Inferential Statistics Report

1.1 Introduction[1]

This report will investigate the median starting salary for graduates from different colleges based of a dataset. The reason of me investigating this dataset is to find out graduates from which major earns the most for their starting salary. Since engineering school is still having the period of major selection, it is relevant to know which major is the best.

I am interested in using this dataset to answer the research questions shown below and using appropriate applications of statistical knowledge we learnt in CA07 to make conclusions based on my calculations.

2.1 Dataset

The dataset is based on the data provided from the Wall Street Journal (https://www.kaggle.com/wsj/college-salaries) with 270 sets of values per variable. It includes salaries of the type of major, starting median salary, mid-career median salary, the name of university, and more.

2.2 Research questions:

1) What is the starting median salary for graduates? Construct a confidence interval of 95% and explain what it means analytically.
2) Is there a **statistical** or **practical significant** difference between the starting median salary of Engineering major compared to starting median salary of Liberal arts major? (take alpha = 0.05) (t-test, n<30)
3) Is it true that on average, having a low median starting salary means we will have a low mid-career median salary?
4) What type of logic is used when making conclusions based on statistics

2.3 variables

Research question 2:

- Independent variable:
    - The type of major in university
    - This is a **qualitative** variable which means the variable cannot be counted. In addition, it is also a **nominal** variable as the variable can be categorised into different major types.
    - Since this is a qualitative, nominal variable, we need to use a bar graph to visualise the data specifically[2]

[1] #professionalism: I have included headers to make reading my work much easier since I did not do very well for last assignment. I divided the assignment into six sections, and labelled each subsection.
[2] #variables: I identified the independent variable as the type of major in university and stated the nature of the variable as qualitative and nominal. I further explained that we need to use a bar graph to visualise the data as it is a qualitative and nominal variable.

- Dependent variable:
    - the starting median salary
    - This is a **quantitative** variable as it has a numerical value assigned to it for the salary. It is also a **discrete** variable as all the numbers are whole numbers with separate values and no decimal place

Research question 3:

- Independent variable:
    - the starting median salary
    - This is a **quantitative** variable as it has a numerical value assigned to it for the salary. It is also a **discrete** variable as all the numbers are whole numbers with separate values and no decimal place
- Dependent variable:
    - the mid-career median salary
    - This is a **quantitative** variable as it has a numerical value assigned to it for the salary. It is also a **discrete** variable as all the numbers are whole numbers with separate values and no decimal place

Confounding variable:

- The location of their job- if someone works in a city, they will tend to get paid more due to the higher cost of living in the central business district (CBD). This variable is not intentionally studied but will influence the outcome of the experiment, thus it is a confounding variable
- The work schedule of the job. Generally, people with longer work hours tend to earn more if they are in the same sector of work, so the study did not take into account the schedule for work and it will affect the starting salary of the graduate from college.[3]

3.0 Methods

The dataset was fed into Python using Pandas for me to tabulate and analyse the data. Firstly, I will tabulate the data and display the summary of statics of the data. Then I will display the sample distributions of the variables as shown in the figures below. Those distributions were created using python and the details are shown in appendix B.

3.1 Research question 1

For this question, we will first need to check if the data fits the assumptions of a normal distribution. The point estimate of the sampling distribution is nearly normal, and assumed the estimate is unbiased from independent individual observations. We can verify this by looking into the data collection method, and if it is from a simple random method, then it is independently observed. However, we also know that there is no perfect case of sampling, and data is more likely to be collected through convenience sampling, which might result in skew. The sample size is relatively large > 30, we can allow slight to moderate skew for our distribution. This satisfies the conditions to

[3] #variables: I have identified possible confounding variables that might influence the data collected. For example, the location and work schedule of the job might affect the salary of the job.

apply central limit theorem for our data and we can approximate the data using a normal distribution.

To compute the confidence interval, I created a simple algorithm so that it is easy to follow[4]:

1) Identify the parameters of the distribution, n=269
2) Calculate mean(μ) and standard deviation(∂) from the given set of values (starting median salary):
    a. $\mu = \frac{\sum x}{n}$
    b. $\sigma = \sqrt{\sum \frac{(x-mean)^2}{n}}$
3) Calculate the standard error of the sample proportion,
    a. $SE = \frac{\sigma}{\sqrt{n}}$
4) Calculate the z-score that corresponds to 95% confidence interval
    a. Z= stats.binom.isf()
5) Calculate the Margin of Error (MOE)
    a. MOE=Z* SE
6) Finally, we can compute the confidence interval (CI) of the unbiased and nearly normal point estimate as:
    a. Confidence interval: [Mean – MOE, Mean + MOE]

3.2 Research question 2:

To answer this question, I calculate some summary statistics for starting median salary for engineering major and liberal arts major

Table 1

Summary statistics for starting median salary ($) for engineering major and liberal arts major		
	Engineering Major	Liberal Arts Major
Sample size	19	47
Mean	59057.89	45746.81
Standard deviation	7633.74	4322.12

3.2.1 statistical significance

To calculate the statistical significance of the two groups of academic major graduates, I must use a t-distribution as the sample size is small, n <30. I will run a hypothesis test to compute the statistical significance. I assumed that the observations are independent and nearly normal when recording the data. For this hypothesis test, I will use the default alpha value, which is 0.05[5].

Hypothesis test algorithm:

[4] #algorithms: I applied algorithmic thinking to solve the problem of calculating the confidence interval in an orderly fashion

[5] #descriptivestats: I calculated the descriptive stats and place them into table 1. Since the sample size is small, I interpreted that we need to use a T distribution to calculate the statistical significance.

