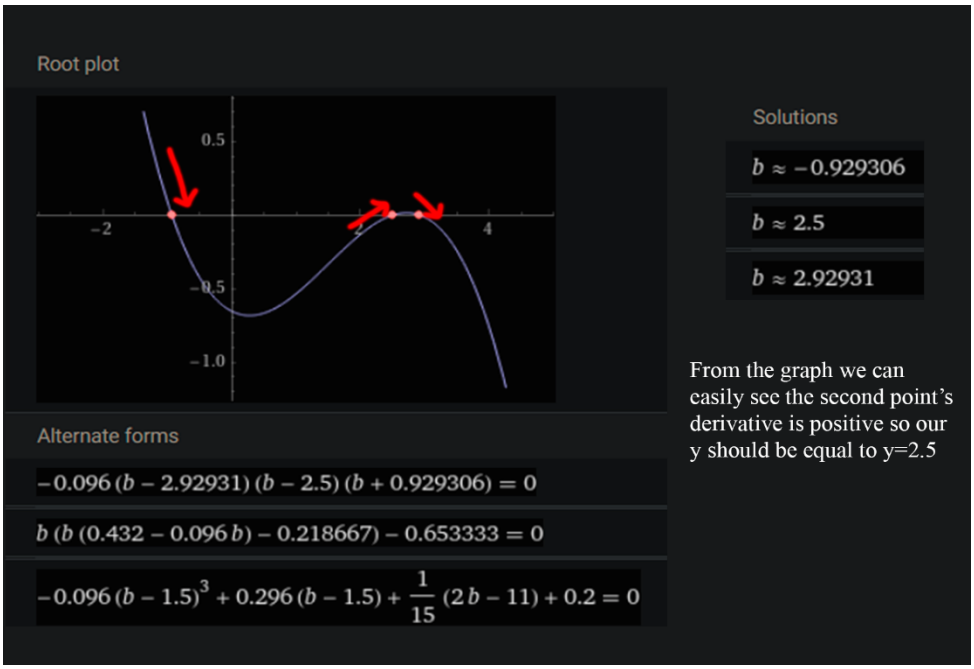
CENG 222 TERM
PROJECT REPORT

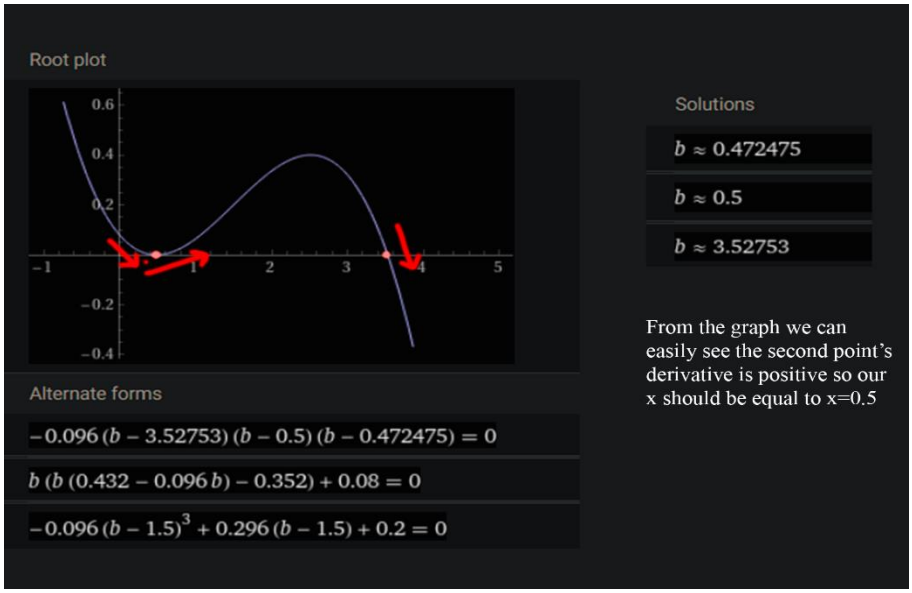
Task-1:

1-1) It looks like geometric distribution because we trying to find the first success. **A = 0.5^(a+1)** with the PMF of **X = p*(1-k)^(k-1)** which means **<u>A = k-1!</u>**

1-2) For finding y you should make the two PDF equal to them and if we calculate this equation we will get **<u>y=5.5!</u>**



For finding x we should make first equation equals to zero from here we will get the equation shown in the figure and we get from here **x=0.5**

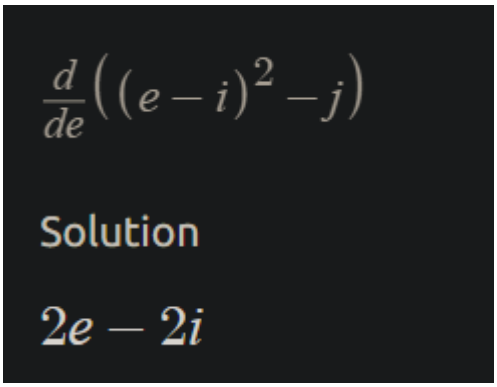


Finding z is the easiest one of these equations we can simply say **(-2b+11)/15 = 0** from here we will get z as **z=5.5**

For finding t we will simply put the y value we found from the equations above to the equation **(-2b+11)/15**. If we use the y value, then we will get **t** as **(-2*2.5+11)/15 = 0.4** so we can say t is equal to the **t = 0.4**

