DS1000 - Assignment 3

Teagan Martins

1. a) The explanatory variable in this experiment was the usage of alcohol/marijuana.

   b) The response variable was the accuracy of their driving, which was rated from 1 to 10.

   c) The experiment was single-blind because only the participants knew what condition they were under (sober, drunk, high). The evaluator did not know.

   d) The experiment was a block design because the participants weren't chosen in pairs and assigned treatment. Instead, the participants were simply chosen at random.

2. No, this does not mean that vitamins cause additional weight loss because the vitamins may have made the participants more motivated to stay healthier overall, and therefore they may have worked out and eaten healthier as well. It's very much like a placebo effect.

3. a) This is a simple random sample and it would not result in a biased sample because each individual has an equal chance of being selected.

   b) This is a stratified sample and it would not result in a biased sample because each person in each group still has an equal chance of being selected, and each group is still being represented by the sample. This is a more complex version of a simple random sample.

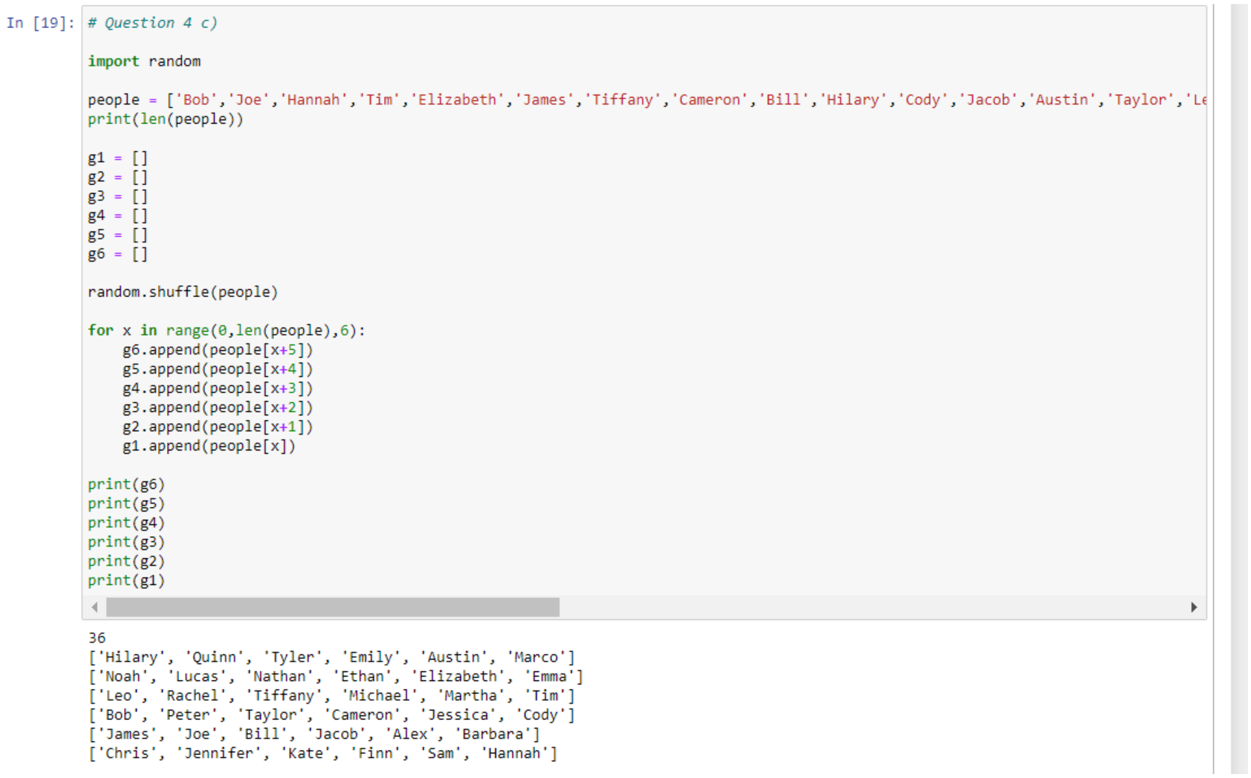
   c) This is a systematic sample and would result in a biased sample because certain cultures have more common first letters for last names, for example, Lee is a very common last name among those of East Asian descent. Because of this, it can result in a ratio of students that doesn't represent the actual ethnicity of the overall population of students.