1) Null hypothesis: the starting median salary of engineering major graduates and liberal arts graduates are the same.
    a. H0: $\mu_E = \mu_A$
    b. null value= 0, difference between both means should be 0 if they are equal
2) Alternative hypothesis: the starting median salary of engineering major graduates are not the same as liberal arts graduates
    a. Ha: $\mu_E \neq \mu_A$
3) Compute the point estimate of interest and the degrees of freedom (DOF):
    a. Point estimate = $\mu_E - \mu_A$
    b. Degrees of freedom = smaller sample size -1
4) Calculate the sample difference of two means (standard error) using standard deviation estimates based on the two samples
    a. $SE_{\mu_E - \mu_A} = \sqrt{SE_{\mu_E}^2 + SE_{\mu_A}^2} = \sqrt{\left(\frac{SD_E^2}{n_E}\right) + \left(\frac{SD_A^2}{n_A}\right)}$
5) Once we obtained the standard error, DOF, and point estimate, we can use it to calculate the T-score from the t distribution
    a. $T = \frac{point\ estimate - null\ value}{SE_{\mu_E - \mu_A}}$
6) Calculate the p value associated with the T-score and compare it with the value of alpha. Then, draw conclusions from the comparison whether to accept or reject the hypothesis.

3.2.2 Practical significance

If we know the statistical significance, it tells us nothing about the effect size between the two data and how they differ from each other. Therefore, we need to calculate practical significance to compare how the difference between the two means are, expressed in the pooled standard deviation.

In this case, I will use Hedge's g to calculate the effect size as in comparison to Cohen's D and Glass's Delta, it contains the standardised effect size with correct upward bias of Cohen's d. This is evident in small sample size and will make the value for practical significance more accurate.

The steps to calculate the effect size are as shown below

1) Compute the pooled standard error
    a. $pooled\ SD = \sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{n_1 + n_2 - 2}}$
2) Calculate Cohen's d using the formula:
    a. $Cohen's\ d = \frac{\mu_E - \mu_A}{pooled\ SD}$
3) Hedge's g formula
    a. Hedge's g = $Cohen's\ d * (1 - \frac{3}{4(n_1 + n_2) - 9})$

3.3 Research question 3

To answer this question, we need to plot both variables on a regression line to see if there is a clear correlation between the dependent and independent variable.

To plot the regression line between starting median salary and mid-career median salary, we need to check if it satisfies the assumptions as listed below:

1) It is a linear relationship between the explanatory and response variable
2) There is nearly normal residuals as outliers can greatly affect the regression line
3) There should be constant variability for the points, and the least square line should be able to show the variability
4) The variables are recorded using independent observations

For this case, we can show that it satisfies the assumptions by plotting the residual plot to show random distribution and histograms of residuals. By plotting the distribution and histogram, we can see if it is suitable to use a regression line. Then, we can compute the r-squared value and draw conclusions from the value to answer this question

4.0 Results and conclusions

4.1 Research Question 1

Using my confidence interval algorithm:

1) Identify the parameters of the distribution, n=269
2) Calculate mean(μ) and standard deviation(∂) from the given set of values (starting median salary):
    a. $\mu = \frac{\sum x}{n} = 46068.4$
    b. $\sigma = \sqrt{\sum \frac{(x-mean)^2}{n}} = 6400.6$
3) Calculate the standard error of the sample proportion,
    a. $SE = \frac{\sigma}{\sqrt{n}} = \frac{6210}{\sqrt{268}} = 390.3$
4) Calculate the z-score that corresponds to 95% confidence interval
    a. Z= stats.binom.isf(0.95) = 1.96
5) Calculate the Margin of Error (MOE), MOE= Z*SE
    a. MOE=1.96 x 390.3 = 764.9
6) Finally, we can compute the confidence interval (CI) of the unbiased and nearly normal point estimate as:
    a. Mean +- MOE, CI: [45303.5,46833.3]

We will first compute the z score in relation to 95% for normal distribution using stats.norm.isf(0.95) which means stats.norm.ppf(0.025), which is 1.96. This means that the confidence interval will be the x value associated with z score of -1.95 to the x value associated with the z score of 1.96.

This means that analytically, we are 95% confident that the population mean of the starting median salary is between $45304 and $46833. The visualisation of the confidence interval is shown in figure 1 below beautifully using python matplotlib package.[6]

Figure 1:

[6] #confidenceintervals: Using my algorithm shown in steps 1 to 6, I calculated the confidence interval of 95% in python and interpreted what it means to have a confidence interval of 95%.

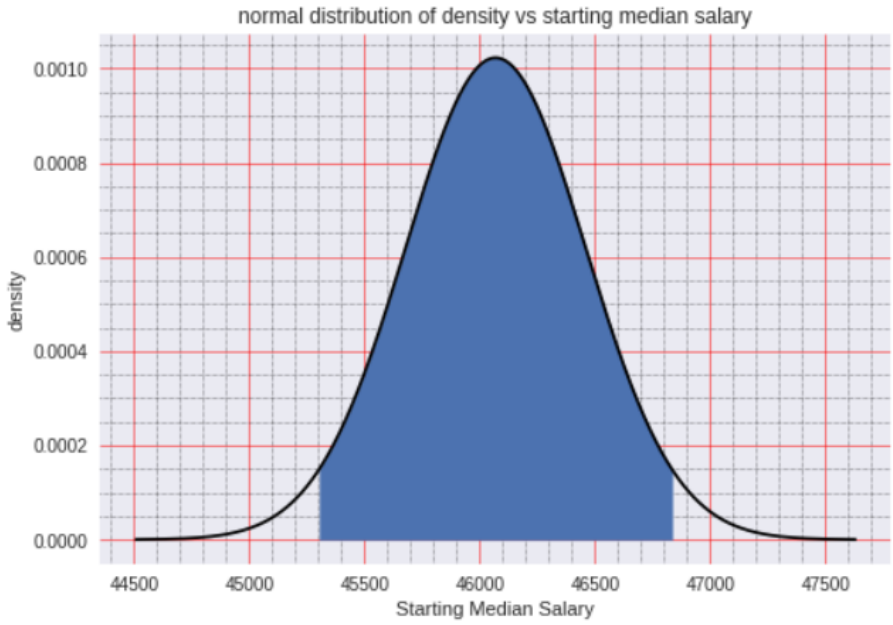


Figure 1 is a visual representation of the probability density normal distribution[7] for the starting median salary of fresh graduates from colleges. The shaded blue region represents the 95% confidence interval.[8]

