GroupReport_CocoaChocolate

April 9, 2024

Hello and welcome to this project. We are going to start. The goal will roam around chocolate! We have chocolate data from 2006 to 2024. We have 616 different manufacturers from 68 different countries. Bean origins(for cocoa) are from 64 different countries. There are 1715 specific bean origins/bar names. There are cocoa percentages of course. And, there is even the ingredients column! We can see how many ingredients we have. For example, we can see whether we have B = Beans, S = Sugar, S* = Sweetener other than white cane or beet sugar, C = Cocoa Butter, V = Vanilla, L = Lecithin, Sa = Salt. There is a column for characteristics. Also, we have ratings for the different chocolates we have. As an ID, we have REF column. The following is the link to the data. https://flavorsofcacao.com/chocolate_database.html

We are trying to answer a specific question by following and researching about this data. Basically, we know that we consume chocolate, some of us, everyday. But which chocolate material is really our favorite. In a better word, what makes a chocolate great in people's minds. Is it the chcoclate itself? or the reality inside; Cocoa! We eat chocolate but are we liking it based on the cocoa. In fact, is it the cocoa that we like or the combination of all ingredients that make the chocolate? If you ask a random person in street, probably they look at you and say: "Yo, what's the difference? I like chocolate and I like Cocoa". But no. You don't necessarily like cocoa. In fact, we don't know what you like. we don't know what cocoa percentage actually satisfies you. So, that's what we are going to answer. The reality of chocolate is cocoa. But what cocoa percentage is the favorite and frequent?

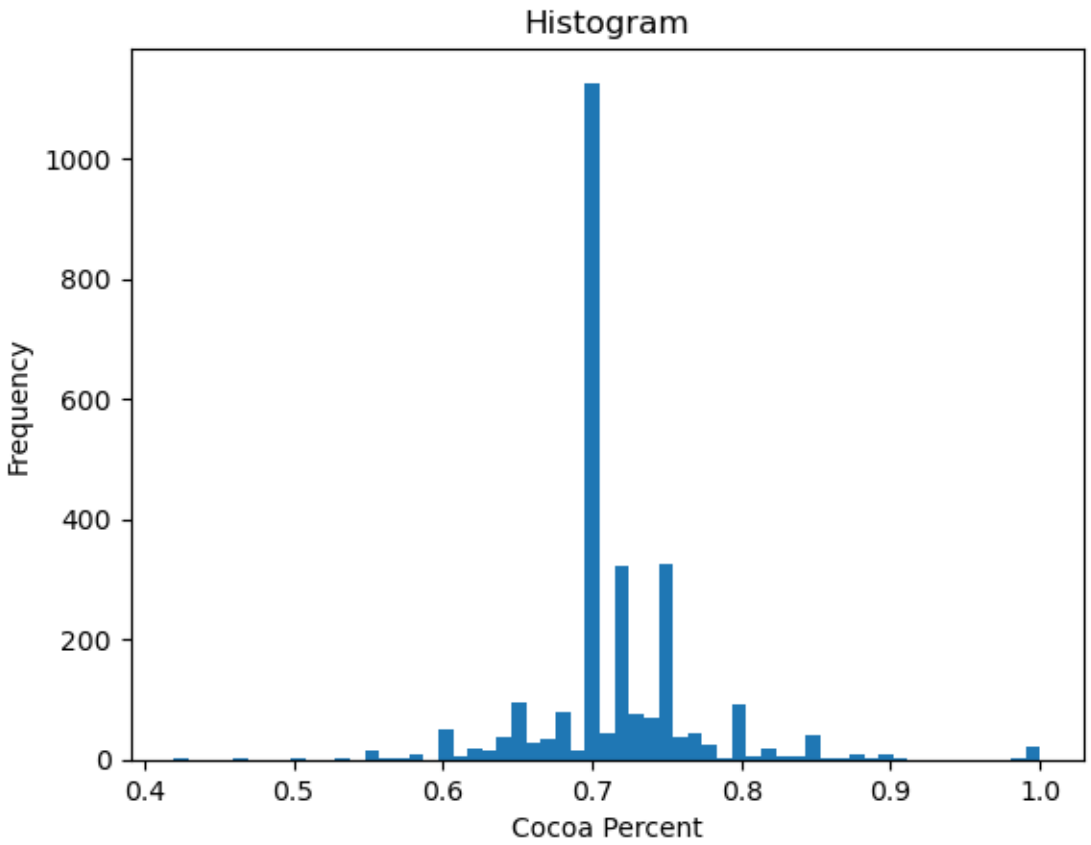
The problem we are going to solve is to let the investors know what they are actually selling or in better words, what they should be selling. Any investor, by a little bit of research will find out that he/she needs to go find cocoa beans in some part of Ivory coast or some other countries. But, who says that you need cocoa to have a favorable chocolate to sell? (JUST ADD SUGAR?). So, we want to solve the problem of how cocoa percentage matters. Our goal is approximately the same. We want to know how important the cocoa percentage is. We will insert rating column to the process. The final story is looking at cocoa column individually, getting as much as possible from it and then compare it to ratings column and decide what cocoa percentage is favorable based on history and also, is there really a positive association between cocoa and rating? Be with us, because we are going to find out.