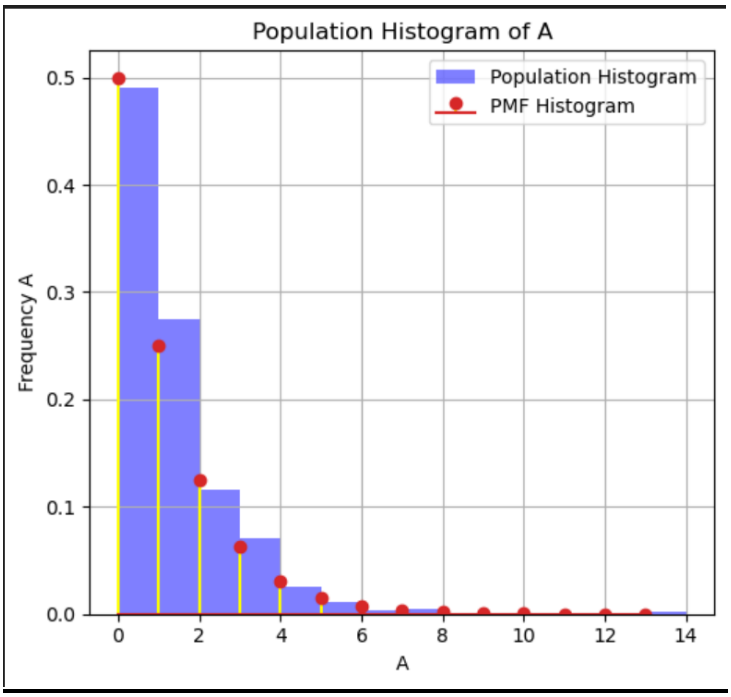
At the end our values are **x = 0.5, y=2.5, z=5.5** and the **t=0.4**

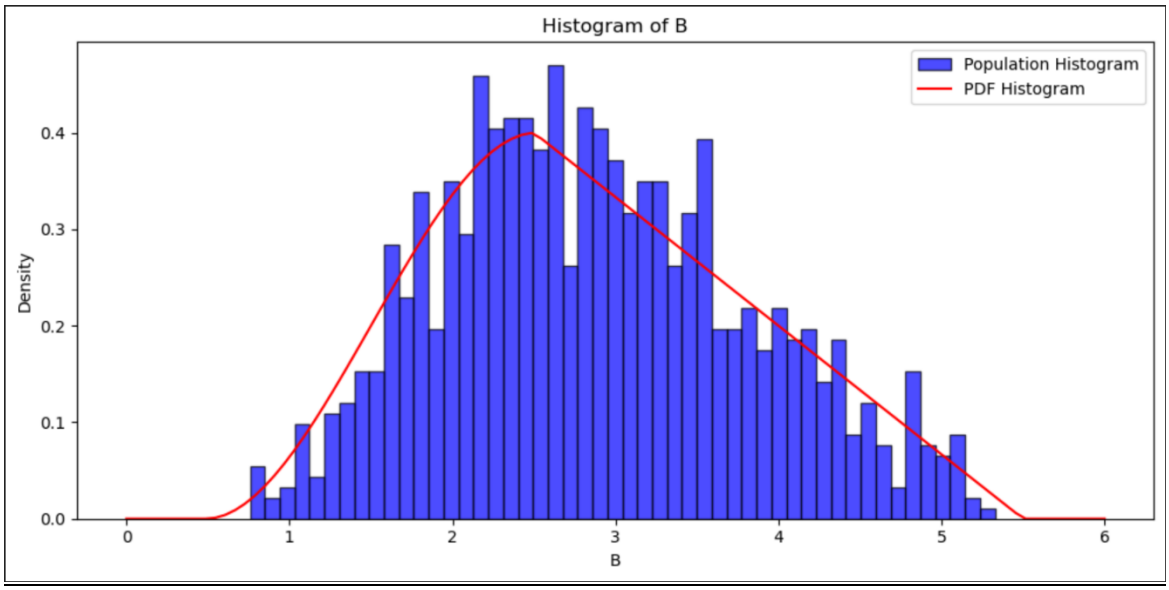
1-3) For finding the probability density function of the e we can use the given cumulative distribution function. The CDF = **(e-i)^2-j**. If we take the derivative of the CDF we will get