4.2 Research Question 2

4.2.1 Statistical significance

Hypothesis test algorithm:

1) Null hypothesis: the starting median salary of engineering major graduates and liberal arts graduates are the same.
    a. H0: $\mu_E = \mu_A$
    b. null value= 0, difference between both means should be 0 if they are equal
2) Alternative hypothesis: the starting median salary of engineering major graduates are not the same as liberal arts graduates
    a. Ha: $\mu_E \neq \mu_A$
3) Compute the point estimate of interest and the degrees of freedom (DOF):
    a. Point estimate = $\mu_E - \mu_A = 59057.89 - 45746.81 = \mathbf{13311.09}$
    b. Degrees of freedom = smaller sample size -1 = 19-1 = **18**
4) Calculate the sample difference of two means (standard error) using standard deviation estimates based on the two samples
    a. $SE_{\mu_E - \mu_A} = \sqrt{SE_{\mu_E}^2 + SE_{\mu_A}^2} = \sqrt{\left(\frac{SD_E^2}{n_E}\right) + \left(\frac{SD_A^2}{n_A}\right)}$

$$= \sqrt{\left(\frac{7633.74^2}{19}\right) + \left(\frac{4322.13^2}{47}\right)}$$

$$= \mathbf{1861.3213}$$

[7] #distributions: I used a normal distribution to approximate the large sample size based from the central limit theorem, and displayed it in figure 1

[8] #dataviz: I represented the calculated confidence interval using python beautifully in the blue shaded region of the normal distribution.

5) Once we obtained the standard error, DOF, and point estimate, we can use it to calculate the T-score from the t distribution
    a. $T = \frac{point\ estimate - null\ value}{SE_{\mu_E - \mu_A}} = \frac{13311.086 - 0}{1861.32} = \mathbf{7.15}$

6) Calculate the p value associated with the T-score and compare it with the value of alpha.
    a. This is a two tails test, so $num\ of\ tails = 2$
    b. $p - value = num\ of\ tails * \left(1 - stats.t.cdf(7.15,18)\right) = \mathbf{1.164 * 10^{-6}}$
    c. $Recap: \alpha = 0.05$

Form the p value= $\mathbf{1.164 * 10^{-6}}$, we can conclude that since it is very small (approximately 0) and it is smaller than α, the null hypothesis is rejected. This means that the starting median salary of engineering major is statistically different from the starting median salary of liberal arts major[9]. The difference is displayed in the figure 3 and figure 4 below.

Figure 3

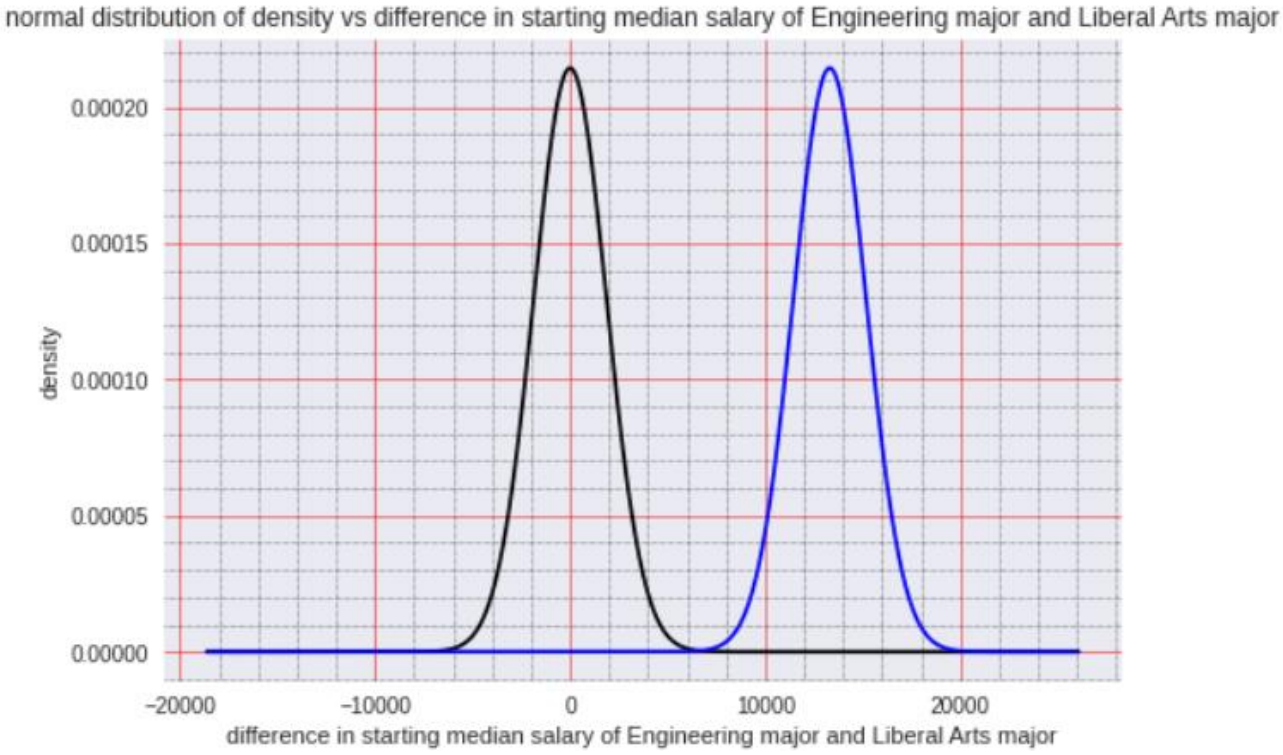


Figure 3 is a visual representation of the hypothesis test that we calculated above for the statistical significance. The black graph represents the distribution of the null value, which is there is no difference between the starting median salary of Engineering major and Liberal Arts major. The blue graph represents the distribution of the point estimate, which centre is at 13311. There is a red shaded area that represents the p-value if the null hypothesis is true given that the sample distribution is at the point estimate. Since the p value is so small, we cannot see the shaded red region on figure 3, so I decided to zoom the graph so that it is visible.