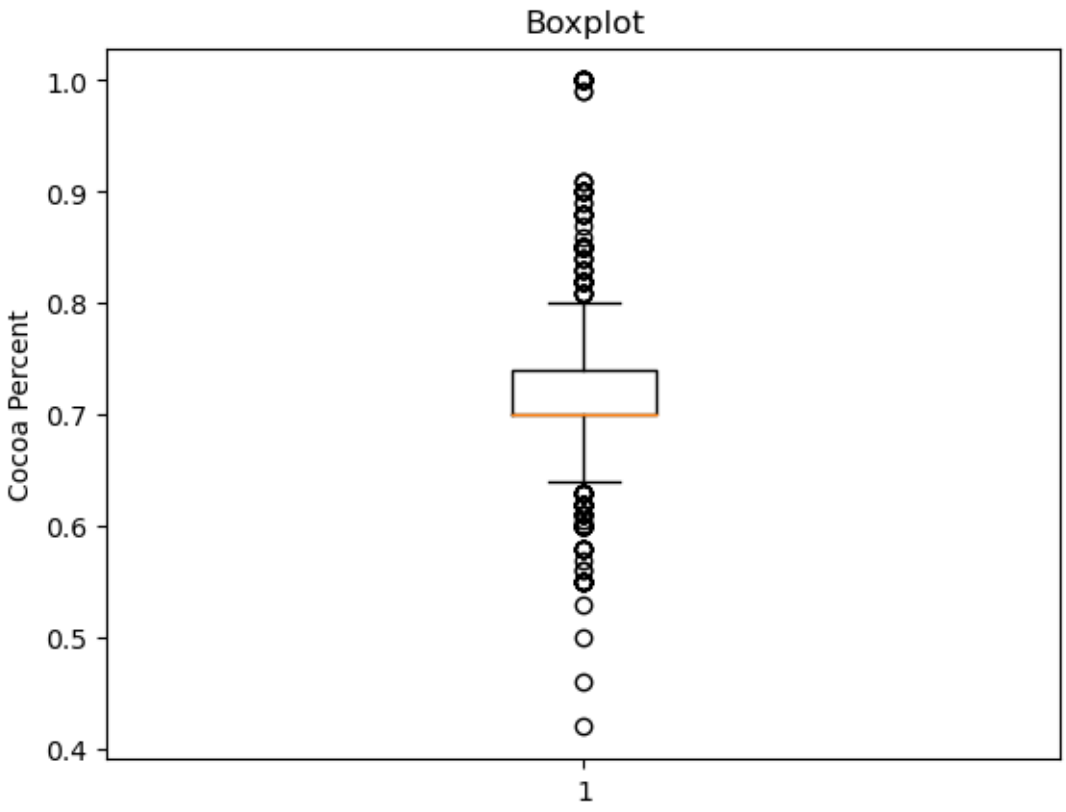
CocoaChocolate is chosen as the title. The reason why is that finally, we see ourselves finding out the true value of the reality of chocolate which is cocoa. We often think about "strawberry chocolate" or "raspberry chocolate" or "lava chocolate" that goes around your mouth while you are trying figure out where to put it and how to eat it… ayeeee. I know you have your upper teeth on your lower lips right now…The point is that whether it is raspberry or vanilla, it is cocoa at first! We need to put some respect back on the core. The core is cocoa. That is, before you eat any flavour of chocolate, know that it is a cocoa chocolate. And that's the title of this research.

