STATISTICAL INFERENCE:
ASSIGNMENT REPORT

CS50: Formal Analysis
Minerva Schools at KGI

Khanh Nguyen
December 2018

STATISTICAL INFERENCE

i. Introduction

This report is based on a dataset about food choice preferences of college students ("Food choices and preferences of college students", n.d.).

The question of exploration for this dataset is: "Does eating veggies affect GPA of college students?". The motivation for this research question is the fact that I am a vegetarian, so I eat veggies very often.

I perform a 2-tailed hypothesis test to determine the statistical significance among the GPA of students who eat veggies frequently and those who do not. In other words, is there convincing evidence that students who eat more veggies have higher or lower GPA than those who do not?

ii. Dataset

This dataset is obtained from Kaggle database with questions about college students eating habits, such as preferences for a cuisine, frequency of eating out, or information about their nutrition. There are a total of 125 respondents in this dataset. The data can be found here.

I'm using this sample data to estimate whether there is a difference in GPA among students who eat veggies to those not do so frequently. Therefore, the only 2 variables of interest is the value of GPA and the likelihood of eating veggies. The value of GPA is a numerical continuous variable, which is based on a 4.0 scale. The value of the likelihood of eating veggies is an ordinal variable with the choice from 1 to 5. 1 indicates the unlikelihood of eating veggies, and 5 means very likely to eat veggies in a day. While GPA is a quantitative variable, the likelihood of eating veggies is a qualitative variable. [1]

As there are only 125 students surveyed, I manually clean the data with removing non-numerical value and mixture of both numerical and non-numerical value (i.e. 3.69 and some comments

[1] #variables: GPA is a continuous variable because GPA is computed from many courses and can take any value from 0 to 4.0. Because it shows the result of mathematical operation and has a mathematical meaning, GPA is a quantitative variable. In contrast, the likelihood of eating veggies is an ordinal qualitative variable because the number from 1 to 5 is just a method to show the degree of students' behavior, with 1 is very unlikely and 5 is very likely. Taken out of context, 1 to 5 doesn't have any mathematical meaning.

about it) in the GPA domain. In contrast, there is no missing value in the self-report of likelihood of eating veggies. As the research question explores the relationship between eating veggies preference and GPA, I deleted any rows of data not reflecting these 2 values. Therefore, the dataset has only 121 results left.

iii. Methods

The software I used to analyze the dataset is Google Spreadsheet for drafting and manipulating data (dividing the GPA data into 2 subsets), Excel for importing data to Python, and Python for mathematical operations and visualization. The reason why I use Google Spreadsheet and Excel simultaneously is that Google Spreadsheet is easier to draft visualization and perform mathematical operations. In addition, Google Spreadsheet saves changes on the cloud as long as I am connected with the internet, whereas Excel has a disadvantage of manual saving requirement.

For Python, I used the pandas package to import the data, numpy for manipulating the list and generate descriptive statistics, and matplotlib to generate visual graph.

a) Preliminary analysis

Here is an overview description of the statistics I'm exploring[2] (table 1). Full calculation can be found in appendix B.

	GPA	Likelihood of eating veggies
Count	n = 121	n = 121
Mean	$\bar{x}_{GPA}$ = 3.419	$\bar{x}_{veggies}$ = 4.025
Median	3.5	4
Mode	3.5	5
Standard deviation	s_{GPA} = 0.388	$s_{veggies}$ = 1.048
Range	1.8	4

Table 1: Overview description of GPA and likelihood of eating veggies among college students

[2] #descriptivestats: This is the description of the statistics of 2 variables I'm exploring.

The GPA variable has a large sample size (n = 121. Its mean and median (3.419 and 3.5) are quite close to each other, but because the GPA is on a 4.0 scale, the distribution is skewed to the left[3] (figure 1 - can be found in appendix C).

