis Testing II-sample, Correlation and Regression. This variousness of methods is intended to output several results that will accurately decide on which feature is important in affecting the graduate admission.

The implementation of theses method will be conducted on Python language, one of the leading programming languages in carrying statistical computing and graphics.

# III. Experiments and Results

## 1. Hypothesis Testing Type 1

This test consists of the following hypothesis: reject the null hypothesis when it is true. This test will be conducted on the GRE score. Also, we are using Z-test, as a statistical test for which the distribution of the test statistic under the null hypothesis can be approximated by a normal distribution. The reason behind opting for this kind of tests is that the variance of the dataset is unknown and also the number of samples is greater than 30.

Our interest here is to determine whether there is an evidence with a significance level of 0.05 to support the fact that the mean GRE score is different from 310.

We can state the hypothesis along with other measurements as follows:

- **Null Hypothesis: $H_0$**: $\mu = 310$ → population mean of GRE score is equal to 310.
- **Alternative Hypothesis: $H_1$:** $\mu \neq 310$ → population mean of GRE score is not equal to 310.
- **Significance Level**: $\alpha = 0.05$

- **Z-Scores:** $z_{0.025} = 1.96$ and $-z_{0.025} = -1.96$

```
# n is the length of the sample
n <- length(df$GRE.Score)
print(n)
```

```
[1] 501
```

```
# s = sample variance
s <- var(df$GRE.Score)
print(s)
```

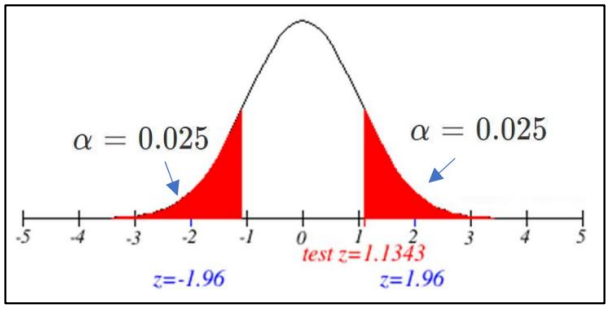
```
[1] 127.3252
```

```
# Z = test statistic
Z <- (mean(df$GRE.Score)-310)/(s/sqrt(n))
print(Z)
```

```
[1] 1.13774
```

```
# p-value of test statistic
P = 2*pnorm(-abs(Z))
print(P)
```

```
[1] 0.2552291
```




We can notice that the Test Statistic Z = 1.13774 is less than the Z-Score $z_{0.025} = 1.95$ and greater than the Z-Score $-z_{0.025} = -1.95$, which makes it fall in the region where the null hypothesis is failed to be rejected. Therefore, we cannot conclude that the mean GRE Score of a student is differs from 310 at significance level of 0.05.

## 2. Hypothesis Testing Type 2

It is also important to consider other variables that contribute in students being admissible for pursuing their graduate studies. Hypothesis testing type 2 discusses whether there is any difference between the mean of chance of admission of students with research experience and the students without research under the t-test of 5% significance level.

I spilt the dataset into two groups based on whether the student has research experience or not. Also, I assumed that the variances of two groups are not equal since I cannot reasonably predict that both of groups of data are having the same variances.

We identify the following variables:

**$\mu_1$** = the sample mean of chance of admission of students who have research experience.

**$\mu_2$** = the sample mean of chance of admission of students who do not have research experience.

**n1** = the number of students who have research experience.

**n2** = the number of students who do not have research experience.

**var1** = the sample variance of chance of admission of students who have research experience.

**var2** = the sample variance of chance of admission of students who do not have research experience.

**$t_0$** = test statistic.

**v** = degree of freedom.

We can state the hypothesis as follows:

- **Null Hypothesis: $H_0$:** $\mu_1 = \mu_2$ → the sample mean of chance of admission of students who have research experience is equal to the sample mean of chance of admission of students who do not have research experience.
- **Alternative Hypothesis: $H_1$:** $\mu_1 \neq \mu_2$ → the sample mean of chance of admission of students who have research experience is not equal to the sample mean of chance of admission of students who do not have research experience.

```
# withResearch = table with students having researches
# withoutResearch = table with students having no researches

withResearch = df[df["Research"] == 1]
withoutResearch = df[df["Research"] == 0]
```

```
# n1 = number of students having researches
n1 = len(withResearch)
print(n1)
```

```
280
```

```
# n2 = number of students having no researches
n2 = len(withoutResearch)
print(n2)
```

```
220
```

```
x1 = withResearch["Chance of Admit "].mean()
print(x1)
```

```
0.7899642857142857
```

```
x2 = withoutResearch["Chance of Admit "].mean()
print(x2)
```

```
0.634909090909091
```

```
var1 = withResearch["Chance of Admit "].var()
print(var1)
```

```
0.01518028545826933
```

```
var2 = withoutResearch["Chance of Admit "].var()
print(var1)
```

```
t0 = (x1 - x2 - 0)/(sqrt((var1/n1) + (var2/n2)))
print(t0)
```

```
14.707274979628917
```

```
v = ((var1/n1) + (var2/n2)) *
      ((var1/n1) + (var2/n2)) /
        ((((var1/n1)*(var1/n1)) / (n1-1)) + (((var2/n2)*(var2/n2)) / (n2-1)))
print(v)
```

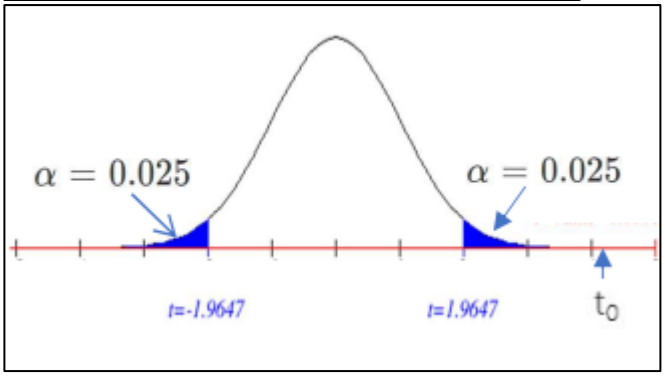
```
487.6049791861994
```

```
np.floor(v)
```

```
487.0
```

```
alpha = 0.05
t_alpha = stats.t.ppf(alpha/2, np.floor(v))
print(t_alpha)
```

```
-1.9648471009786466
```



The above lines of code output the following results:

- **Test statistic: $t_0$** = 14.707274979628917
- **Significance level: $\alpha$** = 0.05
- **Degree of freedom: v** = 487

According to hypothesis test type 2, the null hypothesis will be rejected if $t_0 < -t_{0.024,\ 487} = -1.965$ or if $t_0 > t_{0.025,\ 487} = 1.965$.

We have t0 = 14.7 which is bigger than $t_{0.025,\ 487} = 1.965$. Therefore, we reject the null hypothesis.

Another method can be implemented to confirm the result found before. We use the p-value. The following code shows the result of the p-value:

```
from scipy import stats

# Perform the t-test
t, p = stats.ttest_ind(withResearch["Chance of Admit "], withoutResearch["Chance of Admit "])

# Print the p-value
print(p)
```

```
3.5954935458409782e-40
```

The above result show that p-value = $3.51e^{-40}$ which less than the significance level $\alpha$ = 0.05. Therefore, the null hypothesis is rejected again.

Based on the calculations from two different methods, it appears that there is a significant difference in the mean chance of admission for students with and without research experience. The results of the statistical calculation show that students with research experience have a higher mean chance of admission. Therefore, it can be concluded that students with research experience are more likely to be admitted to a university compared to students without research experience.

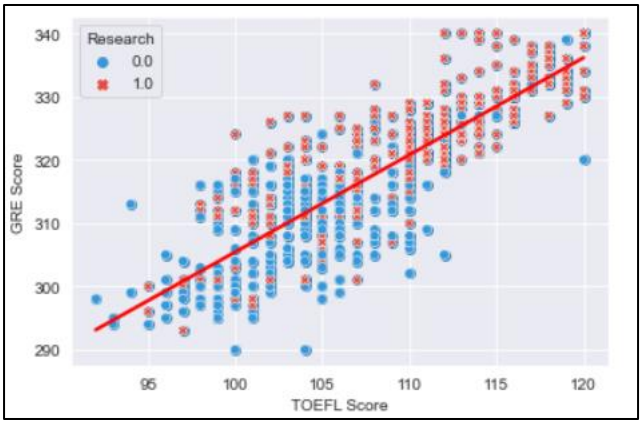
## 3. Correlation

### a. Classical Correlation Approach

A correlation test is a statistical test that is used to determine the strength and direction of the relationship between two variables. Correlation tests are used to determine whether a change in one variable is associated with a change in another variable.