Our intended audience are all of the chocolate investors and manufacturers themselves. They might want to change ingredients after listening to us! The value we are trying to offer to the industry is not just the financial aspect of it and how much cocoa is better to use among ingredients. It's also the abstract feeling of respect for cocoa itself. Whether it turns out that it's better to increase/decrease cocoa percentage in general, whether people like cocoa or not, we could train minds into eating the core rather than liking the sugar.

The belief is that all people whether they are in chocolate industry or not can benefit from our work. The reason is that depending on the conclusion we are going to make, industry can decide how to manufacture. Average random person can decide whether to eat chocolate or not and what kind of chocolate to eat. And it matters because if you are an investor and you want to make money, you need to know how to make money. If you are a manufacturer and want to make chocolate, you need to know how to make chocolate and if you are a random person in street (no offense to random people in streets), you need to know what to explore and what to eat. Health definitely matters.

[3]:
```
#histogram&boxplot:

import pandas as pd
import matplotlib.pyplot as plt
import openpyxl
import numpy as np

file_path = r'C:\Users\shbar\Downloads\stats.xlsx'

try:
    df = pd.read_excel(file_path, usecols=['Cocoa Percent'])
except FileNotFoundError:
    print(f"File '{file_path}' not found. Check the file path.")

data = df['Cocoa Percent']

# Create histogram
plt.hist(data, bins=59)
plt.xlabel('Cocoa Percent')
plt.ylabel('Frequency')
plt.title('Histogram')
plt.show()

# Create a boxplot
plt.boxplot(data)
plt.ylabel('Cocoa Percent')
plt.title('Boxplot')
plt.show()
```


Histogram
Frequency
1000
800
600
400
200
0
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Cocoa Percent



This python-generated histogram is giving us that another look; Frequency vs. Cocoa percent. There is a big difference between chocolates of cocoa percent around 70% and other chocolates. These chocolates have been rated. The first thought is that if around-70%-cocoa chocolate is the most frequent type of chocolate, it is the most favorable type of chocolate too, OR is it? Our number of records is around 2700 chocolates. It seems, around 1600 of them are 70% cocoa chocolates! Second most prominent ones would be the low and mid 70s% chocolates. Then, we have mid and high 60s% and around-80% chocolates competing with each other for the third rank. So, manufacturers are doing something. Whether they know it or not, collectively, they are producing something more than something else. Let's see if they are right.

Our boxplot alongside its outliers is giving us some idea what chocolate not to produce, of course, regardless of the idea that some people might like exceptions. That's a different story to be unique in an industry. Looking at the histogram could help us have some idea about how close Q2 and Q1 would be; Mode has a very high frequency relatively. Now, in boxplot, we can't even see Q1. Of course, they are both there. They might both be the same percentage. We just know there are a lot of data there between Q1 and Q2. However, it makes more sense that there are a lot of data between Q2 and Q3. The truth is, it's just as many data as what's between Q1 and Q2. There is a lot(of data) going on in that around-70% range. We, now, also know that mid 60s% to around-80% is what manufacturers might decide to produce by just looking at the boxplot. But, should they? Another thing to interpret is that, the number of outliers does not seem a lot relatively. But still, it does not seem to be very easy to just ignore them.