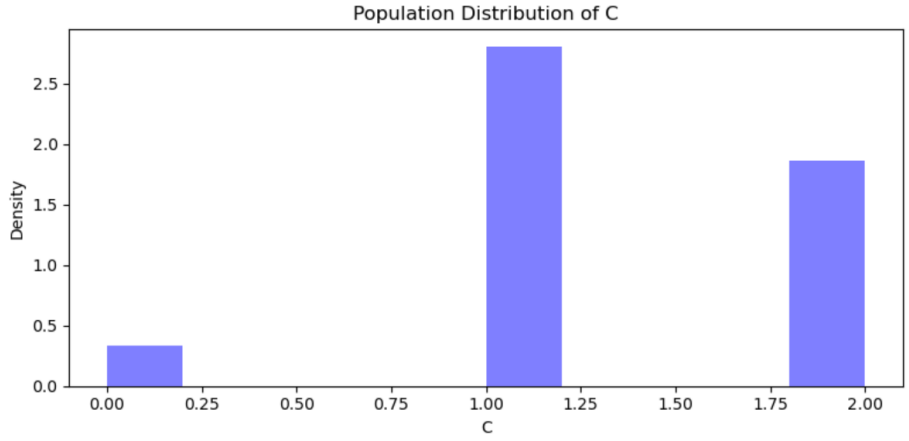


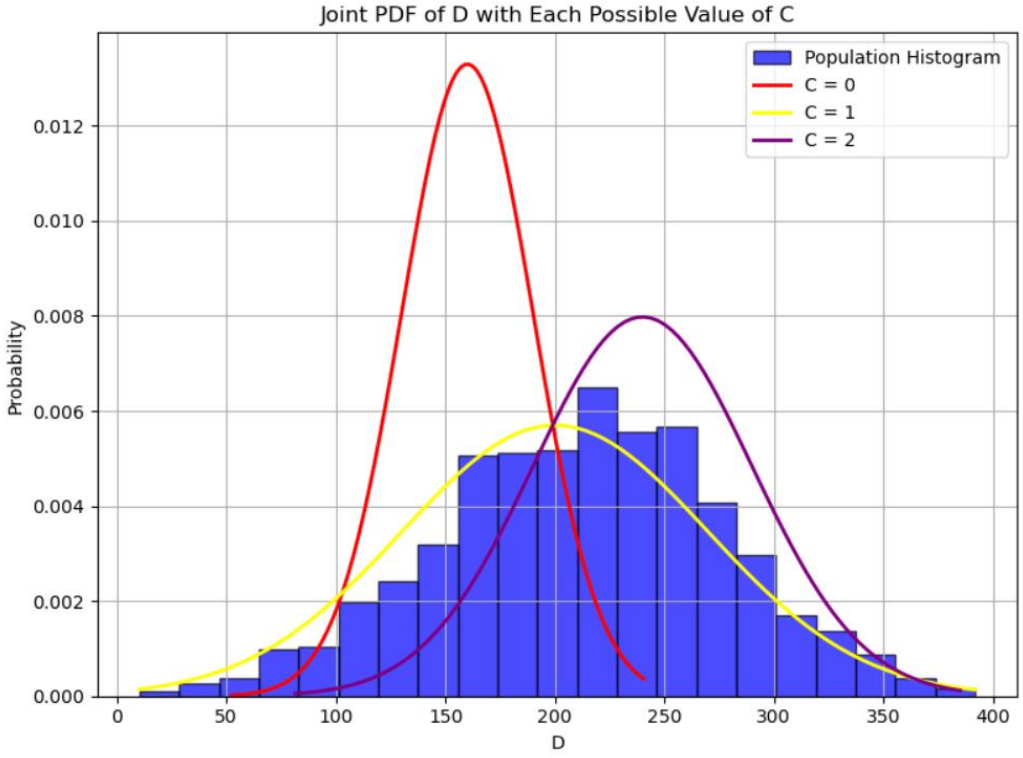
1-4) If we look at the graph and the function. The function is the probability density function and area below the lines should equal the 1 and if make calculations. **l = 1-k**. This solution holds. We can get it from equation **(1-0) * k + (2-1) * l = 1** so **k + l = 1** and **l = 1-k**

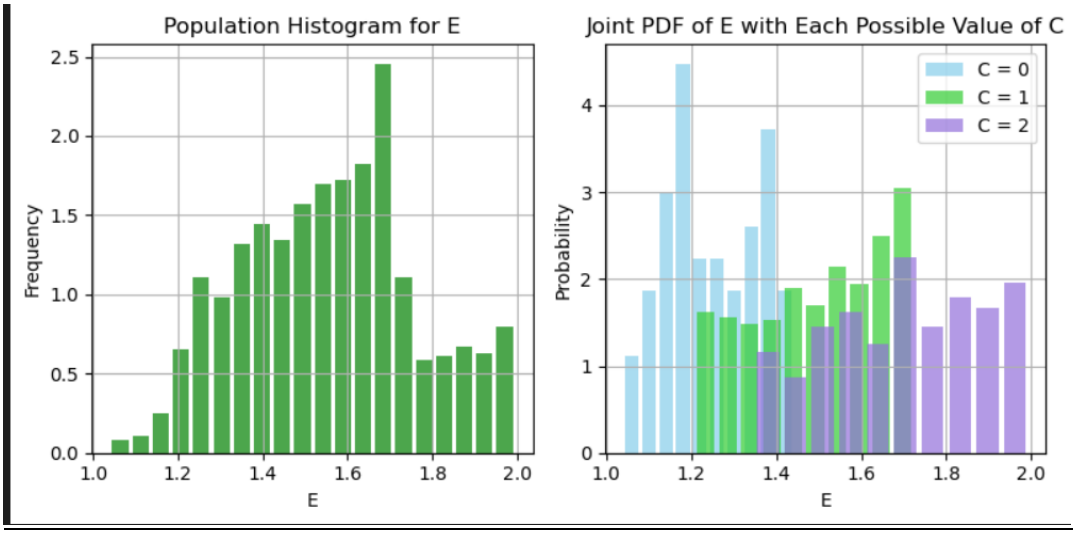
1-5) **The figures of the generated 1000 population:**