There several types of correlation test, and this test relies on Pearson's correlation coefficient, which is the most commonly used correlation test. It measures the strength and direction of the linear relationship between two variables. Let us study the relationship between the TOEFL score than the GRE score using correlation.

```
# Set the Seaborn style
sns.set_style("darkgrid")

# Scatterplot with regression line and confidence interval
sns.regplot(x="TOEFL Score", y="GRE Score", data=df,
            line_kws={"color": "red"}, ci=True)

# Add Pearson's correlation coefficient and p-value to the plot
sns.scatterplot(x="TOEFL Score", y="GRE Score", data=df,
                hue="Research", style="Research",
                palette=["#3498db", "#e74c3c"])

# Set the x-axis label
plt.xlabel("TOEFL Score")

# Set the y-axis label
plt.ylabel("GRE Score")

# Show the plot
plt.show()
```




The following screenshot shows that the correlation coefficient is r = 0.82.

```
# Calculate the correlation coefficient
r, p = stats.pearsonr(df1["TOEFL Score"], df1['GRE Score'])

# Print the correlation coefficient
print(r)
```

```
0.827200403531721
```

The analysis of the data shows that there is a strong positive linear relationship between TOEFL score and GRE score. This is demonstrated by the high correlation coefficient of 0.8272 and the scatter plot of the data, which shows a clear upward trend. The results suggest that as TOEFL score increases, GRE score also tends to increase. This indicates a relatively strong positive association between the two variables.

### b. Significance Test for Correlation

After theoretically proving that the there is a strong linear relationship between the TOEFL score and the GRE score, let us support this fact statistically. A significance test is conducted at significance level of 0.05.

We can state the hypothesis as follows:

- $H_0$: $\rho = 0$ → there is no linear correlation between the TOEFL score and the GRE score.
- $H_1$: $\rho \neq 0$ → there is a linear correlation between the TOEFL score and the GRE score.

The following screenshot shows that the test statistic is $t_0$ = -309.63

```
from scipy import stats

# Perform the t-test
t, p = stats.ttest_ind(withResearch["TOEFL Score"], withoutResearch["GRE Score"])

# Print the t-value
print(t)
```

```
-309.6310088455176
```

The statistic test value $t_0 < -t_{0.025,\ 487.6} = -1.9647$ which yields the rejection of the null hypothesis. Therefore, there is sufficient evidence of linear relationship between TOEFL Score and GRE Score of students at the 5% level of significance.

To summarize, research has shown that there is a strong positive correlation between a student's CGPA and their GRE score, as well as their TOEFL score.