[1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import openpyxl
import numpy as np

file_path = r'C:\Users\shbar\Downloads\stats.xlsx'

try:
    df = pd.read_excel(file_path, usecols=['Cocoa Percent'])
except FileNotFoundError:
    print(f"File '{file_path}' not found. Check the file path.")

data = df['Cocoa Percent']

#mean
mean_value = np.mean(data)

#median
median_value = np.median(data)

#mode
mode_value = data.mode().iloc[0]

#quartiles
q1 = np.percentile(data, 25)
q3 = np.percentile(data, 75)

#range
data_range = np.max(data) - np.min(data)

#population variance
data_variance = np.var(data)

#population standard deviation
data_dev = np.std(data)

num_data = data.count()
#1% trimmed mean
trim1 = np.sort(data)[round(0.1*num_data):round(-0.1*num_data)]  # Remove top␣
 ↪and bottom 1%
trimmed_mean1 = np.mean(trim1)

#2.5% trimmed mean
trim2 = np.sort(data)[round(0.025*num_data):round(-0.025*num_data)]  # Remove␣
 ↪top and bottom 2.5%
trimmed_mean2 = np.mean(trim2)
```

```
print(f"Mean: {mean_value:.4f}")
print(f"Median: {median_value:.2f}")
print(f"Mode: {mode_value:.2f}")
print(f"Q1 (25th percentile): {q1:.2f}")
print(f"Q3 (75th percentile): {q3:.2f}")
print(f"Range: {data_range:.2f}")
print(f"Variance: {data_variance:.4f}")
print(f"Standard Deviation: {data_dev:.4f}")
print(f"1% Trimmed Mean: {trimmed_mean1:.2f}")
print(f"2.5% Trimmed Mean: {trimmed_mean2:.2f}")
```