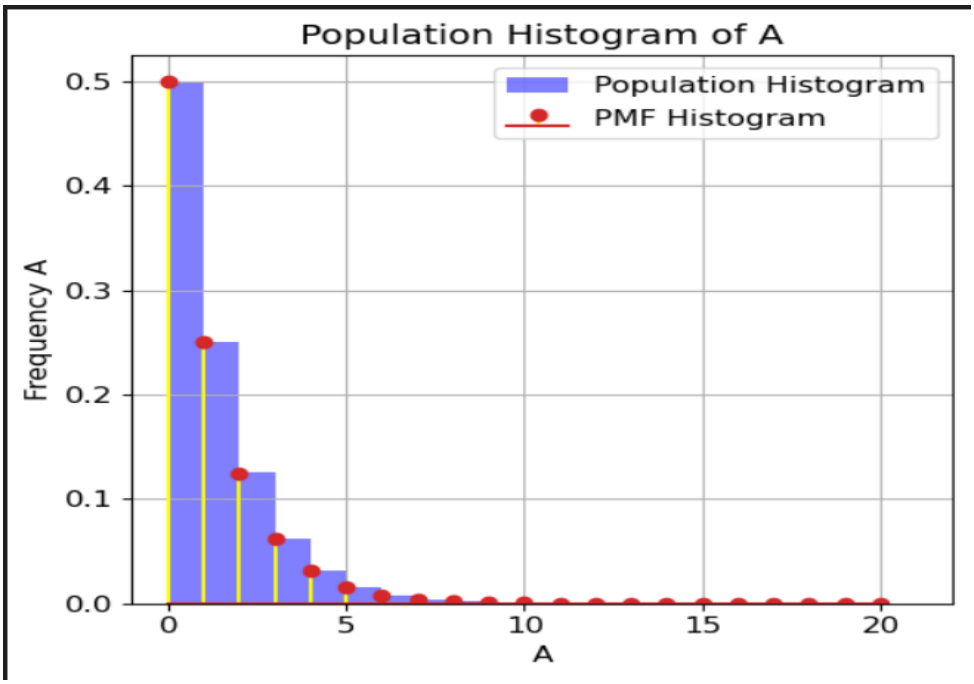


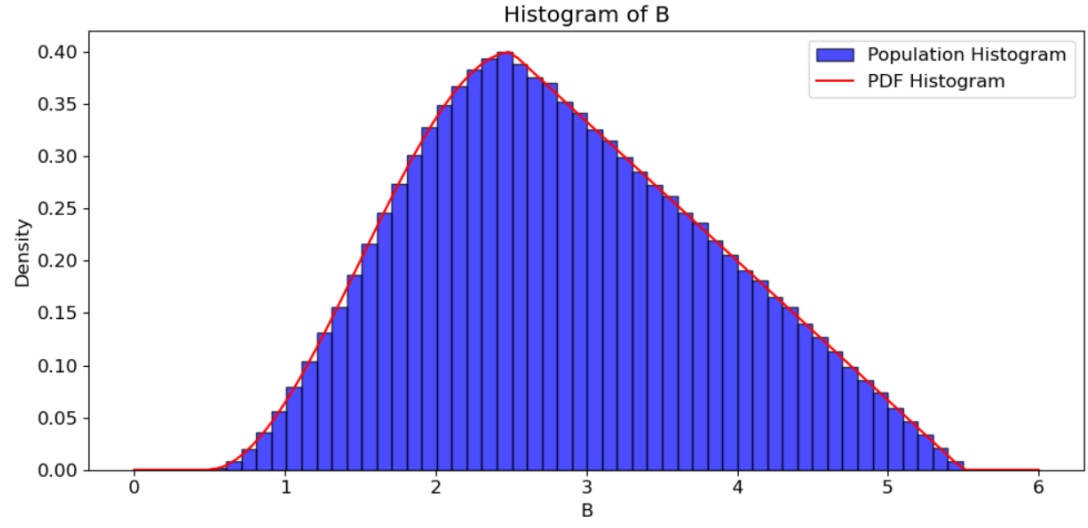

Population Distribution of C
Density
C
0.0
0.5
1.0
1.5
2.0
2.5
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00

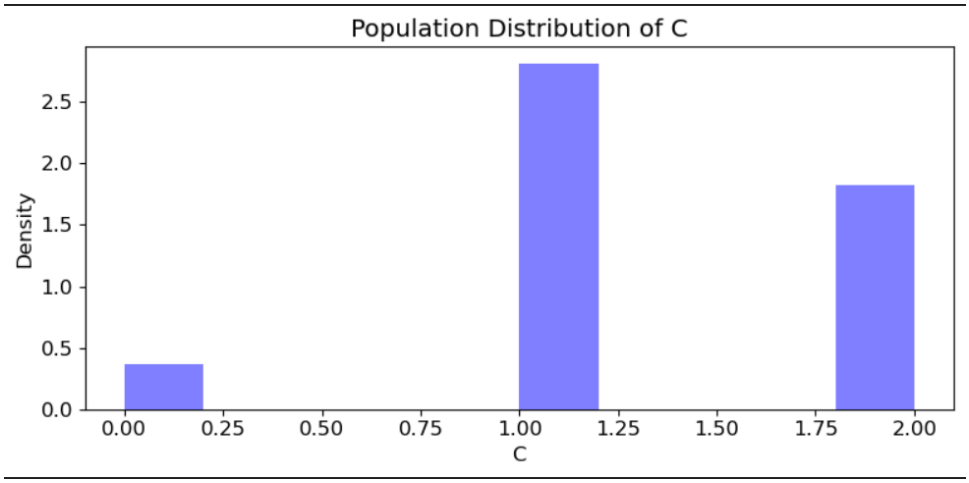

Joint PDF of D with Each Possible Value of C
Population Histogram
C = 0
C = 1
C = 2
Probability
D
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0
50
100
150
200
250
300
350
400

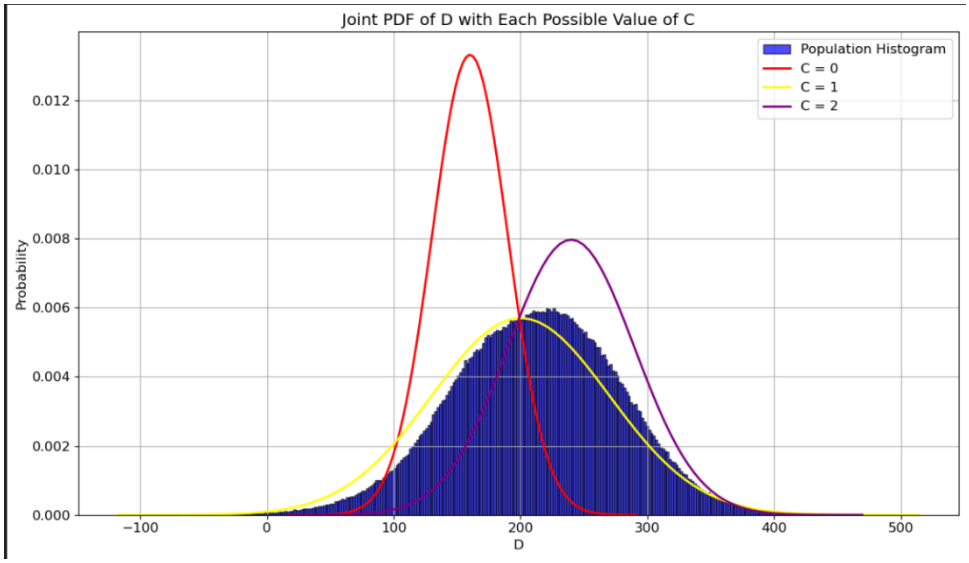

Population Histogram for E
Frequency
E
0.0
0.5
1.0
1.5
2.0
2.5
1.0
1.2
1.4
1.6
1.8
2.0
Joint PDF of E with Each Possible Value of C
C = 0
C = 1
C = 2
Probability
E
0
1
2
3
4
1.0
1.2
1.4
1.6
1.8
2.0