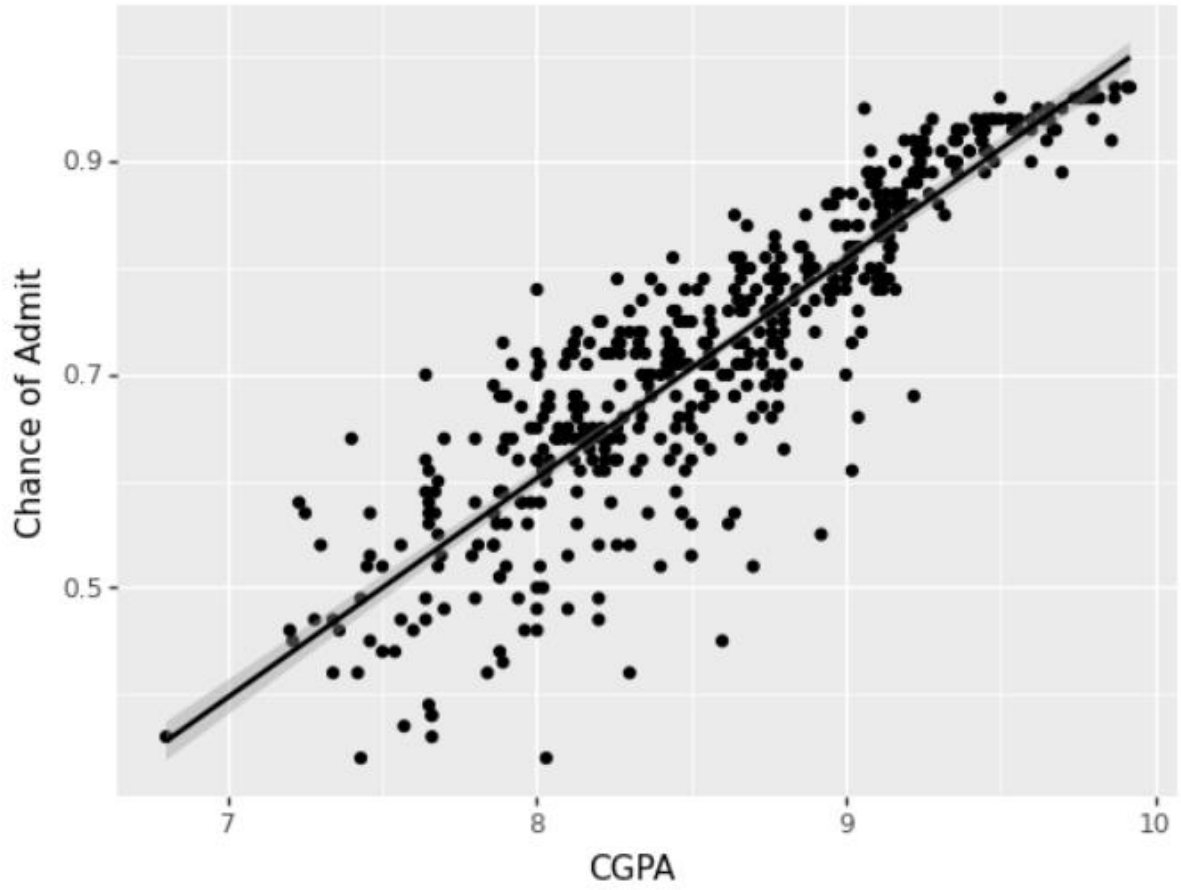
## 4. Regression

### a. Classical and R-Squared Approach

Regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It allows you to make predictions about the dependent variable based on the values of the independent variables.

In this analysis, we are examining the relationship between the dependent variable, Chance of Admit, and the independent variable, CGPA. We are testing whether there is a linear relationship between these two variables, with the goal of using this relationship to make predictions about Chance of Admit based on the value of CGPA.

```
from plotnine import *

(ggplot(df1, aes(x='CGPA', y="Chance of Admit "))
 + geom_point()
 + geom_smooth(method='lm'))
```




Based on the graph, the regression model involves only one independent variable and is referred to as simple regression. The relationship between the variables appears to be positive and linear, and the equation of the model is represented as y = -1.0443 + 0.2059x (please refer to the following screenshot that shows how we got the coefficient and the intercept). This equation can be used to make predictions about the dependent variable based on the value of the independent variable.

```
from sklearn.linear_model import LinearRegression

# Split the data into independent and dependent variables
X = df1[['CGPA']]
y = df1["Chance of Admit "]

# Create a Linear Regression model
model = LinearRegression()

# Train the model on the data
model.fit(X, y)

# Print the model coefficients
print(f'Intercept: {model.intercept_}')
print(f'Coefficient: {model.coef_[0]}')
```

```
Intercept: -1.044334923899038
Coefficient: 0.2059216789132832
```

The coefficient of the intercept, or b0, in the regression model is equal to -1.0443. This value represents the estimated average value of the dependent variable, Chance of Admit, when the value of the independent variable, CGPA, is 0. It is important to note that in this case, a value of 0 for CGPA may not be realistic, as it would likely indicate that the student did not participate in any compulsory academic activities. Therefore, the value of b0 may not have a meaningful interpretation in this context.

The coefficient of the slope, or b1, in the regression model is equal to 0.2059. This value represents the average change in the dependent variable, Chance of Admit, that is associated with a one-unit change in the independent variable, CGPA. In this case, an increase of one unit in CGPA is associated with an average increase of 0.2059 units in Chance of Admit. This suggests that there is a relationship between the two variables, as indicated by the calculated value of b1.

Also, the R-Squared value can give us an idea about how our regression model is performing. Basically, R-squared is a statistical measure that represents the proportion of the variance in the dependent variable that is explained by the independent variables in a regression model. It ranges from 0 to 1, with higher values indicating a better fit. The following lines of code output the value of $R^2$.

```
import statsmodels.formula.api as smf

# Fit the model
model = smf.ols(formula='COA ~ CGPA', data=df1)
results = model.fit()

# Get the R-squared value
r2 = results.rsquared
print(f'R-squared: {r2:.2f}')
```

```
R-squared: 0.78
```

R-squared = 0.78 for this model. This means that the independent variable (CGPA) is able to explain around 78% of the variation in the dependent variable (Chance of Admit). This suggests that there is a weaker linear relationship between the two variables, as not all of the variation in the dependent variable is explained by the independent variable.

