280201033-Deniz KAYA

CENG222

Probability and Statistics

Term project for 2022-2023 Spring

Report

TASK 1: Generating the population

2) A looks like a geometric population with p value 0.5. Geometric population pmf is:

$P(X) = (1-p)^{x-1}.p\ for\ X = 1, 2, \ldots$

When we insert p = 0.5, the new equation becomes:

$P(X) = 0.5^{x}\ for\ X = 1, 2, \ldots$

and we can define A in terms of X such as: A = X – 1 since X stars from 1 and goes to infinity and A starts from 0 and goes to infinity. With that information, P(A) becomes:

$P(A) = 0.5^{A+1}\ for\ A = 0, 1, 2, \ldots$

which is the given pmf of A.

4) For b to be a continuous random variable, the area under the curve must be 1. So, we will integrate it under these curves and find the values x,y,z,t.Also, both functions value when inserted y should be equal. So:

$$-0.096y^3 + 0.432y^2 - 0.352y + 0.08 = \frac{-2y+11}{15}$$

From here, we get $y = \frac{5}{2}$ or $y = \frac{6+\sqrt{134}}{6}$ or $y = \frac{6-\sqrt{134}}{6}$ but since it must be positive, the last one cannot occur for our solution.

First function's value at x is equal to zero. So:

$$-0.096x^3 + 0.432x^2 - 0.352x + 0.08 = 0$$

From this equation, $x = \frac{1}{2}, \frac{6+\sqrt{21}}{3}, \frac{6-\sqrt{21}}{3}$

Second function's value at z is equal to zero. So:

$$\frac{-2z+11}{15} = 0$$

From this equation, $z = 11/2$.

For their integral to be equal to 1, values must be: x = ½, y = 5/2 and z = 11/2.

$$\int_{1/2}^{\frac{5}{2}} (-0.096b^3 + 0.432b^2 - 0.352b + 0.08)\, db + \int_{\frac{5}{2}}^{11/2} \frac{-2b + 11}{15} db = 1$$

And lastly, t value becomes:

$$\frac{-2y + 11}{15} = \frac{-\frac{2.5}{2} + 11}{15} = \frac{2}{5}$$

10)To find the pdf of E from the cdf of it, we need to take the derivative of cdf.So, the pdf becomes:

$f(e; i, j) = F'(e; i, j)$

$f(e; i, j) = (e^2 - 2ei + i^2 - j)' = 2e - 2i \; for \; (i + \sqrt{j}) \leq e \leq (i + \sqrt{j+1})$

13)We will find the l value for each possible k values. For k = 0.1, the equations becomes:

$f(h) = 0.1, 0 \leq h \leq 1$
$f(h) = l, 1 < h \leq 2$
$f(h) = 0 \; elsewhere.$

The total integral must be 1. So:

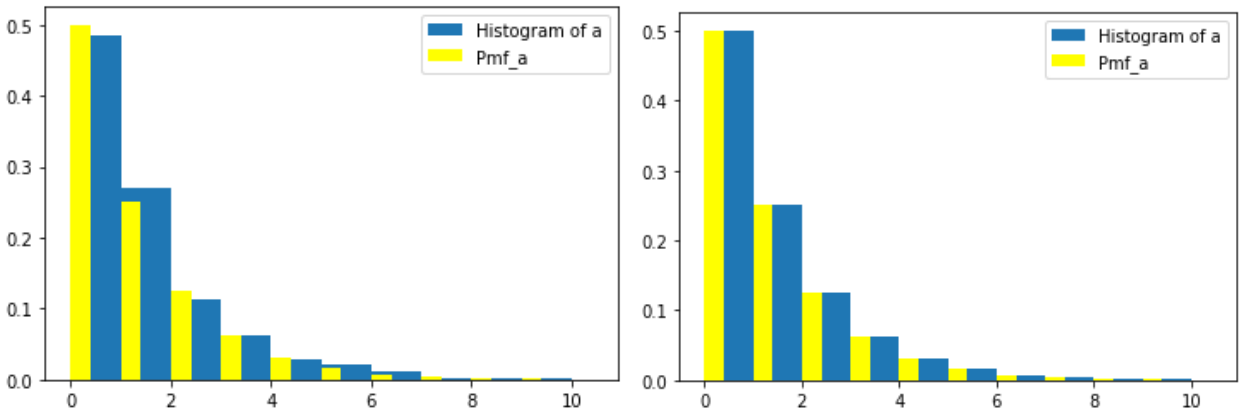
$$\int_0^1 0.1\, dh + \int_1^2 l\, dh = 1$$
$0.1.(1 - 0) + l(2 - 1) = 1$
$0.1 + l = 1$
$l = 0.9$