[9] #significance: I calculated the statistical significance for my hypothesis test to verify if engineering major graduates and liberal art major graduates have the same starting median salary. Since the p-value is smaller than the significance value, I rejected the null hypothesis and concluded that the engineering major and liberal arts major have statistically different starting median salaries.

Figure 4

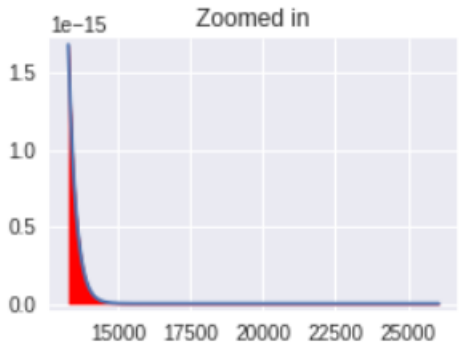


Figure 4 is a super zoomed in section of figure 3 to show that the shaded red region is extremely small, and therefore the p-value is very very small[10]

As mentioned earlier, the p-value is shown as the shaded red region in figure 3, but since it is so small, the graph needed to be zoomed for it to be visible. The P-value is $\mathbf{1.164 * 10^{-6}}$, and the point estimate is to the right of the null value. Since α =0.05, it means that we are 95% confident that the starting salary of Engineering major is different from the starting salary of Liberal Arts major, and Engineering major graduates earn more than Liberal Arts major graduates.

4.2.2 Practical Significance

The steps to calculate the effect size:

1) Compute the pooled standard error
   a. $pooledSD = \sqrt{\frac{(n_1-1)SD_1^2+(n_2-1)SD_2^2}{n_1+n_2-2}} = \sqrt{\frac{((19-1)*7633.74^2+(47-1)*4322.13^2)}{19+47-2}}$

      $= 5460.43$
2) Calculate Cohen's d using the formula:
   a. $Cohen's\ d = \frac{\mu_E-\mu_A}{pooled\ SD} = \frac{59057.89-45746.81}{5460.43} = 2.4377$
3) Hedge's g formula
   a. Hedge's g = $Cohen's\ d * \left(1 - \frac{3}{4(n_1+n_2)-9}\right) = 2.4377 * \left(1 - \frac{3}{4(19+47)-9}\right) = 2.41$

Since the Hedge's g is 2.41 (more than 0.8), we can conclude that the difference between the two means, engineering major median starting salary and Liberal Arts major median starting salary have a **large effect size**. In the context of this dataset, it means that the difference between the two means is 2.41 pooled standard deviations away from each other. This implies that the starting median salary for Engineering major graduates have practically large significance compared to Liberal Arts major graduates.[11]

[10] #dataviz: figures 3 and 4 is used to represent the calculated statistical significance on the distribution. Since the shared region for p-value cannot be seen in figure 3, I used python to zoom into a smaller section of the graph to show that the p-value is very very small.

[11] #significance: I calculated the practical significance between engineering major starting salary and liberal art starting salary, and applied hedge's g formula as it contains the standardised effect size with correct upward bias of Cohen's d. Since the effect size is 2.41, I concluded that it has a large effect.

4.3 Research Question 3

First, we will plot the three plots:

1) Scatter plot of mid-career median salary vs starting median salary (figure 5)
2) Residual plot of the residuals (figure 6)
3) Histogram of residuals (figure 7)