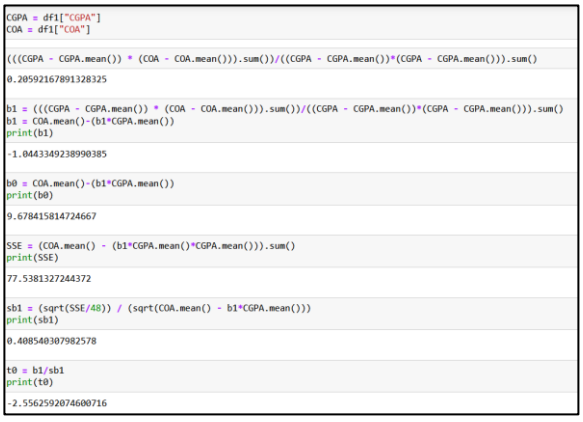
### b. Inference about the Slope.

Regression analysis has been conducted to further test the possibility of a linear relationship between variable x and y.

We can state the hypothesis as follows:

- $H_0$: $\beta_1 = 0$ → There is no linear relationship between CGPA and Chance of Admit
- $H_1$: $\beta_1 \neq 0$ → There exists a linear relationship between CGPA and Chance of Admit

The following screenshots show some statistical results. On the left we concluded that the test statistic value $t_0$ = -2.56, while on the right p-values = $1.76 \times 10^{-88}$ and $3.3 \times 10^{-165}$.



```
CGPA = df1["CGPA"]
COA = df1["COA"]

(((CGPA - CGPA.mean()) * (COA - COA.mean())).sum())/((CGPA - CGPA.mean())*(CGPA - CGPA.mean())).sum()

0.20592167891328325

b1 = (((CGPA - CGPA.mean()) * (COA - COA.mean())).sum())/((CGPA - CGPA.mean())*(CGPA - CGPA.mean())).sum()
b1 = COA.mean()-(b1*CGPA.mean())
print(b1)

-1.0443349238990385

b0 = COA.mean()-(b1*CGPA.mean())
print(b0)

9.678415814724667

SSE = (COA.mean() - (b1*CGPA.mean()*CGPA.mean())).sum()
print(SSE)

77.5381327244372

sb1 = (sqrt(SSE/48)) / (sqrt(COA.mean() - b1*CGPA.mean()))
print(sb1)

0.408540307982578

t0 = b1/sb1
print(t0)

-2.5562592074600716
```

```
results.pvalues

Intercept     1.761419e-88
CGPA         3.396545e-165
dtype: float64
```

Again, both p-values are less than the significance value 0.05, which allows us to reject the null hypothesis. Based on the statistical test, it is likely that CGPA has a significant effect on a student's likelihood of being admitted.

## IV. Conclusion

To sum up, academic success plays a significant role in the admissions process for universities. While other factors may also be considered, the results of this study demonstrate that a student's academic performance is a major factor in determining their chances of being accepted. It is important for students to prioritize their academic performance while also participating in extracurricular activities, as these can help develop skills and abilities. Ultimately, a well-rounded student who excels both academically and in extracurricular pursuits is likely to have the best chances of gaining admission to a university. In short, academic performance and graduate admissions are closely connected.

# References


Mohan, A. (November, 2018). Graduate Admission 2. Version 2. Retrieved December, 02 2022 from https://www.kaggle.com/datasets/mohansacharya/graduate-admissions?select=Admission_Predict_Ver1.1.csv

Bhandari, P. (2022, November 11). Type I & type II errors: Differences, examples, visualizations. Scribbr. Retrieved December 16, 2022, from https://www.scribbr.com/statistics/type-i-and-type-ii-errors/

James, A. (2020). Correlation. JMP. Retrieved December 16, 2022, from https://www.jmp.com/en_ca/statistics-knowledge-portal/what-is-correlation.html#:~:text=What%20is%20correlation%3F,statement%20about%20cause%20and%20effect.

Beers, B. (2022, September 9). What is regression? definition, calculation, and example. Investopedia. Retrieved December 16, 2022, from https://www.investopedia.com/terms/r/regression.asp