For k = 0.4, l becomes 0.6 since the area under the curve is simply k for (0,1) and l for (1,2). Similarly, for k = 0.7, l becomes 0.3. So, the generalized pdf becomes:

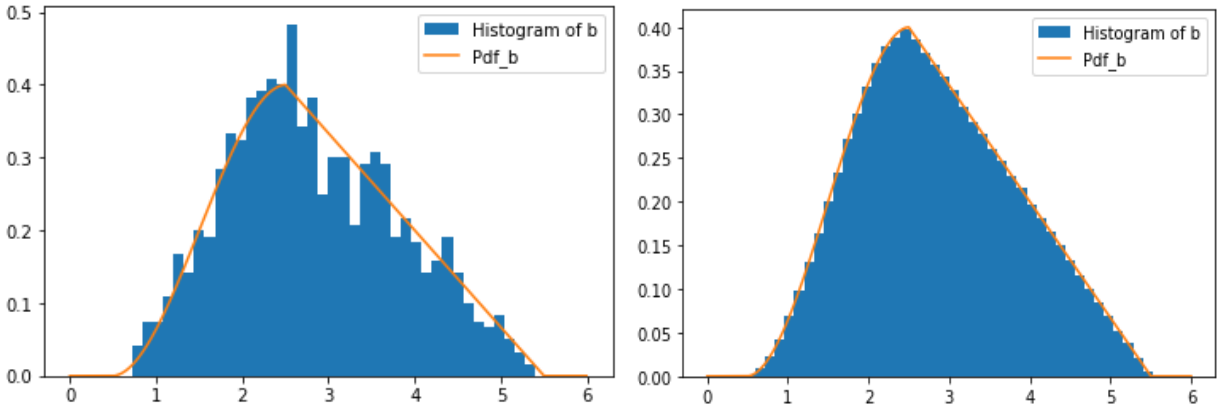
$f(h) = k, 0 \leq h \leq 1$
$f(h) = 1 - k, 1 < h \leq 2$
$f(h) = 0 \; elsewhere.$

19) The graphs on the left is from population size 1000 and the ones on the right are from population size 1000000.

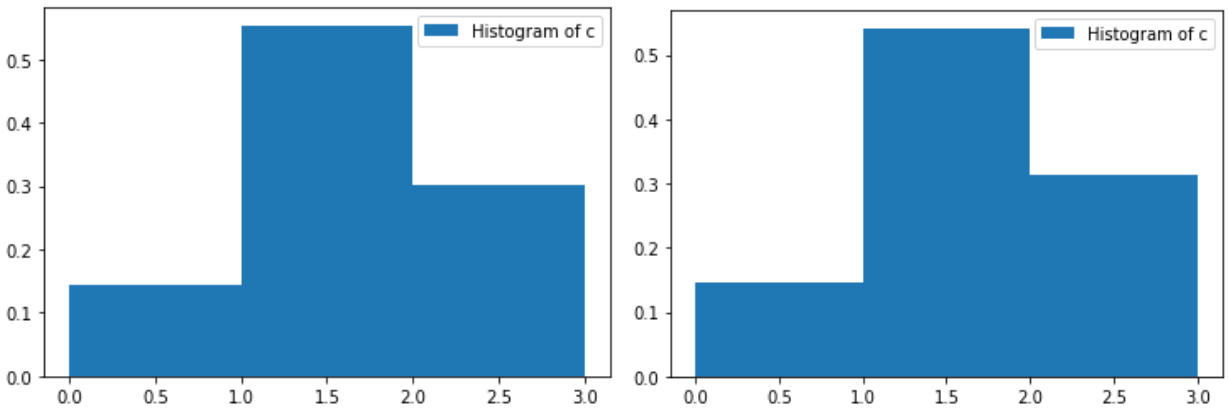
Part A:



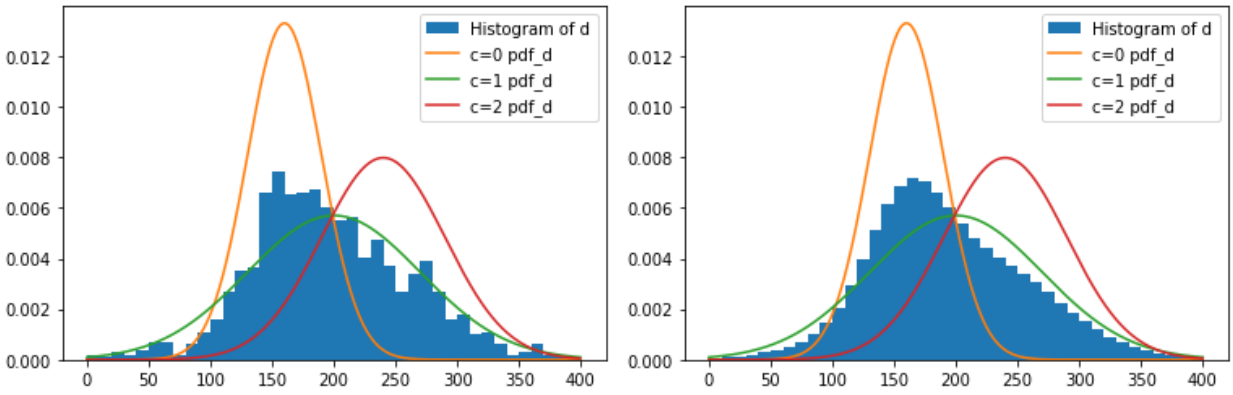
Part B:



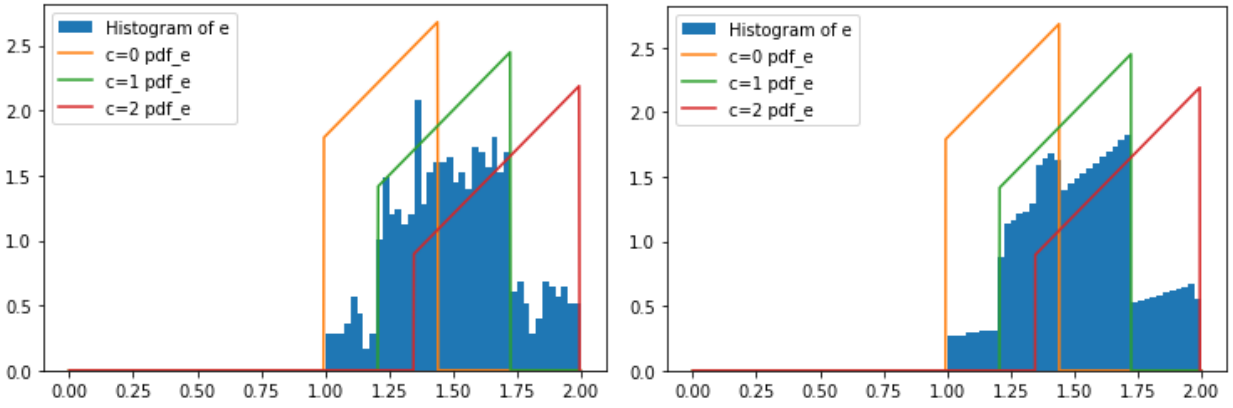
Part C:



Part D:



Part E:



TASK 2: Sampling and Descriptive Statistics

6)Variable A is from a Geometric distribution with p value 0.5 and X = A+1. We know that estimated value of geometric distribution is 1/p so E[X] = 2. If we plug A+1 for X, we get that:

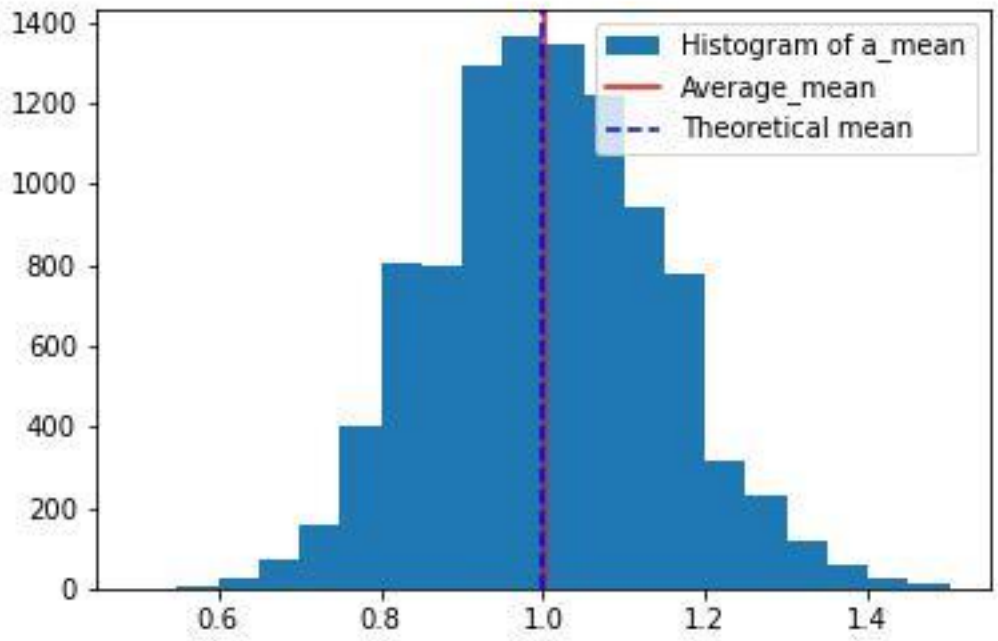
E[A + 1] = 2, E[A] = 1.

For the variance, we know that a geometric distribution's variance is $\frac{1-p}{p^2}$. From here, we get that:

Var(X) = $\frac{1-0.5}{0.5^2} = 2$. If we plug in A+1 for X, we get that:

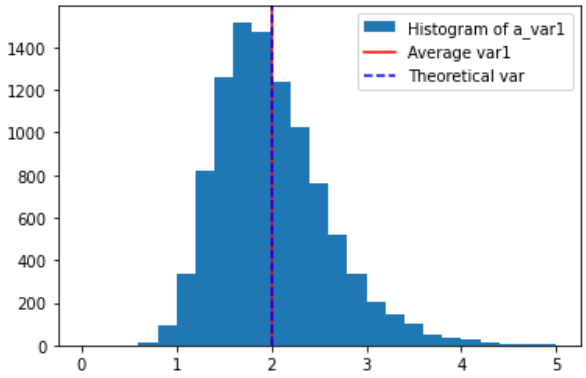
Var(A + 1) = Var(A) = 2.

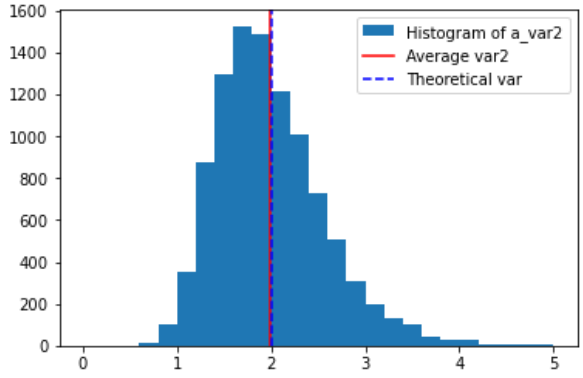
8)Part A:



From the figure we can see that no matter the distribution is we get the mean is, if we take enough samples from it, the distribution of means becomes a normal distribution.

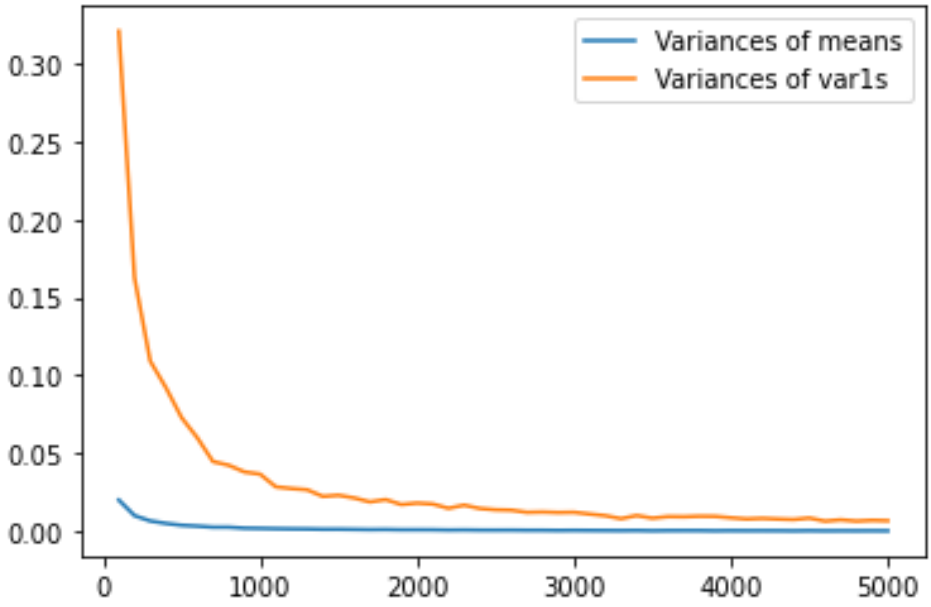
Part b and Part c side by side:





From these figures, we see that dividing the sum to n-1 instead of n gives us a more accurate estimation of variance.

10)



We can see from the graphs that the variance of means and type1 variances converge to zero as the sample size gets closer to the population size.

TASK 3:Parameter Estimation

1)Maximum Likelihood estimation of a normal distribution parameters are:

For the mean of the distribution equals to:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Which is the estimated mean of the distribution. The variance of the distribution equals to:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \hat{\mu})^2$$

Which is the estimated variance but instead of n-1 as divisor, it has n as divisor.

3)To estimate the parameters of distribution E, we first take the first moment which is the sample mean of the distribution.

$$m_1 = \int_{i+\sqrt{j}}^{i+\sqrt{j+1}} e * f(e)\, de$$

We know f(e) from the task1.10. So:

$$m_1 = \int_{i+\sqrt{j}}^{i+\sqrt{j+1}} e * (2e - 2i) de$$

After the calculations, i becomes:

$$i = m_1 - \frac{2}{3}((j+1)^{\frac{3}{2}} - j^{\frac{3}{2}})$$

So our g(j) is:

$$g(j) = m_1 - \frac{2}{3}((j+1)^{\frac{3}{2}} - j^{\frac{3}{2}})$$

We need two equations to find two parameters so we take the second moment.

$$m_2 = \int_{i+\sqrt{j}}^{i+\sqrt{j+1}} e^2 * f(e)\, de$$

$$m_2 = \int_{i+\sqrt{j}}^{i+\sqrt{j+1}} e^2 * (2e - 2i)de$$

After the calculations, equation becomes:

$$m_2 = \frac{4i}{3}\left((j+1)^{\frac{3}{2}} - j^{\frac{3}{2}}\right) + i^2 + j + \frac{1}{2}$$

When we put g(j) instead of i, we get:

$$0 = \frac{4(m_1 - \frac{2}{3}((j+1)^{\frac{3}{2}} - j^{\frac{3}{2}}))}{3}\left((j+1)^{\frac{3}{2}} - j^{\frac{3}{2}}\right) + (m_1 - \frac{2}{3}((j+1)^{\frac{3}{2}} - j^{\frac{3}{2}}))^2 + j + \frac{1}{2} - m_2$$

So our f(j) becomes:

$$f(j) = \frac{4(m_1 - \frac{2}{3}((j+1)^{\frac{3}{2}} - j^{\frac{3}{2}}))}{3}\left((j+1)^{\frac{3}{2}} - j^{\frac{3}{2}}\right) + (m_1 - \frac{2}{3}((j+1)^{\frac{3}{2}} - j^{\frac{3}{2}}))^2 + j + \frac{1}{2} - m_2$$

5)The only parameter for f(h) is k since l is equal to 1-k. This comes from that the area under the pdf must be equal to 1 so from geometry, k + l = 1. So, the pdf becomes:

f(h) = k , 0<x<1

f(h) = 1-k, 1<x<2.