Figure 5

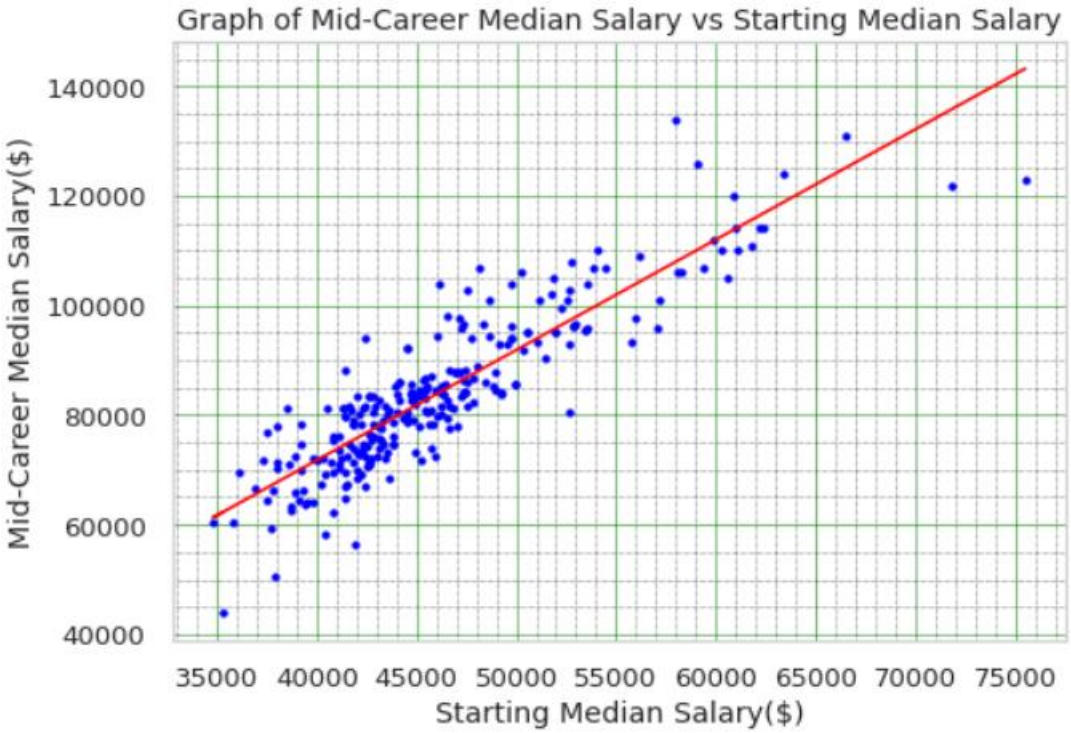


Figure 5 is a scatterplot between the mid-career median salary and starting median salary. There is a regression line plotted in the scatterplot, shown in red.[12]

Figure 6

[12] #dataviz: figures 5,6,7 are data visualisations that I used to show that the data fits the necessary conditions to be displayed on a scatterplot, which is having constant variability between residuals and a histogram of approximately normally distributed residuals

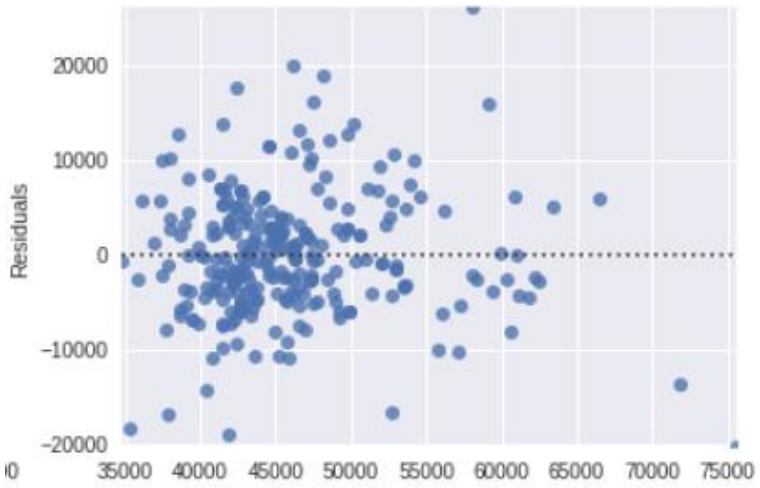


Figure 6 is a residual plot that shows the distribution of residuals. We can observe random variability between all residuals of the variables.

Figure 7

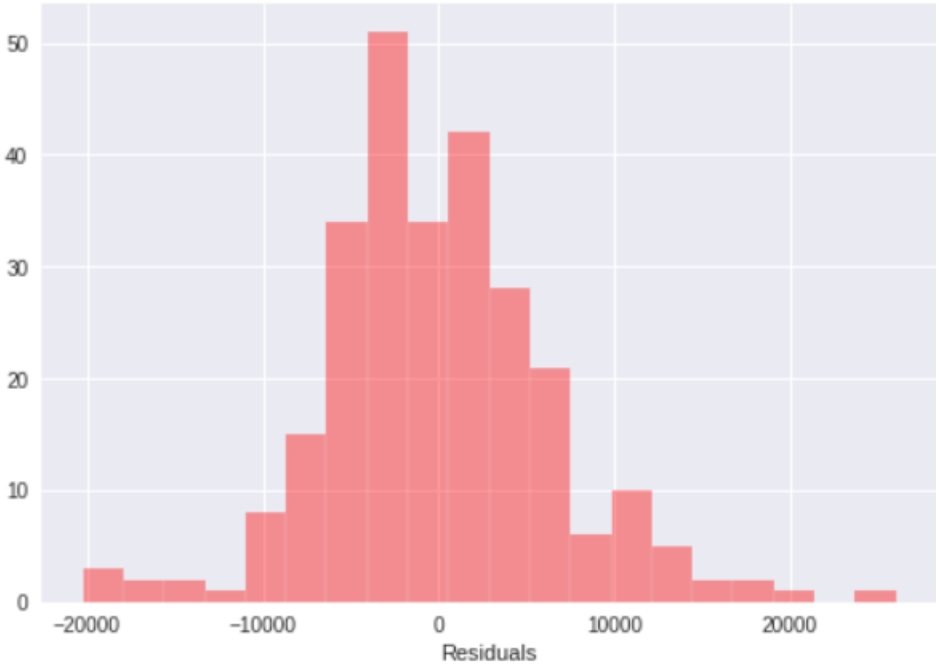


Figure 7 shows the histogram of residuals for the scatterplot. We can observe an approximately normal distribution for the residuals with peak at approximately 0.

From figure 6, we can observe that the residual plot shows constant variability between the points of residuals dispersed around the plot. This means that a linear regression model is most appropriate for the data that is measured. In figure 7, the histogram of residuals also shows an approximately normal distribution. This implies that the variance is normally distributed and implicates a linear football shaped homoscedastic scatterplot, which is shown in figure 5. As a result, the least squares line is appropriate to fit the points in the scatterplot. The scatterplot also shows a linear relationship between the x and y variables, and the observations are assumed to be independent, so the regression line is plotted in figure 5 as the assumptions are satisfied.

For the scatterplot, the R-squared value is computed using python and it is 0.787. This means that 78.7% of the variability of the mid-career salary can be explained by the starting median salary. The

other 21.3% of the variability might be dependent on other factors such as working location, working hours, type of job, and so on. From the scatterplot, the R-value is 0.887, which also shows that the data shows strong correlation between the variables.