```
Mean: 0.7160
Median: 0.70
Mode: 0.70
Q1 (25th percentile): 0.70
Q3 (75th percentile): 0.74
Range: 0.58
Variance: 0.0030
Standard Deviation: 0.0551
1% Trimmed Mean: 0.71
2.5% Trimmed Mean: 0.71
```

It's time..., our values...checking or not checking...That's the question...Well, the trimmed means are lowering with respect to the mean. It means what we cut from top was more effective than what we cut from bottom. Something was pulling the mean up. Now, it's gone and that's why it's down to 0.71. The most frequent data (mode) is 0.70. We thought so!. Q2(median) and Q1 are same. We guessed so! Range makes sense with our histogram. Also, IQR is 0.04. Additionally, there is not much standard deviation in data.

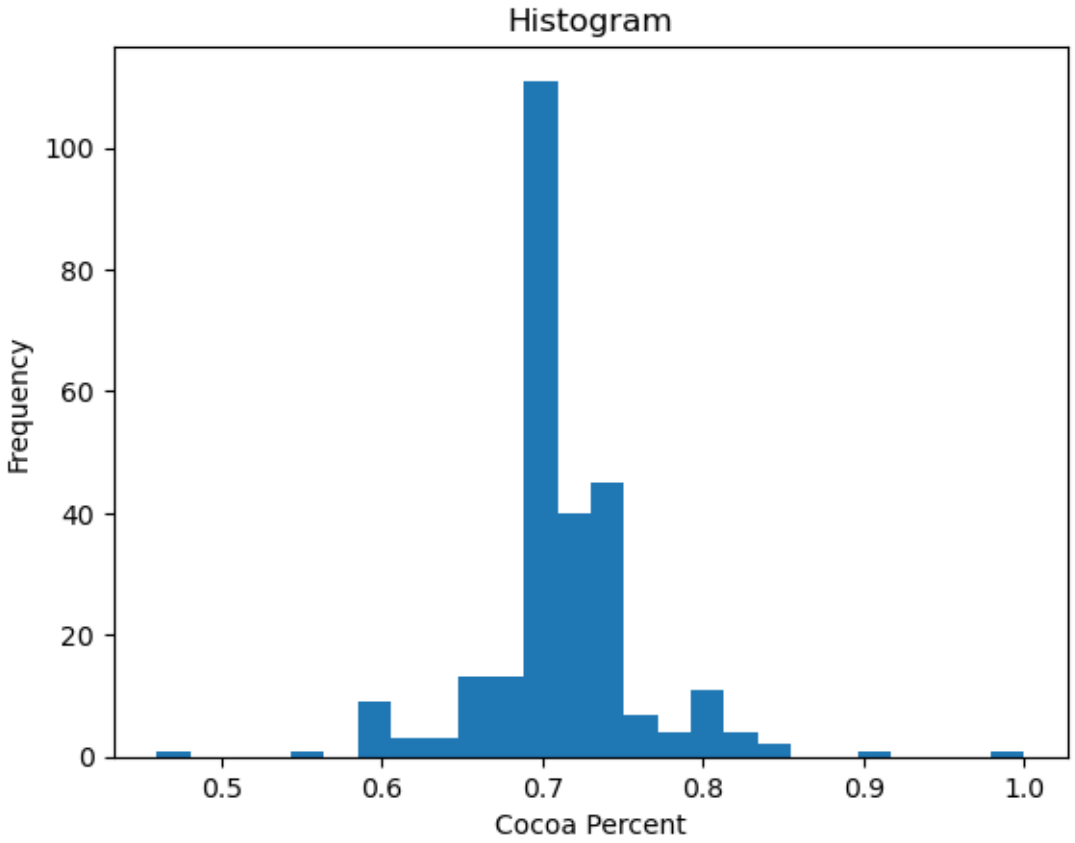
Our graphical representations align with our numerical findings. We had a lot of data around 0.7 in histogram. See that mode is 0.7. Our second most frequent data was after 0.7. See that mean rounds up to 0.72. (We will talk later why we decided to round to 4 decimal places for mean in the code). They pulled it up! Apparently, we have so many 70%s that median is 0.7 too. Q1, Q2, Q3 make sense with visual presentation of boxplot. Range makes sense with histogram. (Do you remember we had 59 bins in the code? It is basically stemming from the idea of range being 0.58)

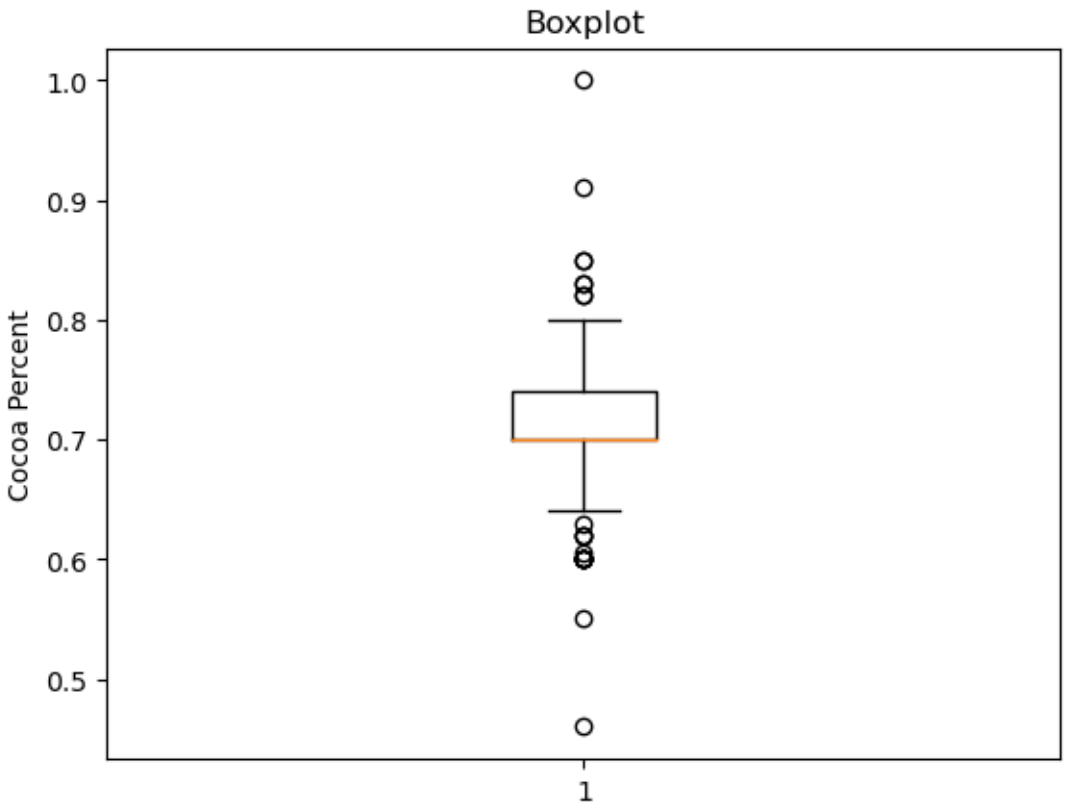
[2]:
```
import pandas as pd
import matplotlib.pyplot as plt
import openpyxl
import numpy as np
import statistics