n_1 = number of samples between 0 and 1

n_2 = number of samples between 1 and 2

n = total number of samples

Then, the likelyhood function becomes:

$$\prod_{0<x_i<1} k * \prod_{1<x_i<2} (1-k)$$

$$k^{n_1} * (1-k)^{n_2}$$

If we take the logarithm of this equation, we get:

$$n_1 \log(k) + n_2 \log(1-k)$$

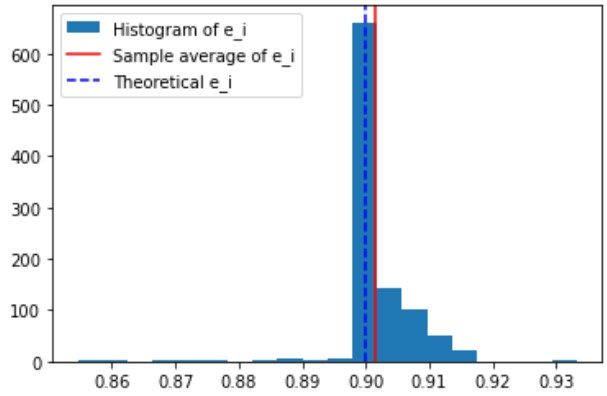
And if we take the derivative of this and equate it to zero, we get:

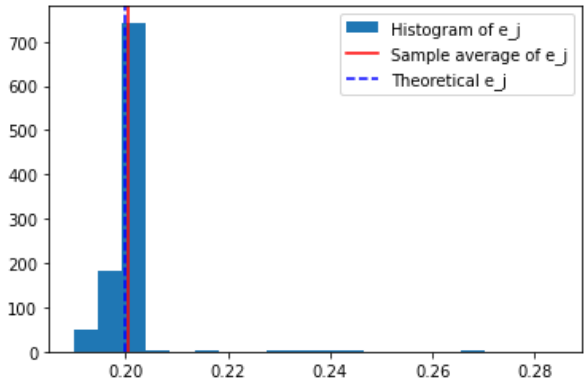
$$\frac{n_1}{k} - \frac{n_2}{1-k} = 0$$

$$k = \frac{n_1}{n}$$

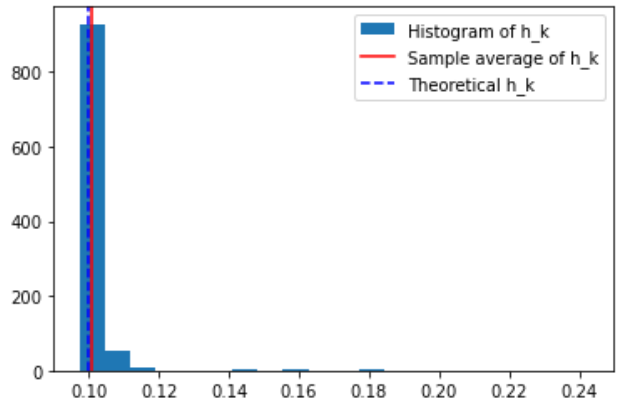
9)

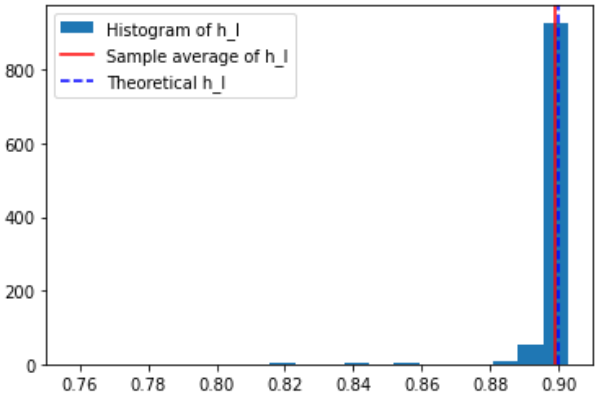
Part A and B in order:





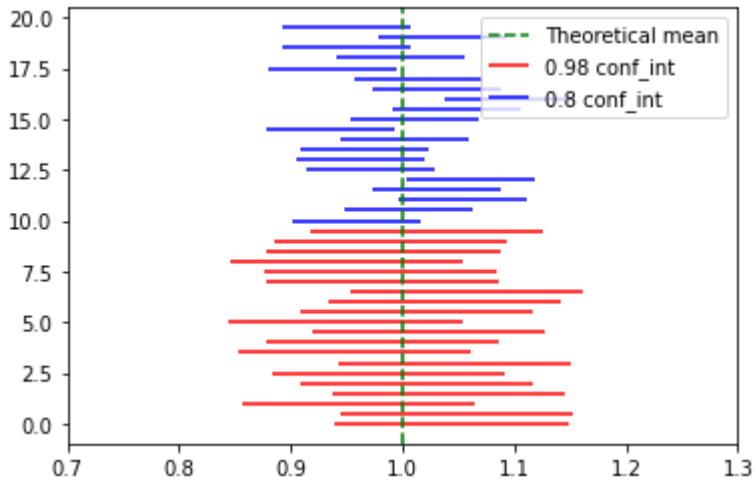
Part C and D in order:





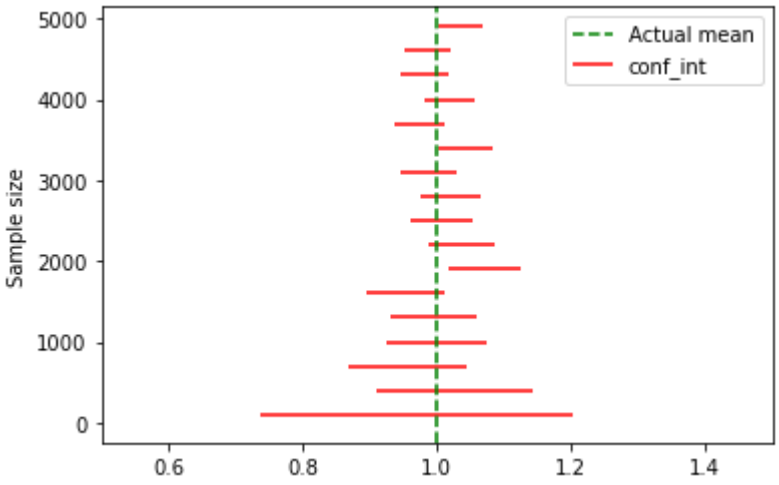
TASK 4:Confidence Intervals

3)



As the confidence level increase, the interval becomes wider but possibility of finding end values of an interval which include the actual mean increases. On the other hand, as the confidence interval decreases, the interval becomes thinner but the possibility of finding end values of an interval which include the actual mean decreases.

5)



As the sample size gets larger, confidence interval gets thinner with the same possibility of finding end values of an interval which includes the actual mean

TASK 5:Hypothesis testing

Our alternative hypothesis h_a is mean being increased. The null hypothesis h_0 is that the mean has not changed.Since the sample size 500 which is big enough, we can assume that sample mean has a normal distribution.This is a one sided test and our significance level is %3 so we will reject a sample if $Z > z_{0.03}$. Value of $z_{0.03}$ is 1.88. Also we know that the distribution's variance is 2 so its standart deviation is $\sqrt{2}$. So, with all these information, we can calculate the Z value using the formula:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{1.2 - 1}{\frac{\sqrt{2}}{\sqrt{500}}} = 3.162$$

Our Z value is 3.162 which falls into rejection region so we reject the null hypothesis meaning that with the given data, we can say that the exercise frequency of patients has increased.

TASK 6:Naive Bayes classifier

2)We are asked to compute P(C|D,E,H) for every possible C values. Using the Bayes Rule, we can rewrite it like:

$$P(C|D,E,H) = \frac{P(D,E,H|C) * P(C)}{P(D,E,H)}$$

Since D,E and H are indepentent from each other, we can rewrite it like:

$$P(C|D,E,H) = \frac{P(D|C) * P(E|C) * P(H|C) * P(C)}{P(D) * P(E) * P(H)}$$

D, E and H are continuous variables so their pdf would be zero so we will rewrite them using very small δs:

$$P(C|D,E,H)$$
$$= \frac{f(D-\delta_1 > D > D+\delta_1|C) * f(E-\delta_2 > E > E+\delta_2|C) * f(H-\delta_3 > H > H+\delta_3|C) * P(C)}{f(D-\delta_1 > D > D+\delta_1) * f(E-\delta_2 > E > E+\delta_2) * f(H-\delta_3 > H > H+\delta_3)}$$

Since δs are very small, we can rewrite them all like:

$$P(C|D,E,H) = \frac{2\delta_1 * f(D|C) * 2\delta_2 * f(E|C) * 2\delta_3 * f(H|C) * P(C)}{2\delta_1 * f(D) * 2\delta_2 * f(E) * 2\delta_3 * f(H)}$$

We can divide both with respective 2δs and the final version of the equation would be:

$$P(C|D,E,H) = \frac{f(D|C) * f(E|C) * f(H|C) * P(C)}{f(D) * f(E) * f(H)}$$