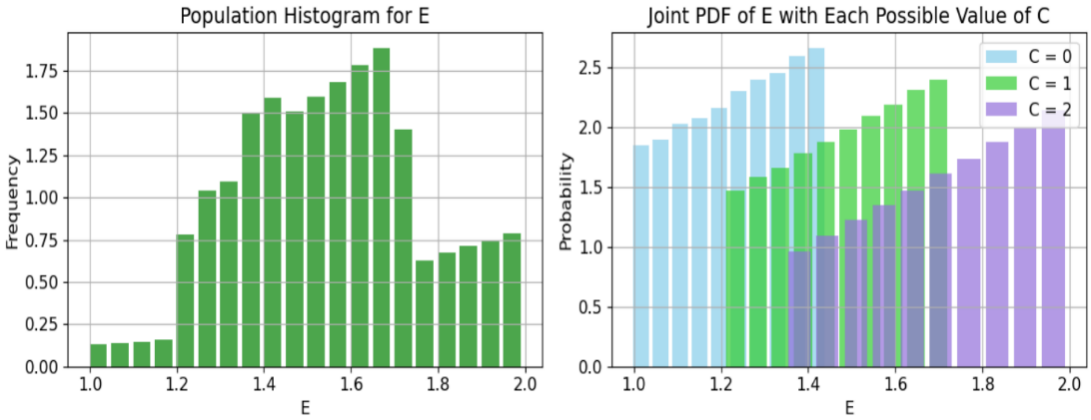
1-5) **The figures of the generated 1 million population:**






Population Distribution of C
Density
C
0.0
0.5
1.0
1.5
2.0
2.5
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00


Joint PDF of D with Each Possible Value of C
Population Histogram
C = 0
C = 1
C = 2
Probability
D
0.000
0.002
0.004
0.006
0.008
0.010
0.012
−100
0
100
200
300
400
500


Population Histogram for E
Frequency
E
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
1.0
1.2
1.4
1.6
1.8
2.0
Joint PDF of E with Each Possible Value of C
C = 0
C = 1
C = 2
Probability
E
0.0
0.5
1.0
1.5
2.0
2.5
1.0
1.2
1.4
1.6
1.8
2.0

Task-2:

2-1) For finding the theoretical mean we will have the formula

Mean (μ) = Σ(x * p(x)):

From this formula if we apply the "pmf_a" which is equals to $P(a) = 0.5^{(a+1)}$ from this equation we will get approximately the solution

Variance (σ^2) = Σ((x - μ)^2 * p(x)):

Formula given below I found the variance with using this

Variance (σ^2) = $E[x^2] - E[x]^2$

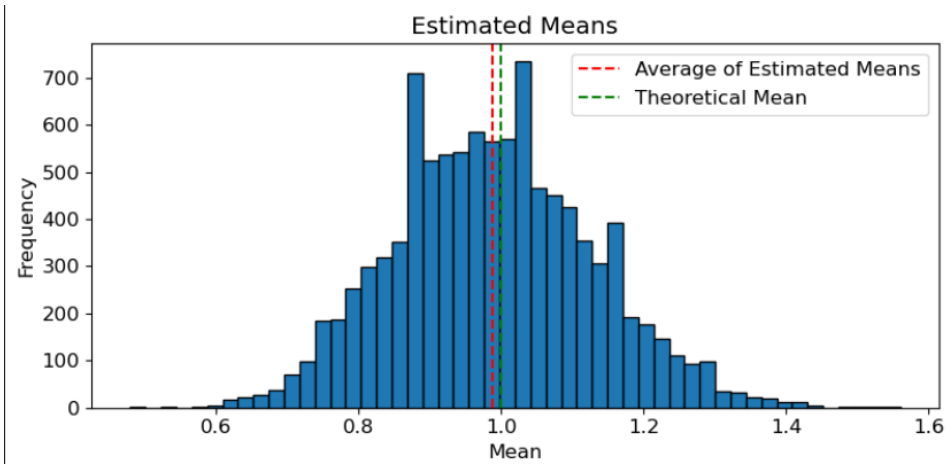
I applied the formula below and I found the values

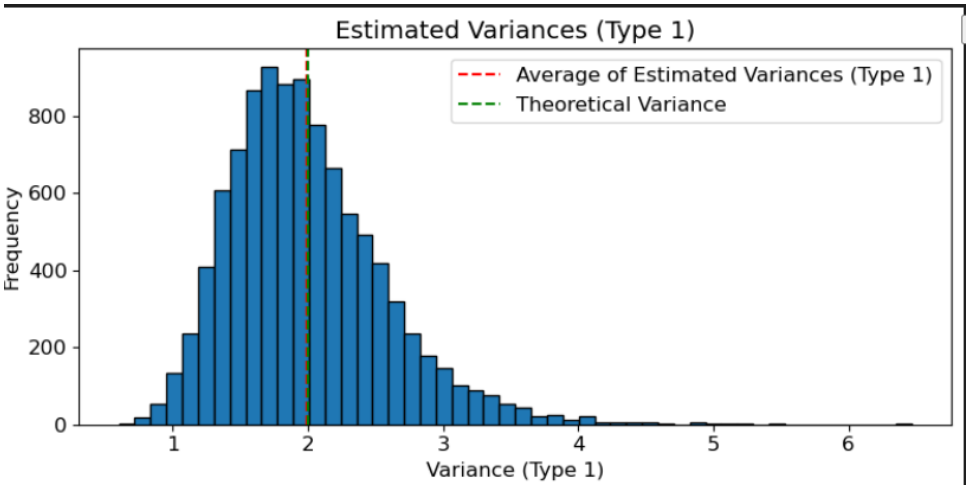
```
expected_A = 0
expected_A_square = 0

for a in range(1000000):
    expected_A += a * (0.5 ** (a + 1))

for a in range(1000000):
    expected_A_square += (a ** 2) * (0.5 ** (a + 1))

theoretical_mean = expected_A
theoretical_variance = expected_A_square - (expected_A ** 2)

print("Theoretical Mean:", theoretical_mean)
print("Theoretical Variance:", theoretical_variance)
```