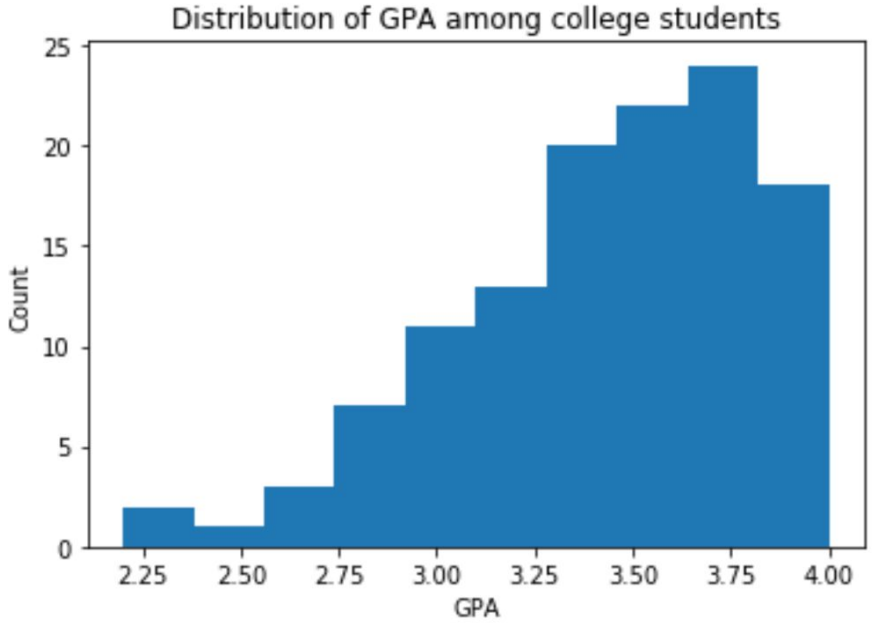


Figure 1: Distribution of GPA among college students

Similarly, the likelihood of eating veggies variable also has large sample size (n = 121) and similar mean and median (4.025 and 4). The mode is 5 while this variable is also on a scale of 5. Therefore, the distribution of this variable is skewed to the left (figure 2 - can be found in appendix C).

[3] #dataviz: There is 4 histograms in this paper to describe the distribution of the variables of interest.

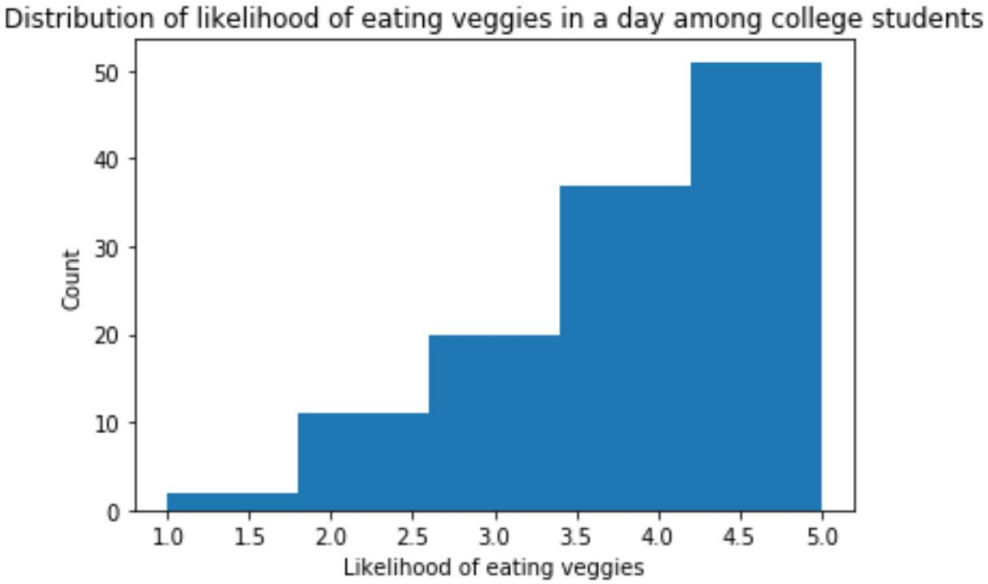


Figure 2: Distribution of likelihood of eating veggies among college students

To estimate the GPA of college students, I perform a confidence interval operation on the mean of GPA of this sample. I use the t-statistics to construct the confidence interval of 95% because this sample meets the following condition:

- Random: I assume that the experiment to collect this sample was done randomly
- Normal: The sample size is large (bigger than 30) with a normal distribution
- Independent: The sample is independent from the population, assuming that the sample size is less than 10% of the whole college students population

In addition, the t-statistics is used instead of the z-statistics because I'm using the standard deviation from the sample, not from the population. Following this, the standard error is computed by the sample standard deviation divided by the square root of sample size (n = 121 according to table 1). Detailed calculation of this confidence interval construction can be found in appendix D[4].

b) Full-scale analysis

Based on the likelihood of eating veggies, I divided the GPA data into 2 subsets: 1 subset is "1-very unlikely" or "2 unlikely" to eat veggies, and the other subset is "4-likely" or "5-very likely" to eat veggies.

[4] #confidenceintervals: This is the part that I constructed the confidence interval for the mean of GPA variable.

An overview of the 2 subset can be summarized in the table below (refer to appendix B for calculation)

	GPA of group unlikely eating veggies	GPA of group likely eating veggies
Count	n = 13	n = 88
Mean	$\bar{x}_{GPA}$ = 3.295	$\bar{x}_{GPA}$ = 3.449
Median	3.4	3.5
Mode	None	3.5
Range	1.72	1.6
Standard deviation	$s_{less\ veggies}$ = 0.534	s_{GPA} = 0.358

Table 2: Descriptive statistics of 2 subsets: GPA of the group eating less veggies and GPA of the group eating more veggies

According to table 2, the group of students eating less veggies is quite small (n = 13).The distribution as figure 3 below (appendix C) is skewed to the left with an outlier.

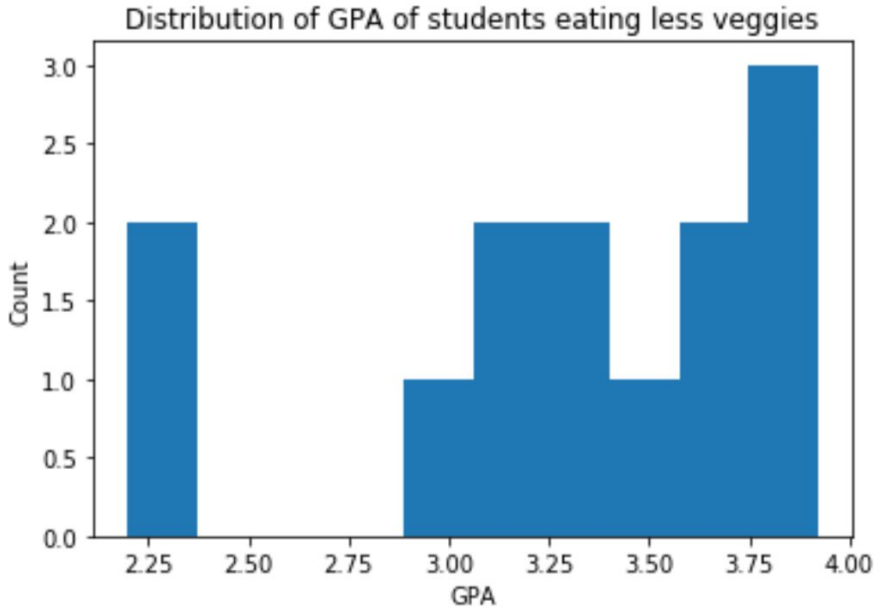


Figure 3: Distribution of GPA of students eating less veggies

Meanwhile, the group of students eating more veggies is larger (n = 88). The mean and median of the subset is also very close to one another with little standard deviation. This results in the skewed to the left distribution as figure 4 below (appendix C)[5]:

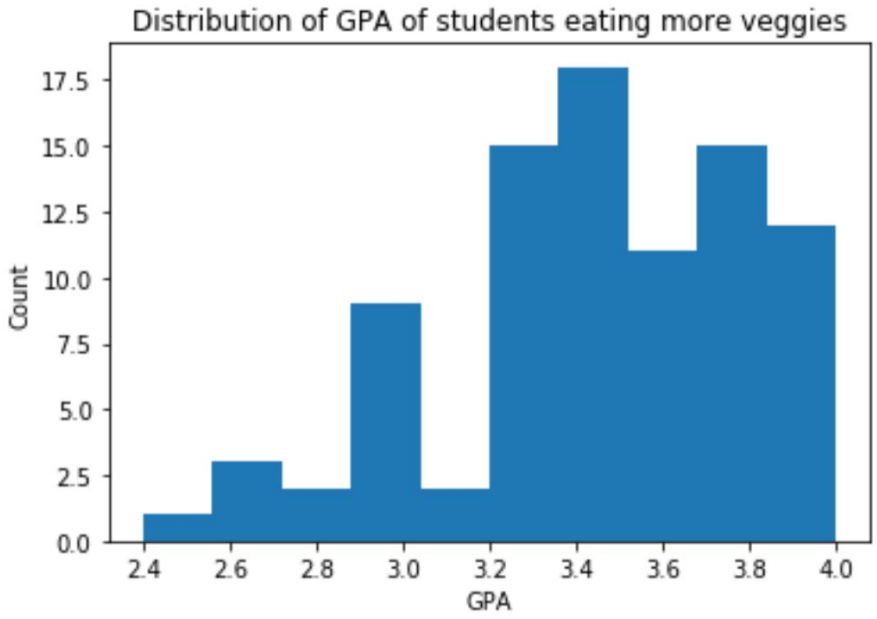


Figure 4: Distribution of GPA of students eating more veggies

To answer the research question: Is there a difference in GPA between students eating more or less veggies, I'm conducting a hypothesis test (t-test). This is a 2-tailed test because I am interested in finding whether there is any difference in both of the spectrum. The significance level is 5% (alpha = 0.05) as a rule of thumb.

The null and alternative hypothesis is as follow:

- Null hypothesis: There is no difference in GPA between students eating less or more veggies.
- Alternative hypothesis: There is a difference in GPA between students eating less or more veggies.

I am using the "difference between 2 means" method to test this hypothesis. The standard error is computed by taking the square root of the sum of the variance divided by the sample size. From the standard error, I have the t-score by dividing the difference between 2 means by the standard error. The degree of freedom is obtained by subtracting 1 from the smaller sample size, which is

[5] #distribution: There is a description of each histogram with the type of distribution.

the sample of eating less veggie in this case. Lastly, I use the scikit package to compute the p-value.

For effect size, I computed the pooled standard deviation and calculate the effect size by the difference between two means divided by the pooled standard deviation[6].

Detailed calculation of this significance test can be found in appendix E.

iv. Results and Conclusions

a) Confidence interval:

With the confidence level of 95%, the GPA range is [3.35, 3.488]. This means that we are 95% confidence that the true mean GPA of college students population is within this range. However, we must take note that there is still a 5% chance that this interval does not capture the true mean of the population because we have only constructed the confidence interval with 1 sample.

b) Hypothesis test:

With the standard error of 0.153 and t-score of 1.007, we have the two-tailed p-value of 0.334 > 0.05. A high p-value (out of 1) means that there is a high chance of similarity between the two means that we are comparing. On the basis of this data, we are failing to reject our null hypothesis with this high p-value. The effect size of 0.401 according to Cohen's d also suggests the same thing.

We must take note that this conclusion is a generalization from a hypothesis test. Therefore, the conclusion is inductive[7]. No matter how much data we gather, statistical inference is always deductive, as there is always the possibility of a Type I or Type II error, even if that possibility is small.

On another hand, this research has value in providing an overview of the relationship in GPA between students eat more and less veggies. More sampling can be done to give more validation to this research question.

[6] #significance: I perform a hypothesis test between 2 means with the t-distribution and effect size according to Cohen's d.