The regression equation is : y=2.014 x - 8818.102

The y intercept implies that there will be -8818.102 for mid-career salary if starting median salary is 0. Of course, this does not make any sense as it is only used for adjusting the regression line and it is meaningless.

The slope of the regression equation is 2.104, which implies that for every one dollar increase in starting median salary, there will be an increase of 2.104 for mid-career median salary.

To answer the question, the scatterplot shows that 78.7% of the variability can be explained by starting median salary, and if one have a low median starting salary, it is 78.7% likely that they will also obtain a low mid-career median salary solely from their starting median salary. [13]

4.4 Research Question 4

The logic used to make conclusions for previous questions are inductive, as the premises only imply the conclusion and does not definitely show the conclusion. This is because we did statistical calculations and conclude based on the predicted probability of the event occurring. Since it is a probability, it is not 100% certain that it will happen, and still have a chance of the opposite result happening. However, the probability of that is very small, for example, in research question 3, the probability of the Engineering major graduates and Liberal Arts graduates earning the same starting median salary is 0.00000116, which is really small, and smaller than alpha value, therefore we concluded that they are not equal based on our calculations.[14]

5.0 References

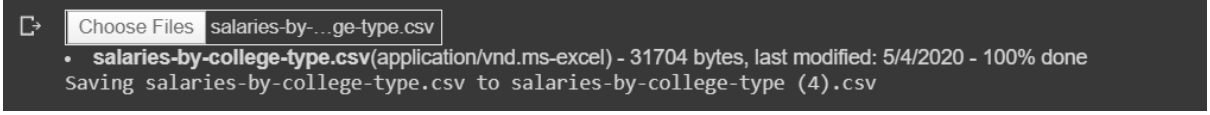
1) 2012, The Wall Street Journal, *Where it Pays to Attend College*, Kaggle, retrieved from: https://www.kaggle.com/wsj/college-salaries

6.0 Appendix

All calculations for statistics and visualisations are all done using python, and the code is shown below.

Part 1: importing CSV

```
[ ]  from google.colab import files
     uploaded=files.upload()
```



[13] #regression: I applied and interpreted what it means by having an R-squared value of 78.8, and shown analytically what the regression line equation means, then used that to answer research question 4 as the regression line shows that lower starting median salary means lower mid-career median salary.
[14] #induction: I applied inductive knowledge to make conclusions using probability and statistics.

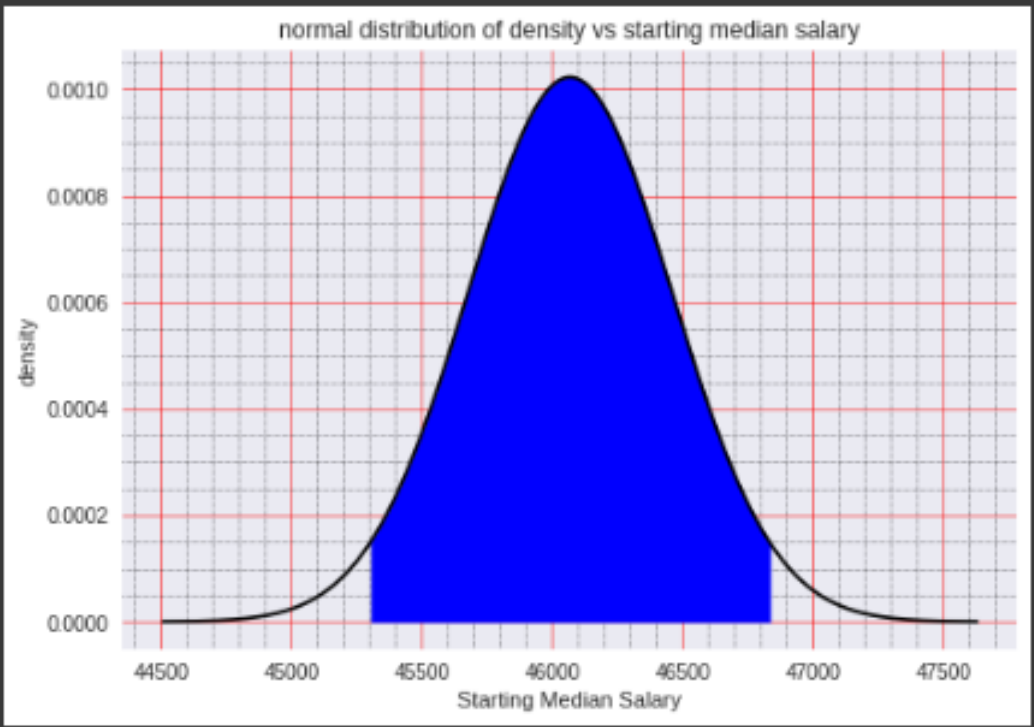
Part 2: interpreting data

```
import csv
import matplotlib.pyplot as plt
import pandas as pd
import scipy.stats as stats
from scipy.stats import norm
from statistics import mean
import numpy as np

df=pd.read_csv(io.BytesIO(uploaded['salaries-by-college-type.csv']),nrows=269)

x_label='Starting Median Salary'
y_label='Mid-Career Median Salary'
x=df[x_label]
y=df[y_label]
x1=[]
y1=[]

for i in range (269):
    x1.append(int(x[i][1:-7]+x[i][-6:-3]))
for i in range (269):
    y1.append(int(y[i][1:-7]+y[i][-6:-3]))

xn=np.array(x1,dtype=np.float64)
x_mean=np.mean(xn)

yn=np.array(y1,dtype=np.float64)
y_mean=np.mean(yn)
```