4. a)

		Ambient Scent		
		Seashore	Firewood	No Scent
Product Density	Crowded	Group 1	Group 2	Group 3
	Empty	Group 4	Group 5	Group 6

b) I would use software that would randomly shuffles the list of all participants. From there I would assign each person into 1 group. The 1st participant in the list would be in Group 1, the 2nd participant in the list would be in Group 2… the 6th participant in the list would be in Group 6, the 7th participant in the list would be in Group 1, and so on until all 36 people have been randomly assigned to a group. Doing it this way ensures each group has an equal number of people (6). This would be systematic sampling but it would be unbiased as each person is still randomly selected.

c)



```
In [19]: # Question 4 c)

         import random

         people = ['Bob','Joe','Hannah','Tim','Elizabeth','James','Tiffany','Cameron','Bill','Hilary','Cody','Jacob','Austin','Taylor','Le
         print(len(people))

         g1 = []
         g2 = []
         g3 = []
         g4 = []
         g5 = []
         g6 = []

         random.shuffle(people)

         for x in range(0,len(people),6):
             g6.append(people[x+5])
             g5.append(people[x+4])
             g4.append(people[x+3])
             g3.append(people[x+2])
             g2.append(people[x+1])
             g1.append(people[x])

         print(g6)
         print(g5)
         print(g4)
         print(g3)
         print(g2)
         print(g1)
```

```
36
['Hilary', 'Quinn', 'Tyler', 'Emily', 'Austin', 'Marco']
['Noah', 'Lucas', 'Nathan', 'Ethan', 'Elizabeth', 'Emma']
['Leo', 'Rachel', 'Tiffany', 'Michael', 'Martha', 'Tim']
['Bob', 'Peter', 'Taylor', 'Cameron', 'Jessica', 'Cody']
['James', 'Joe', 'Bill', 'Jacob', 'Alex', 'Barbara']
['Chris', 'Jennifer', 'Kate', 'Finn', 'Sam', 'Hannah']
```

5. a) probability that a randomly selected seedling was damaged by deer = total damaged by deer/total seedlings

(60+76+44+29)/(60+76+44+29+151+158+177+176) = 209/871 = 0.239954075 which we can round to 0.24 or 24% chance.

b) probability that a randomly selected seedling was damaged by deer given level of cover = total damaged by deer given level of cover/total seedlings given cover

No Cover

60/(60+151) = 60/211 = 0.284360189 which we can round to 0.284 or 28.4%.

≤1/3 Cover

76/(76+158) = 76/234 = 0.324786324 which we can round to 0.325 or 32.5%

1/3 to 2/3 Cover

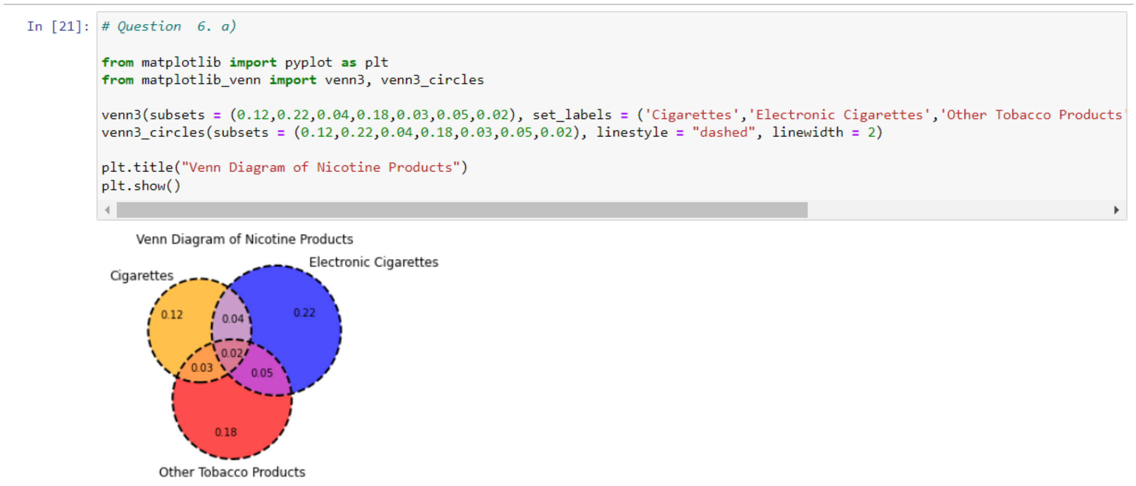
44/(44+177) = 44/221 = 0.199095022 which we can round to 0.199 or 19.9%

>2/3 Cover

29/(29+176) = 29/205 = 0.141463414 which we can round to 0.141 or 14.1%

c) Cover and damages are not independent because as the cover increases, the deer damage decreases. If they were independent, then each conditional probability would be roughly the same, however, this is not the case. There is a very clear pattern that one can see.