file_path = r'C:\Users\shbar\Downloads\stats.xlsx'

try:
    df = pd.read_excel(file_path, usecols=['Cocoa Percent'])
except FileNotFoundError:
    print(f"File '{file_path}' not found. Check the file path.")
```

```
data = df['Cocoa Percent']

population_mean = np.mean(data)

sample = np.random.choice(data, size=int(0.10 * len(data)), replace=False)

plt.hist(sample, bins=26)
plt.xlabel('Cocoa Percent')
plt.ylabel('Frequency')
plt.title('Histogram')
plt.show()

plt.boxplot(sample)
plt.ylabel('Cocoa Percent')
plt.title('Boxplot')
plt.show()

#mean
sample_mean = np.mean(sample)

#median
median_value = np.median(sample)

#mode
mode_value = statistics.mode(sample)

#quartiles
q1 = np.percentile(sample, 25)
q3 = np.percentile(sample, 75)

#range
data_range = np.max(sample) - np.min(sample)

#sample variance
data_variance = np.var(sample)

#sample standard deviation
data_dev = np.std(sample)

num_data = len(sample)
#1% trimmed mean
trim1 = np.sort(sample)[round(0.1*num_data):round(-0.1*num_data)]
trimmed_mean1 = np.mean(trim1)

#2.5% trimmed mean
trim2 = np.sort(sample)[round(0.025*num_data):round(-0.025*num_data)]
```

```
trimmed_mean2 = np.mean(trim2)

print("For sample:")
print(f"Mean: {sample_mean:.4f}")
print(f"Median: {median_value:.2f}")
print(f"Mode: {mode_value:.2f}")
print(f"Q1 (25th percentile): {q1:.2f}")
print(f"Q3 (75th percentile): {q3:.2f}")
print(f"Range: {data_range:.2f}")
print(f"Variance: {data_variance:.4f}")
print(f"Standard Deviation: {data_dev:.4f}")
print(f"1% Trimmed Mean: {trimmed_mean1:.2f}")
print(f"2.5% Trimmed Mean: {trimmed_mean2:.2f}")

import scipy.stats as st

#alphas
a1=(1-0.75)
a2=(1-0.85)
a3=(1-0.95)

confidence75 =[sample_mean-(-st.norm.ppf(a1/2)*data_dev),sample_mean+(-st.norm.
 ↪ppf(a1/2)*data_dev)]
confidence85 =[sample_mean-(-st.norm.ppf(a2/2)*data_dev),sample_mean+(-st.norm.
 ↪ppf(a2/2)*data_dev)]
confidence95 =[sample_mean-(-st.norm.ppf(a3/2)*data_dev),sample_mean+(-st.norm.
 ↪ppf(a3/2)*data_dev)]

print(f"75% Confidence Interval: ({confidence75[0]:.3f}, {confidence75[1]:.
 ↪3f})")
print(f"85% Confidence Interval: ({confidence85[0]:.3f}, {confidence85[1]:.
 ↪3f})")
print(f"95% Confidence Interval: ({confidence95[0]:.3f}, {confidence95[1]:.
 ↪3f})")

import math

t_statistic=(sample_mean-population_mean)/(data_dev/(math.sqrt(num_data)))
print(f"t statistic is {t_statistic:.4f}")

from scipy.stats import t

p_value = 2 * (1 - t.cdf(abs(t_statistic), num_data-1))
print(f"pvalue is {p_value:.3f}")
if p_value <=0.05:
    print("Null hypothesis is rejected")
```