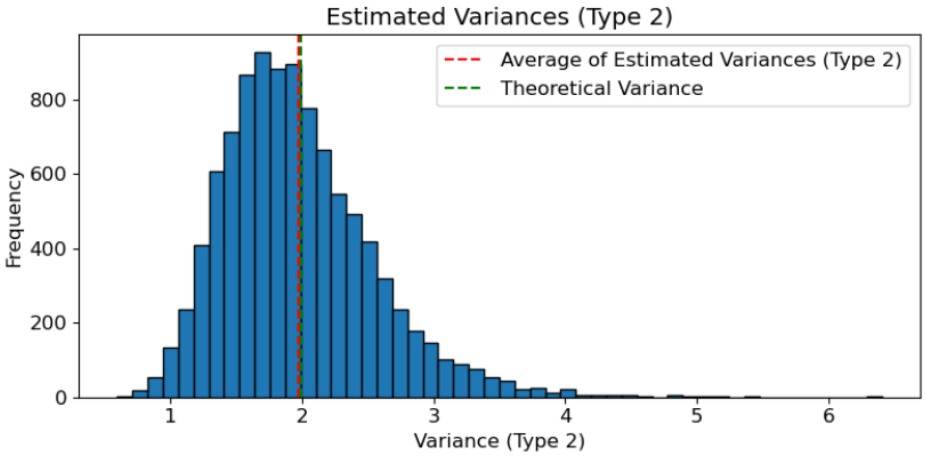
```
Theoretical Mean: 0.9999999999999999
Theoretical Variance: 1.9999999999999998
```

From here we can get the theoretical mean and theoretical variances

2-2) **The figures from the step 7**

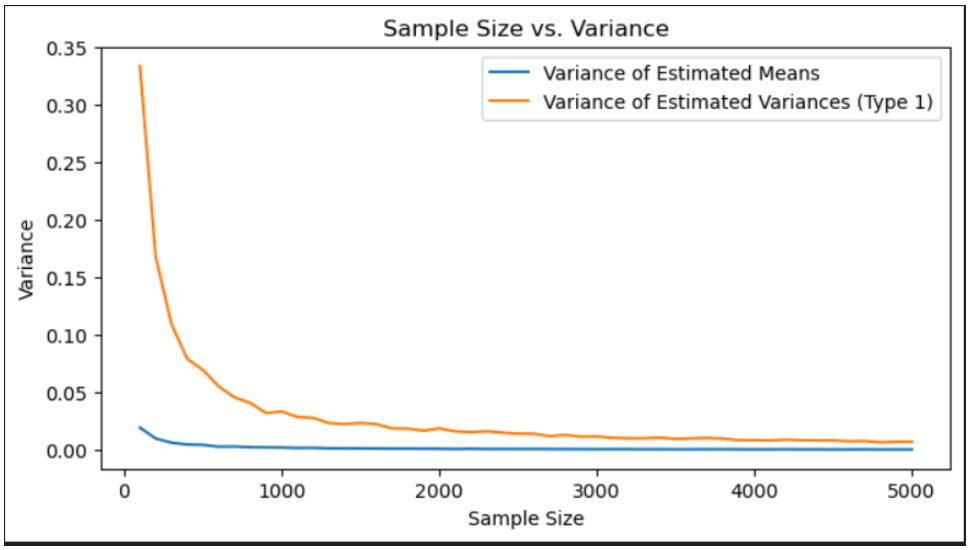






From these plots, we see that even if we do not use all the data, we can obtain results very close to the true mean and variance values by using a random part of the data. Also, we can see the variances are too close to them.

2-3) **The figures from the step 9**



If we use data with a small number of samples, it would be normal for the variance to be high because there will be outliers. If we use a dataset with many samples, the variance value will be much lower as the values will be close to what they should be.

Task-3:

3-1)

$$\sum_i^n \frac{\partial}{\partial\theta} \ln\langle f(x_i|\theta)\rangle = 0$$

$$\sum_i^n \frac{\partial}{\partial\theta} \ln\langle f(x_i|\theta)\rangle = \sum_i^n \nabla_{\mu,\sigma} \ln\langle f(x_i|\mu,\sigma)\rangle = 0$$

$$\sum_i^n \nabla_\mu \ln\langle \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}} \rangle = 0$$

$$\sum_i^n \nabla_\mu \ln\langle \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}} \rangle = \sum_i^n \nabla_\mu -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(x_i-\mu)^2}{2\sigma^2}$$

$$\sum_i^n \nabla_\mu -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(x_i-\mu)^2}{2\sigma^2} = -\frac{1}{2\sigma^2}\sum_i^n \nabla_\mu (x_i-\mu)^2$$

$$= -\frac{1}{2\sigma^2}\sum_i^n -2(x_i-\mu) = \frac{1}{\sigma^2}\sum_i^n (x_i-\mu)$$

$$\frac{1}{\sigma^2}\sum_i^n (x_i-\mu) = 0 \rightarrow \hat{\mu}_{MLE} = \frac{1}{n}\sum_i^n x_i$$

From these equations we can get the maximum likelihood estimation of the parameter d. I used the last equation in my code to implement **"estimate_mml_d**".

3-2) For estimating i and j values we will use method of moments. First need to calculate the PDF from the given CDF we will get

(*I wrote x for e and a for i from here because calculators can not recognize*)

$$2(x - \mathrm{i})$$

Then we need to find moment 1 and moment 2 for 2 unknown variable and make them equal to the integral of **PDF * e** and **PDF * e**2** in given lower and upper bound.

$$\int_{a+\sqrt{j}}^{a+\sqrt{1+j}} (2x - 2a)\, x\, dx = a + \frac{2}{3}\left(-j^{3/2} + \sqrt{j+1}\; j + \sqrt{j+1}\right)$$

From the first integral which is **PDF * e** in given lower and upper bound we get this result and know that from the method of moments it is equals to the moment 1 which is the mean

$$\int_{a+\sqrt{j}}^{a+\sqrt{1+j}} (2x - 2a)\, x^2\, dx = a^2 + \frac{4}{3}a\left(-j^{3/2} + \sqrt{j+1}\; j + \sqrt{j+1}\right) + j + \frac{1}{2}$$

From the second integral which is **PDF*e**2** in the given lower and upper bound we get this result and know that from the method of moments it is equals to the moment 2. From this two equation we can get i and it is equals to

```
i = m1 - 2/3*(-j**(3/2) + sp.sqrt(j + 1)*j + sp.sqrt(j + 1))
```

If we put the known i value to the second equation, we will get equation which have only one unknown and it will be the j. After making simplifications we able to calculate the j value with using **sp.solve()**.