6. a)



```
In [21]: # Question  6. a)

         from matplotlib import pyplot as plt
         from matplotlib_venn import venn3, venn3_circles

         venn3(subsets = (0.12,0.22,0.04,0.18,0.03,0.05,0.02), set_labels = ('Cigarettes','Electronic Cigarettes','Other Tobacco Products'
         venn3_circles(subsets = (0.12,0.22,0.04,0.18,0.03,0.05,0.02), linestyle = "dashed", linewidth = 2)

         plt.title("Venn Diagram of Nicotine Products")
         plt.show()
```

b) The probability that a randomly selected high school student did not use any tobacco product can be calculated by calculating the total percentage that the venn diagram takes up. 1 – (0.22 + 0.12 - 0.04 + 0.18 – 0.05 – 0.01) = 1 – 0.42 = 0.58 or 58%.

We do 0.12 – 0 .04 to remove the already accounted for area in 0.22. Doing this leaves us with 0.12 + 0.01. The 0.01 is the overlapping area between cigarettes and other tobacco products without the middle part.

We do 0.18 – 0.05 – 0.01 to remove the already accounted for area in the other two circle.

c) The probability that a randomly selected high school student used electronic cigarettes but not any other tobacco product can be calculated by subtracting any overlapping areas from the total area of electronic cigarettes.

0.22 – 0.05 – 0.02 = 0.15 or 15%.

We do 0.22 – 0.05 to remove the overlapping area between electronic cigarettes and other tobacco products.

We do 0.22 – 0.02 to remove the overlapping area between electronic cigarettes and cigarettes but without the very middle part of the venn diagram.

d) $P(C|A) = P(A \& C)/P(A) = 0.03/0.18 = 0.166666$ or 16.66666%
The conditional probability of those that use other tobacco products also smoking cigarettes is 16.66666%.

$P(A|C) = P(C \& A)/P(C) = 0.03/0.12 = 0.25$ or 25.0%
The conditional probability of those that smoke cigarettes also using other tobacco products is 25.0%.

7. a) We know that the approximate distribution = standard deviation/ $\sqrt{n}$
Therefore, in this case, the approximate distribution = $1.1/ \sqrt{24} = 0.224536559$ which we can round to 0.22.

b) $z = (8 – 8.9)/0.224536559 = -4.008255956$ which we can round to -4.01
$P(x<8) = P(Z<(8-8.9))$
$P(x<8) = P(Z<-4.01)$
$P(x<8) = 0.00003$ which we can consider 0.

```
import scipy.stats as st
st.norm.cdf(-4.008255956)
```

```
Out: 3.058440239438628e-05
```

c) z = (100/12-8.9)/((0.224536559)(
$P(x<100/12) = P(z<(100/12-8.9)/0.224536559)$
$P(x<8.33333) = P(z<-2.523716713)$
$P(x<8.33333) = 0.0059$

8. a)