[7] #induction: The conclusion of the paper is a generalization based on a hypothesis test. To make an deductive argument, we need to know the data of the whole population, which we don't have, to make conclusion of a specific case.

v. References

Kaggle. (n.d.). Food choices and preferences of college students. Retrieved from https://www.kaggle.com/borapajo/food-choices

vi. Appendix

Appendix A: Import data and packages to Python. Create subset

In [76]:
```
#import needed packages
import pandas as pd
import numpy as np
import statistics as stats
import matplotlib.pyplot as plt
from scipy import stats

foodchoice_df = pd.read_csv("food_coded - Sheet1.csv") #import the data from c
sv file
foodchoice_df = foodchoice_df.dropna(subset = ['GPA','veggies_day'])

GPA = foodchoice_df['GPA']
veggies_day = foodchoice_df['veggies_day']

#Convert the data into list for easier manipulation
list_GPA = list(GPA)
list_veggies_day = list(veggies_day)

#Create list of 2 subset of GPA: Eat less veggies or eat more veggies
less_veggies = []
more_veggies = []

for i in range(len(list_veggies_day)):
    if list_veggies_day[i] <3: #GPA conressponding to those rate 1 or 2 on the
 likelihood of eating veggies
        less_veggies.append(GPA[i])
    if list_veggies_day[i] >3: #GPA conressponding to those rate 4 or 5 on the
 likelihood of eating veggies
        more_veggies.append(GPA[i])

foodchoice_df.head(10) #shows the first 10 rows of the data
```

Out[76]:

	GPA	veggies_day
0	2.25	1
1	3.80	1
2	2.20	2
3	3.87	2
4	3.20	2
5	3.70	2
6	3.10	2
7	3.40	2
8	3.50	2
9	3.92	2

Appendix B: Descriptive statistics for the dataset

```
#Descriptive statistics for the dataset
#Function to calculate mean, mode, range, and standard deviation of the datase
t
def mean(lst):
    total = 0 # variable to calculate the sum of all values in the list
    for i in range(len(lst)): #repeat this loop for each value of the list
        total += lst[i] # sum is the current sum plus the value of the list
    return round(total/len(lst),3) #mean is the total sum divided by the numbe
r of values of the list

def my_count(lst, val): # function to count number of occurences of 1 value in
 a list
    count = 0 #set initial occurence to 0
    for i in range(len(lst)): # the loop to repeat for each value of the list
        if lst[i] == val: # to check whether the current value matches our des
irable value "val"
            count += 1 #if it's true, the it is counted as 1
    return count

def mode(lst):
    occurence = 0 # number of occurences of the mode value
    mode = 0 # mode of the input list
    for i in range(len(lst)):
        if my_count(lst, lst[i]) > occurence: #check whether the current value
 has higher number of occurences than the previous value
            mode = lst[i] #if yes, then the mode is the current value
            occurence = my_count(lst, lst[i]) #update the maximum number of oc
curences according to the current mode
    if occurence == 1: #if occurence is 1, meaning that all number in the list
 only occurs exactly 1 and do not repeat
        return print("This list does not have a mode") #then this list does no
t have a mode
    else:
        return round(mode, 3)

def find_range(lst):
    lst_range = max(lst) - min(lst) #range of a list is the difference between
 the maximum value and minimum value
    return round(lst_range,3)

def std(lst):
    total = 0
    for i in range(len(lst)):
        total += (lst[i]-mean(lst))**2 #find the difference between the curren
t value and the mean of the list, then multiply it to the exponent of 2
    std = (total/len(lst))**0.5 #divide the total above with the number of ite
ms in the list and find square root of it
    return round(std, 3)

#Build a function to calculate descriptive statistics based on Python inbuilt
 function
def descriptive_stats(data):
    print("Count: ", len(data))
    print("Mean: ", mean(data))
    print("Median: ", np.median(data))
    print("Mode: ", mode(data))
```