Part 3: calculating confidence interval

```
[ ] from matplotlib import style
    #select a beautiful style to use
    style.use('seaborn')

    #initialising variables
    meu=x_mean #change these variables later
    SD=np.std(xn)
    SE=SD/np.sqrt(269)


    def confidence_interval(xbar,SE,percentage):
        global z
        k=(1-percentage)/2
        z=stats.norm.isf(k)
        #inverse survival function is finding the z score when we know what is 1-cdf
        lowbound = xbar - z*SE
        highbound = xbar + z*SE
        return lowbound,highbound,z*SE
    k=confidence_interval(meu,SE,0.95)


    #shading the region for confidence interval
    x=np.arange(meu-z*SE,meu+z*SE,0.001)#enter the range of x for the shaded region
    y=norm.pdf(x,meu,SE)#find the y value for the corresponding x value in the z score
```

```
#build the plot
#setting the domain of the plot,-4SE to 4SE away from mean
domain=np.linspace(meu-4*SE,meu+4*SE,1000)
plt.plot(domain, norm.pdf(domain,meu, SE),color="black")
plt.title("normal distribution of density vs starting median salary")
plt.xlabel(x_label)
plt.ylabel("density")

#add the shaded area to represent probability
plt.fill_between(x,y,color='b')

#turning on minor ticks on the distribution for better visualisation
plt.minorticks_on()

#having a grid on the distribution
plt.grid(which='major',linestyle='-',linewidth='0.5',color='red')
plt.grid(which='minor',linestyle=':',linewidth='0.5',color='black')
plt.show()

print("the point estimate is :", meu)
print("the standard deviation is :", SD)
print("the standard error is :", SE)
print("The z score is: ", z)
print("the margin of error is : ",  k[2] )
print("the confidence interval is :",k[0],k[1])
print("The confidence interval is 95%")
```




```
the point estimate is : 46068.40148698885
the standard deviation is : 6400.68578244928
the standard error is : 390.25669983769586
The z score is:  1.959963984540054
the margin of error is :  764.8890764073423
the confidence interval is : 45303.512410581505 46833.29056339619
The confidence interval is 95%
```

Part 4: statistical significance test

```
#statistical significance test
#initialise variables
x_label='Starting Median Salary'
x=df[x_label]
x1=[]
engineering=[]
liberal_arts=[]
for i in range (0,268):
    x1.append(int(x[i][1:-7]+x[i][-6:-3]))

xn=np.array(x1,dtype=np.float64)

#find the values for engineering major
for j in range (19):
  engineering.append(int(x[j][1:-7]+x[j][-6:-3]))
#print(engineering)
#find mean of engineering major
engineering_mean=np.mean(engineering)
#find standard deviation of salary
engineering_SD=np.std(engineering)
#find sample size
engineering_size=int(len(engineering))

#find values for liberal arts major
for j in range (39,86):
  liberal_arts.append(int(x[j][1:-7]+x[j][-6:-3]))
#print(liberal_arts)
#find mean of liberal arts major
arts_mean=np.mean(liberal_arts)
#find standard deviation of salary
arts_SD=np.std(liberal_arts)
#find sample size
arts_size=int(len(liberal_arts))


print("the mean of engineering major starting median salary is : ", engineering_mean)
print("the standard deviation for engineering major starting median salary is : ", engineering_SD)
print("the sample size for engineering graduates is : ", engineering_size)

print("the mean of Liberal Arts major starting median salary is : ", arts_mean)
print("the standard deviation for Liberal Arts major starting median salary is ; ", arts_SD)
print("the sample size for Liberal Arts graduates is : ",arts_size)
```

```
the mean of engineering major starting median salary is :  59057.89473684211
the standard deviation for engineering major starting median salary is :  7633.741980215117
the sample size for engineering graduates is :  19
the mean of Liberal Arts major starting median salary is :  45746.8085106383
the standard deviation for Liberal Arts major starting median salary is ;  4322.127680522185
the sample size for Liberal Arts graduates is :  47
```

```
[ ]  #two tailed test
     tails=2
     print("the null hypothesis is mean of starting median salary engineering =\
      liberal")
     print("Ha: the means are different")
     null_value= 0
     point_estimate= engineering_mean-arts_mean
     print("PE : ",point_estimate)
     DOF = engineering_size-1
     engineering_SE= engineering_SD/ np.sqrt(engineering_size)
     print("Eng SE : ",engineering_SE)
     arts_SE= arts_SD/ np.sqrt(arts_size)
     print("Art SE : ",arts_SE)
     SE=np.sqrt(engineering_SE**2+arts_SE**2)
     print("pooled SE : ",SE)
     T=(point_estimate - null_value)/SE
     print("T-value : ",T)
     print("p-value : ",tails*(1-stats.t.cdf(T,DOF)))
```

```
the null hypothesis is mean of starting median salary engineering = liberal
Ha: the means are different
PE :  13311.086226203806
Eng SE :  1751.3005185695663
Art SE :  630.4471173721832
pooled SE :  1861.3213247972333
T-value :  7.151417677790736
p-value :  1.164461445490872e-06
```

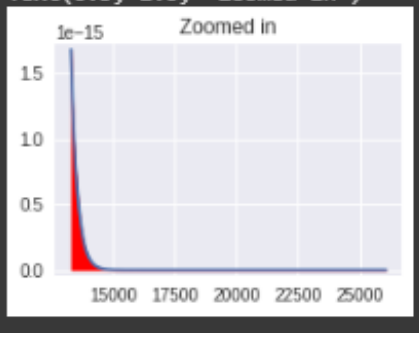
Part 5: plotting for statistical significance

```
[ ]  #shading the region
     x=np.arange(null_value+T*SE,null_value+14*SE,0.001)#enter the range of x for the shaded region
     y=norm.pdf(x,null_value,SE)#find the y value for the corresponding x value in the z score
     x_label="difference in starting median salary of Engineering major and Liberal Arts major"
     #build the plot
     #setting the domain of the plot,-4SE to 4SE away from mean
     domain=np.linspace(null_value-8*SE,null_value+12*SE,1000)
     plt.plot(domain, norm.pdf(domain,null_value, SE),color="black")
     plt.plot(domain, norm.pdf(domain,point_estimate, SE),color="blue")
     plt.title("normal distribution of density vs difference in starting median \
     salary of Engineering major and Liberal Arts major")
     plt.xlabel(x_label)
     plt.ylabel("density")


     #add the shaded area to represent probability
     plt.fill_between(x,y,color='r')

     #turning on minor ticks on the distribution for better visualisation
ZXC  plt.minorticks_on()

     #having a grid on the distribution
     plt.grid(which='major',linestyle='-',linewidth='0.5',color='red')
     plt.grid(which='minor',linestyle=':',linewidth='0.5',color='black')
     plt.show()
```