```
else:
    print("Null hypothesis is not rejected")
```






```
For sample:
Mean: 0.7122
Median: 0.70
Mode: 0.70
Q1 (25th percentile): 0.70
Q3 (75th percentile): 0.74
Range: 0.54
Variance: 0.0027
Standard Deviation: 0.0516
1% Trimmed Mean: 0.71
2.5% Trimmed Mean: 0.71
75% Confidence Interval: (0.653, 0.772)
85% Confidence Interval: (0.638, 0.787)
95% Confidence Interval: (0.611, 0.813)
t statistic is -1.1914
pvalue is 0.235
Null hypothesis is not rejected
```

In general, the histogram does not represent much difference. The most frequent chocolate is still around 70% cocoa. Obviously, there are less outliers in the boxplot. Values are not that different. It makes sense. At least in this case. Sample mean is a little smaller. Median, mode, Q1 remain the same. Q3 is also the same, so it has not moved forward or backward because of us removing data

in order to get to a 10% sample. Range decreases by 0.04. Range could differ a lot. It's random. Variance decreases by 0.0003. Standard deviation decreases by 0.0035. The reduction in variance means that the variability between data points decreases. Trimmed means are the same again!

Confidence intervals...We know the higher your percent, the higher your confidence and therefore, the bigger your interval. Here, for instance, we are 95 percent confident that our mean is between 0.611 & 0.813. We are 85 percent confident that our mean is between 0.638 & 0.787. we are 75 percent confident that our mean is between 0.653 & 0.772.

Decision is to have H0: u=0.72. Population mean is 0.72 when rounded to 2 decimal places. Sample mean is equal to 0.71 when rounded to 2 decimal places. The only reason the means are rounded to 4 decimal places is to make sense for t statistic as we have sample mean - population mean in the numerator. (There were instances of sample mean having the same mean as the population mean when rounded to 2 decimal places in the code). We basically calculate the t statistic and using the degrees of freedom (num_data-1), we get to p value. Pvalue (0.235) turns out to be more than 5% and therefore, we fail to reject the null hypothesis.

[11]:
```
import pandas as pd
import matplotlib.pyplot as plt
import openpyxl
import numpy as np

file_path = r'C:\Users\shbar\Downloads\stats.xlsx'

try:
    df = pd.read_excel(file_path, usecols=['Cocoa Percent'])
    df2 = pd.read_excel(file_path, usecols=['Rating'])
except FileNotFoundError:
    print(f"File '{file_path}' not found. Check the file path.")

data = df['Cocoa Percent']

from scipy.stats import pearsonr

data2 = df2['Rating']

correlation_coefficient, _ = pearsonr(data, data2)
coefficient_of_determination = correlation_coefficient**2

print(f'Correlation Coefficient: {correlation_coefficient:.2f}')
print(f'Coefficient of Determination: {coefficient_of_determination:.2f}')
```