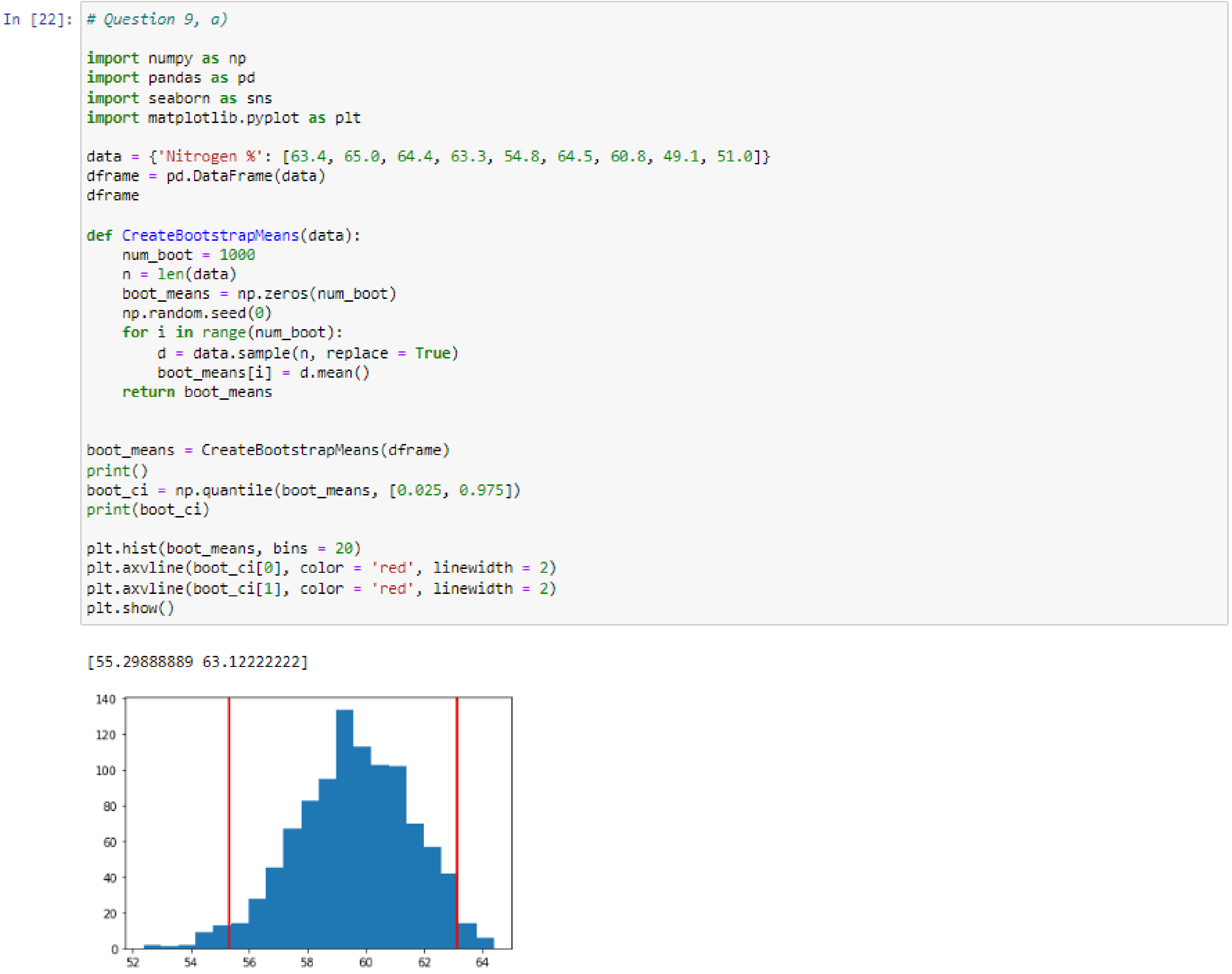
Stem	Leaves
-8	3
-7	0 8
-6	2 5 5 8 8
-5	1 2 2 3 3 6 7 9
-4	0 3 4 7 7 9 9
-3	0 1 3 6 8
-2	0 1 1 2 2 3 5 5 7
-1	0 0 8
-0	3 8
0	2 3 4
1	7
2	2

Key: -2 | 2 = -2.2

b) Lower Limit: -3.587 (mean) – 1.96 x 2.8/√47 = -4.387507073
Upper Limit: -3.587 (mean) + 1.96 x 2.8/√47 = -2.786492927
The interval is (-4.387507073, -2.786492927).

c) Yes, it would be correct to say that the probability is 95% that the mean percentage change in the population lies in the interval I computed in part b. It is safe to say this because the 95% interval does not touch zero.

9. a)



In [22]:

```
# Question 9, a)

import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

data = {'Nitrogen %': [63.4, 65.0, 64.4, 63.3, 54.8, 64.5, 60.8, 49.1, 51.0]}
dframe = pd.DataFrame(data)
dframe

def CreateBootstrapMeans(data):
    num_boot = 1000
    n = len(data)
    boot_means = np.zeros(num_boot)
    np.random.seed(0)
    for i in range(num_boot):
        d = data.sample(n, replace = True)
        boot_means[i] = d.mean()
    return boot_means


boot_means = CreateBootstrapMeans(dframe)
print()
boot_ci = np.quantile(boot_means, [0.025, 0.975])
print(boot_ci)

plt.hist(boot_means, bins = 20)
plt.axvline(boot_ci[0], color = 'red', linewidth = 2)
plt.axvline(boot_ci[1], color = 'red', linewidth = 2)
plt.show()
```

```
[55.29888889 63.12222222]
```

b) Because the confidence interval is (55.29888, 63.12222) and 75.4 is out of range of the interval, it is safe to say that the ancient air differs significantly from the present atmosphere as there is a difference greater than 12 between the upper limit and the number we are looking at (75.4).