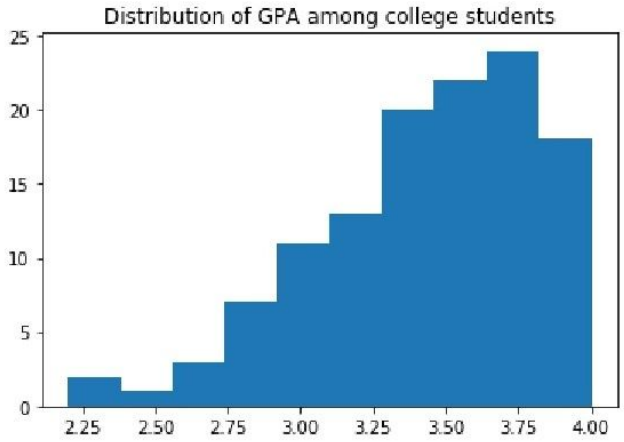
```
    print("Standard deviation: ", std(data))
    print("Range: ", find_range(data))

print("Descriptive stats for GPA")
descriptive_stats(list_GPA)

print("\nDescriptive stats for likelihood of eating veggies")
descriptive_stats(list_veggies_day)

print("\nDescriptive stats for GPA of students eating less veggies")
descriptive_stats(less_veggies)

print("\nDescriptive stats for GPA of students eating more veggies")
descriptive_stats(more_veggies)
```

```
Descriptive stats for GPA
Count:  121
Mean:  3.419
Median:  3.5
Mode:  3.5
Standard deviation:  0.388
Range:  1.8

Descriptive stats for likelihood of eating veggies
Count:  121
Mean:  4.025
Median:  4.0
Mode:  5
Standard deviation:  1.048
Range:  4

Descriptive stats for GPA of students eating less veggies
Count:  13
Mean:  3.295
Median:  3.4
This list does not have a mode
Mode:  None
Standard deviation:  0.534
Range:  1.72

Descriptive stats for GPA of students eating more veggies
Count:  88
Mean:  3.449
Median:  3.5
Mode:  3.5
Standard deviation:  0.358
Range:  1.6
```

Appendix C: Data visualization of the dataset

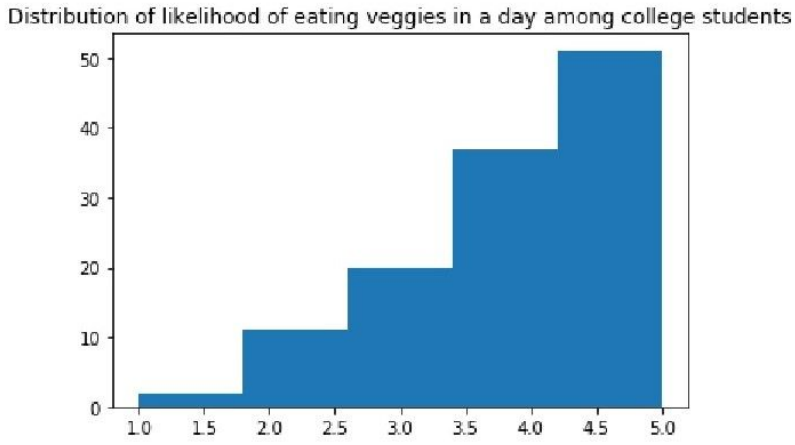
```
In [78]: #Data visualization of the dataset
         plt.hist(list_GPA)
         plt.title("Distribution of GPA among college students")
```

```
Out[78]: Text(0.5,1,'Distribution of GPA among college students')
```