Part 6: zoom in plot

```
[ ]  #zooming in the graph
     ax2 = plt.subplot(222)
     plt.fill_between(x,y,color='r')
     ax2.margins(x=0.05, y=0.025)
     ax2.plot(x, y)
     ax2.set_title('Zoomed in')
```

```
Text(0.5, 1.0, 'Zoomed in')
```



Part 7: calculating statistical significance

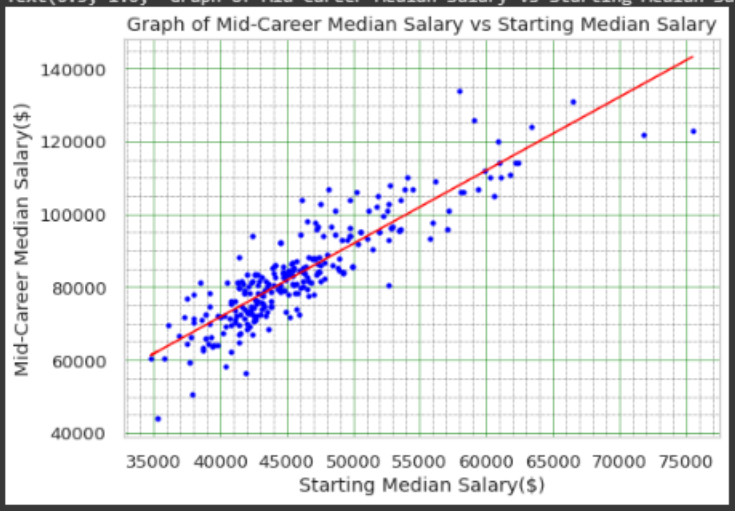
```
[ ]  #caluclating practical significance
     pooled_sd=np.sqrt(((engineering_size-1)*engineering_SD**2+(arts_size-1)*\
                        arts_SD**2)/(engineering_size + arts_size-2))
     print("the pooled standard deviation is : ", pooled_sd)

     #cohen's d calculation
     cohenD=(engineering_mean-arts_mean)/pooled_sd
     print("the cohen's d is : ", cohenD)

     #hedge's g calculation
     hedgeG=cohenD*(1-3/(4*(engineering_size+arts_size)-9))
     print("the hedge's g is : ", hedgeG)
```

```
the pooled standard deviation is :  5460.438015796608
the cohen's d is :  2.437732318853525
the hedge's g is :  2.409053115102307
```

Part 8: plotting regression line

```
[ ]  import seaborn as sns # very nice plotting package

     # style settings
     sns.set(color_codes=True, font_scale = 1.2)
     sns.set_style("whitegrid")

     def best_fit_slope_and_intercept(xn,yn):
         m = ((x_mean*y_mean)-mean(xn*yn))/(((x_mean)**2)-mean(xn**2))
         b= y_mean - m*mean(xn)
         return(m,b)
     m,b = best_fit_slope_and_intercept(xn,yn)
     regression_line=[(m*x1)+b for x1 in xn]

     print(plt.title("Graph of "+ y_label +" vs "+x_label  ))

     plt.ylabel(y_label+"($)")
     plt.xlabel(x_label+"($)")
     plt.minorticks_on()
     plt.grid(which='major',linestyle='-',linewidth='0.5',color='green')
     plt.grid(which='minor',linestyle=':',linewidth='0.5',color='black')
     plt.scatter(x1,y1,s=10,color='blue')
     plt.plot(xn,regression_line, color='red')
     plt.show()
```

```
Text(0.5, 1.0, 'Graph of Mid-Career Median Salary vs Starting Median Salary')
```



Part 9: Plotting residual plot and histogram of residuals

```
import statsmodels.api as statsmodels # useful stats package with regression functions
def regression_model(column_x, column_y):
    # this function uses built in library functions to create a scatter plot,
    # plots of the residuals, compute R-squared, and display the regression eqn

    # fit the regression line using "statsmodels" library:
    X = statsmodels.add_constant(column_x)
    Y = column_y
    regressionmodel = statsmodels.OLS(Y,X).fit() #OLS = "ordinary least squares"

    # extract regression parameters from model, rounded to 3 decimal places:
    Rsquared = round(regressionmodel.rsquared,3)
    slope = round(regressionmodel.params[1],3)
    intercept = round(regressionmodel.params[0],3)

    # make plots:
    fig, (ax1, ax2) = plt.subplots(ncols=2, sharex=True, figsize=(12,4))
    sns.regplot(x=column_x, y=column_y, data=df, marker="+", ax=ax1) # scatter plot
    sns.residplot(x=column_x, y=column_y, data=df, ax=ax2) # residual plot
    ax2.set(ylabel='Residuals')
    ax2.set_ylim(min(regressionmodel.resid)-1,max(regressionmodel.resid)+1)
    plt.figure() # histogram
    sns.distplot(regressionmodel.resid, kde=False, axlabel='Residuals', color='red')

    # print the results:
    print("R-squared = ",Rsquared)
    print("Regression equation: y =",slope, "x + ",intercept)
regression_model(xn, yn)
```

```
R-squared =  0.787
Regression equation: y = 2.014 x +  -8818.102
```