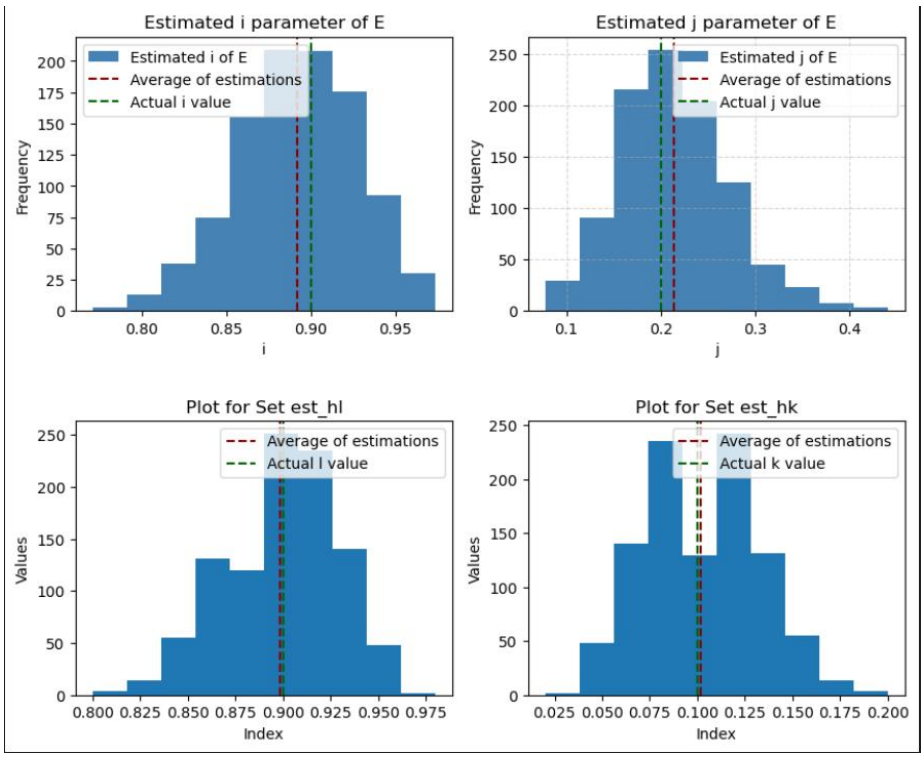
3-3) To estimate the values of k and l using the Maximum Likelihood Estimation (MLE) method, we need to formulate an equation using each h in the sample. We pass the unknown value of k to the **"pdf_h"** method, which returns either k or 1-k. We then multiply all of these probabilities, resulting in an expression like **k^a * (1-k)^(len-a)**, where "a" represents a specific count.

Estimating the value of k using this equation can be challenging. However, by taking the logarithm and differentiating it to maximize the value of k, we can solve the equation and obtain a close approximation for k.

$$\begin{aligned}\ell(\mathbf{w}) &= \log p(y_1, \mathbf{x}_1, y_2, \mathbf{x}_2, \ldots, y_n, \mathbf{x}_n|\mathbf{w}) \\ &= \log \prod_{i=1}^{n} p(y_i, \mathbf{x}_i|\mathbf{w}) \\ &= \sum_{i=1}^{n} \log p(y_i, \mathbf{x}_i|\mathbf{w})\end{aligned}$$

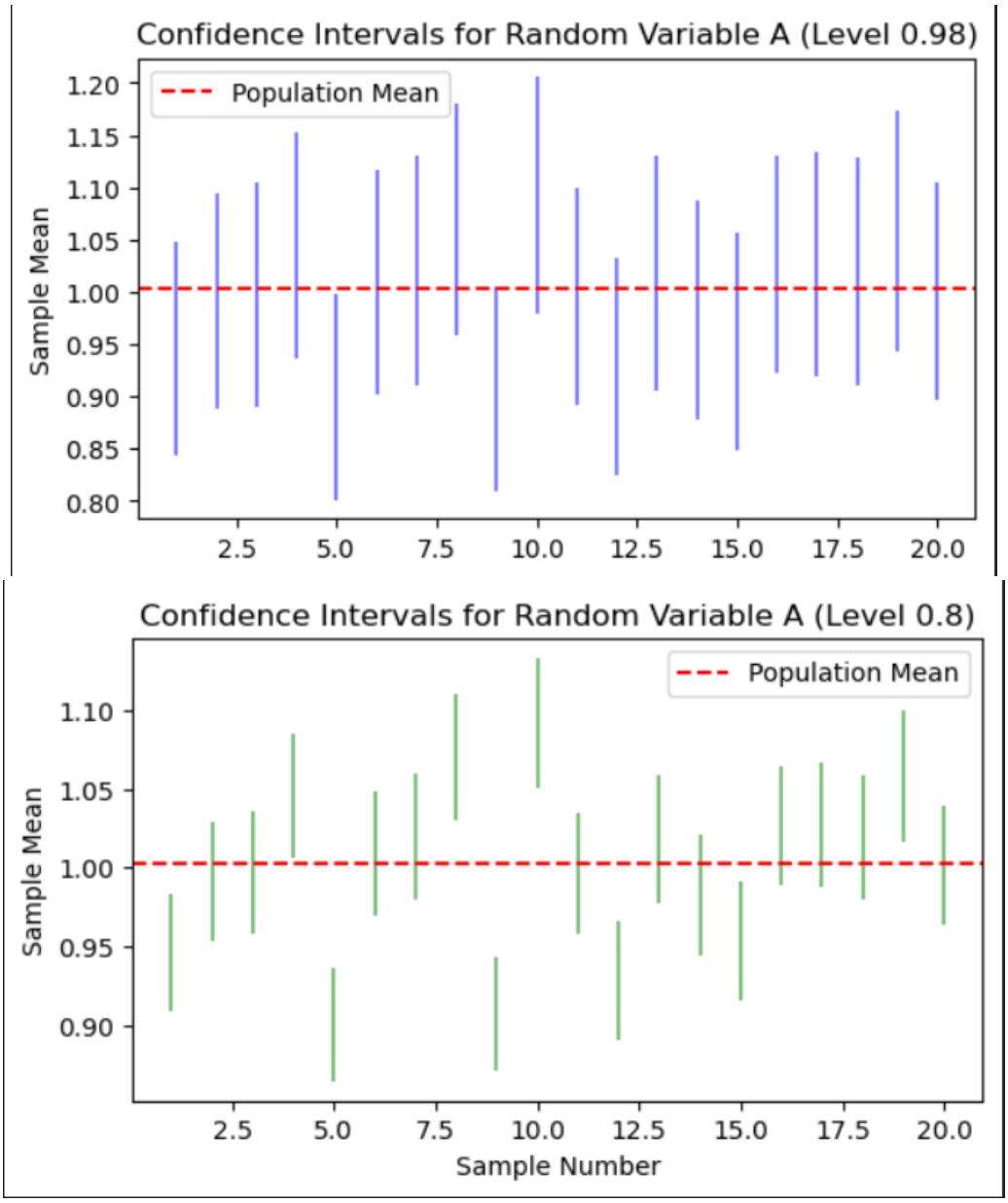
This is the main approach for this method and it is so easy to implement and solve.