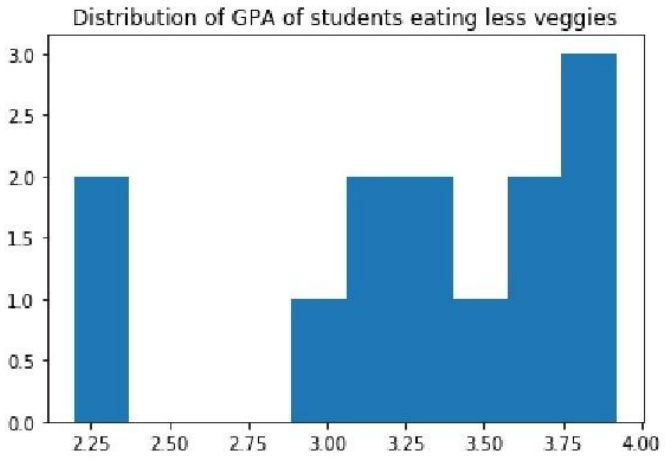

```
In [79]: plt.hist(list_veggies_day, 5)
         plt.title("Distribution of likelihood of eating veggies in a day among college
          students")
```

```
Out[79]: Text(0.5,1,'Distribution of likelihood of eating veggies in a day among colle
         ge students')
```



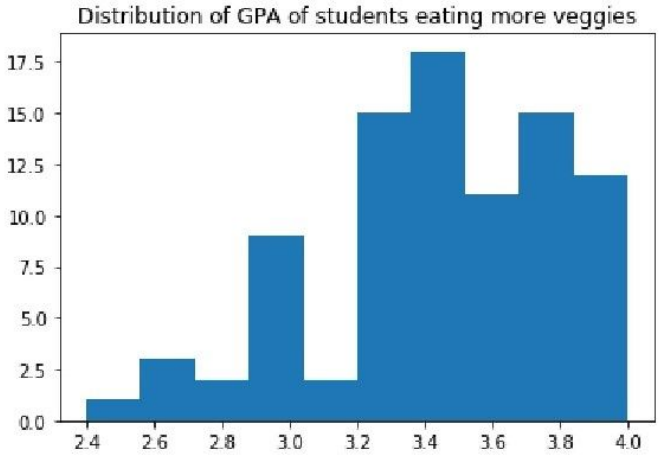
```
In [80]: plt.hist(less_veggies)
         plt.title("Distribution of GPA of students eating less veggies")
```

```
Out[80]: Text(0.5,1,'Distribution of GPA of students eating less veggies')
```




```
In [81]: plt.hist(more_veggies)
         plt.title("Distribution of GPA of students eating more veggies")
```

```
Out[81]: Text(0.5,1,'Distribution of GPA of students eating more veggies')
```



Appendix D: Confidence interval

In [82]:
```
#Constructing confidence interals of the mean of GPA
#Standard error is the standard deviation divided by the square root of the sample size
standard_error = std(list_GPA)/(len(list_GPA)**0.5)

#Upper bound of the confidence interval
#1.96 is the conressponding z-score of the 95% confidence level
upper_limit = mean(list_GPA) + 1.96*standard_error

#Lower bound of the confidence interval
lower_limit = mean(list_GPA) - 1.96*standard_error

print("The 95% confidence interval is from ", round(lower_limit, 3), " to ", round(upper_limit, 3))
```

```
The 95% confidence interval is from  3.35  to  3.488
```

Appendix E: Hypothesis test

In [83]:
```
#Constructing the hypothesis test

#Data from the less_veggies group
n1 = len(less_veggies)
std1 = std(less_veggies)
mean1 = mean(less_veggies)

#Data from the more_veggies group
n2 = len(more_veggies)
std2 = std(more_veggies)
mean2 = mean(more_veggies)

#Standard error
se = ((std1**2/n1)+(std2**2/n2))**0.5
t_score = abs(mean1-mean2)/se

#Degree of freedom is subtracting 1 from the smaller sample size
degree_of_freedom = (min(len(less_veggies), len(more_veggies))) - 1
p_value = 2*stats.t.cdf(-t_score, degree_of_freedom)
SD_pooled = (((std1**2)*(n1-1)+(std2**2)*(n2-1))/(n1+n2-2))**0.5
effect_size = (mean2 - mean1)/SD_pooled

print("Standard error: ", round(se,3))
print("T-score: ", round(t_score, 3))
print("p-value: ", round(p_value, 3))
print("Effect size: ", round(effect_size,3))
```

```
Standard error:  0.153
T-score:  1.007
p-value:  0.334
Effect size:  0.401
```