```
Correlation Coefficient: -0.14
Coefficient of Determination: 0.02
```

Finally, we have an idea of what our conclusion would be. Value of r shows that there is a negative association between Cocoa Percent and Rating of the chocolate. The more cocoa you have, the less people are likely to like it! However, it is obvious that the correlation coefficient is not that much in terms of the strength of the linear relationship. Following it, Coefficient of determination is as low as 0.02. After all, although higher cocoa percentage seems to be a bad factor on ratings, it is

not leaving a huge negative impact on the chocolate rating.

Final thoughts: We started this project. It took a lot of time to find a good dataset. We went ahead with chocolates. We had to make sure the chocolates are from around the world and the dataset is reliable enough. When it became obvious that we have data from 2024 as well, it was a positive aspect. We already mentioned the number of countries and the years' interval the data is from. We had to think what goal we want to acheive. We have the column cocoa percentage. Later, we found out 70% cocoa chocolate is the most frequent type of chocolate in our dataset and our dataset is from all over the world. But we needed to find value in it. The propuse was to find out what effect cocoa percentage has on the chocolate's rating. But before that, considering the first column only, we could offer the value of information about chocoates around the world to the manufacturers and investors, also people. For example by looking at our histogram, You can find out what has realy been happenning in terms of chocolate production. If you are an investor and investing in 80% cocoa chocolate, now you know based on a comprehensive dataset that has gone after chocolates, 80% cocoa chocolate is the third rank in terms of frequency. So, you get some idea of the possible competition you get into if you are investing there.

Histogram helped with the most frequent chocolates. Boxplot showed the concentration of a lot of the data between 25th percentile and median up to the point where Q1 and Q2 are indistinguishable. If you are an investor or manufacturer, you know that working in the 70% cocoa chocolate field can bring you money but it's hard as high competition is expected. Although, median, mode and the first quartile are the same, the mean turned out to be more than them. However, trimmed means are lower. Of course we used python to calculate all these including standard deviation, variance and range. There is no lower cocoa percentage than 42%. Now, the marketing audience knows that producing chocolates with lower percentages could start something new in the industry. It is interesting to know there is 100% cocoa chocolate in this dataset. Imagine how bitter it would be. Either way, numerics and graphics aligned. Then, 10% of the population was chosen as a sample. Initially, the approach was to calculate confidence intervals and conduct hypothesis test individually. But then we realized that for the project to be consistent, we will only work on one sample and if there is any association between a CI or a HT with a sample, it is going to be the initial sample we created.

In the last step, we come back to what we were having in our mind as the second column (Rating). Considering that, correlation coefficient and coefficient of determination were caluclated. The value of r showed us almost everything we needed to know. -0.14...negative association...but weak...It is very interesting to work on a project and come up with a conclusion that you would not necessarily expect. Probably whether you are a person in chocolate industry or a chocolate consumer, you might guess that well, the more cocoa, the more bitter and therefore, the less people would like the chocolate. And that is what we came up with. But that is not all. The truth is that it does not matter that much. On paper, the less cocoa you use, the more chocolate you sell as the rating is higher historically for that type of a chocolate. But it doesn't mean that you necessarily need to go in that direction. The linear relationship is a weak negative association. If you want to start a chocolate company, you wouldn't want to put so much cocoa but you don't need to make the cocoa percentage as low as possible. Your answer as an investor or industry person is "equilibrium leaning towards less cocoa" but not "less cocoa". In other words, make chocolate. But don't be conservative about using cocoa. You can use cocoa.

One last thing: It might not be required by this project but, it might be worth mentioning. We want to introduce a different idea for chocolate industry people. Once you have a linear relationship like this that is not strong, then you know "change" is possible. That is the nature of a relationship. You

can train people's minds. Now you know that people do not hold strong ties with cocoa percentage of a chocolate. So why not making chocolate, putting enough cocoa in the chocolate, but this time, just a little more than what you have previously planned. It is suggested and thought that after a long-enough time frame of doing this, in let's say 5 years, if you assess the updated dataset of chocolates around the world again, correlation coefficient will be lower if not 0. And if you can make the tradition and make this linear relationship so weaker up to the point where you could say "no linear relationship holds anymore", you can call yourself a champion for training people's minds to like cocoa rather than other used ingredients such as sugar. That will help you, that will help me, and that will help the health of chocolate lovers. Let's go… CocoaChocolate!

Reference

Brelinski, B. (2011). Over 2700 plain dark chocolate bars rated!. Flavors of cacao - chocolate database. http://flavorsofcacao.com/chocolate_database.html