3-4) The figures from the Task-2 step 8



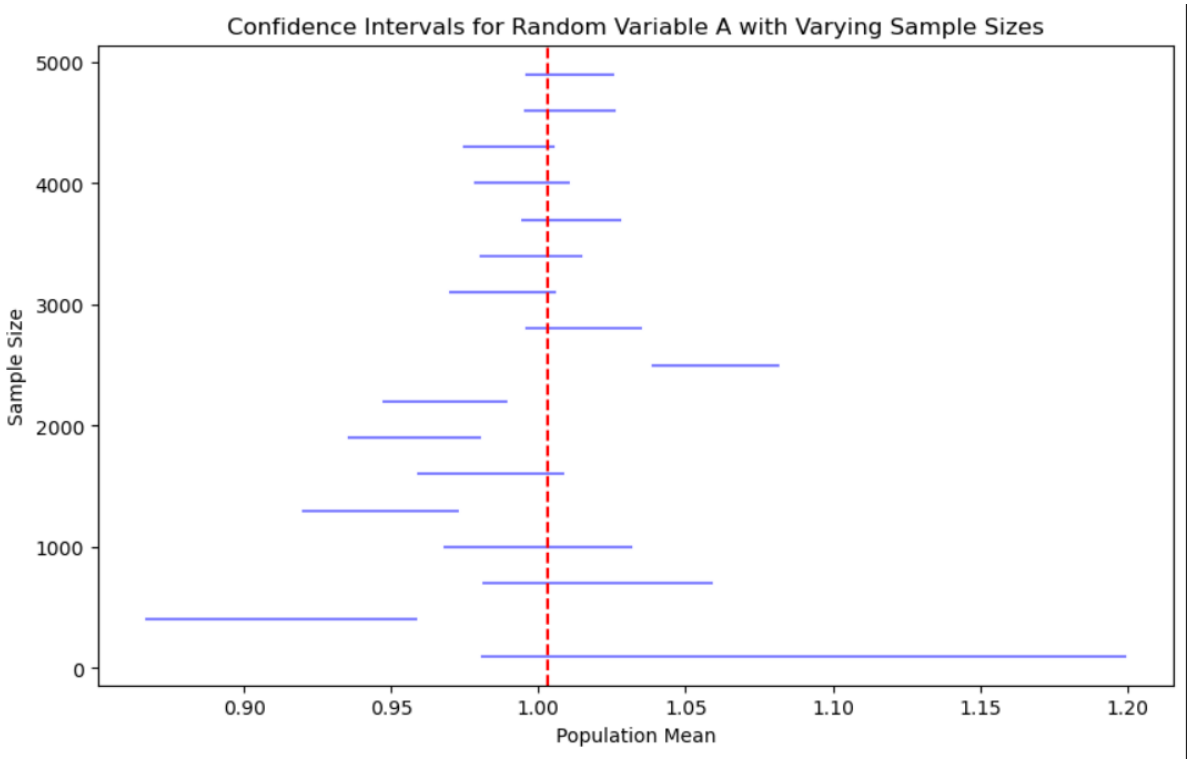
Task-4:

4-1) The figures from the Task-4



Confidence intervals vary depending on confidence levels. As the confidence level increases, the confidence interval generally widens. The higher the confidence level, the larger it will be necessary to allow for a wider range of values to capture the correct parameter value.

4-2) The figures from the Task-4



When sample size is large the variance should be lower, and the results will be closer to the mean when at the bottom of the plot the variance is higher and the confidence interval is more wider but at the top of the plot variance is lower and interval is more narrow.

Task-5:

We will have 2 hypotesis **H0** and **H1** and we need to formulate them.

Null Hypotesis(H0): The exercise frequency has not increased which means

mu <= mu0 (*mu0 is the new exercise frequency*)

Alternative Hypotesis(H1): The exercise freqency has increased.

mu > mu0

Then determine the critical values with the given **significance level 0.03**. Then perform the one tailed test in positive direction. If the sample size is large we should use Z-test and with this method we can easily determine the critical value for the significance level.

Standard deviation is equals to 2

With using this formula **Z = (x-mu)/(std/sqrt(n))** and this is equals to "**3.162** which is test statistic.

After that we can estimate the critcal value with right-tailed method 0.03 significance is 1.8808 and we can easily say that [**1.8808, inf]** and from here we get the last z-score is in this interval so:

The null hypotesis is rejected! Alternative hypotesis is accepted!

Task-6:

Our formula is **P(C | D, E, H) = (P(D, E, H | C) * P(C)) / P(D, E, H)** but in our calculations we dont need to divide with **P(D, E, H)** this its not effect the solution so much and the calculating this value is not easy to determine. So we are not using this. To calculate to **P(C | D, E, H)** we will use the formula:

P(C | D, E, H) = (P(D, E, H | C) * P(C))

So if we use this formula for every category we will have the array which is posterior probabilities and if we apply the "**np.argmax"** to this array we will get the prediction