Statistical Inference: Assignment #2

Reyhaneh Ahani
reyhaneh.ahanii@gmail.com

University of Tehran — December 21, 2023

Introduction

Initially, all the theoretical questions are addressed, followed by the programming section that includes output figures and interpretations of the results.

1 Theoretical Section

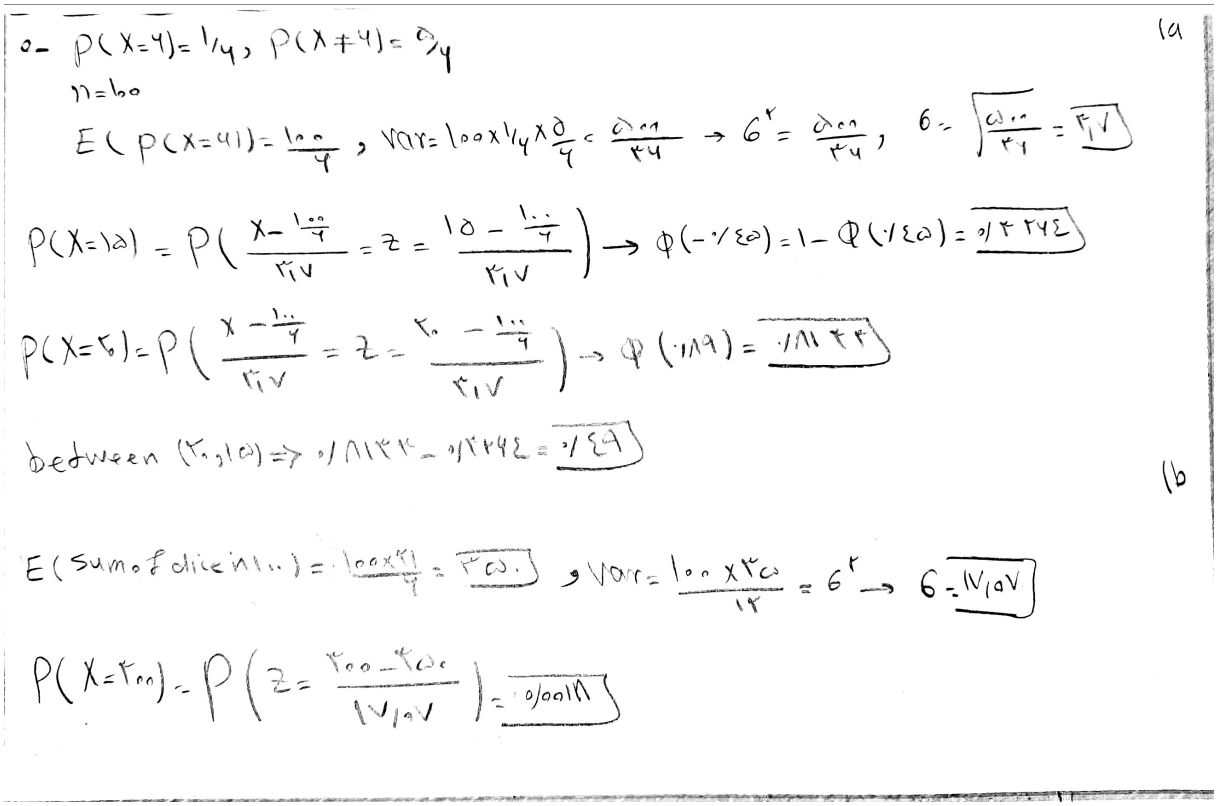


Figure 1: Problem 0

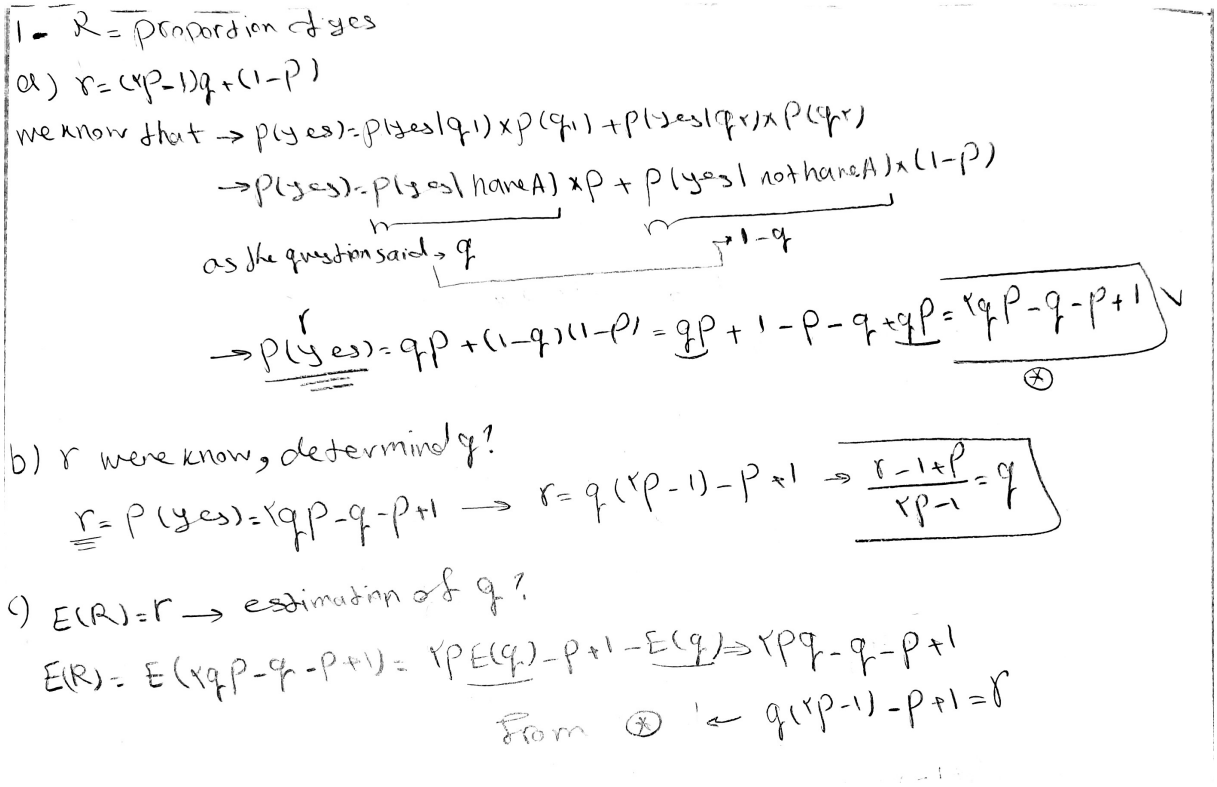


Figure 2: Problem 1) a

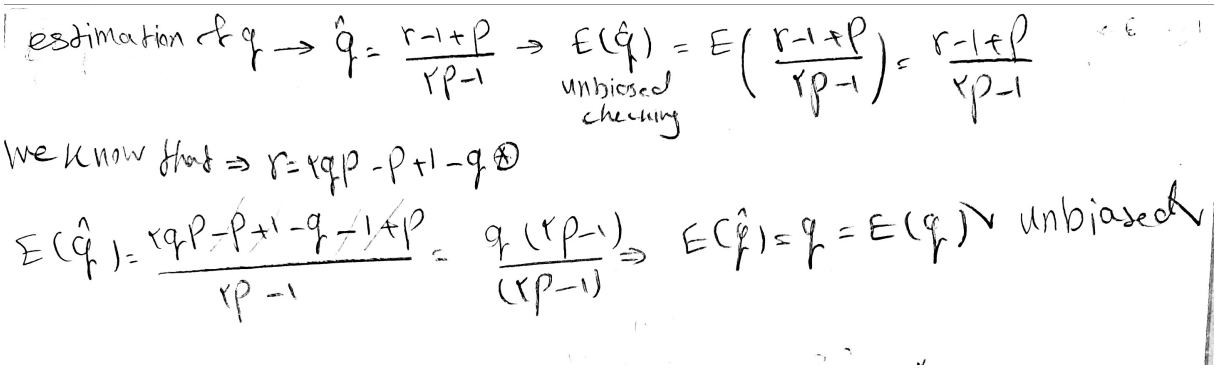


Figure 3: Problem 1) b

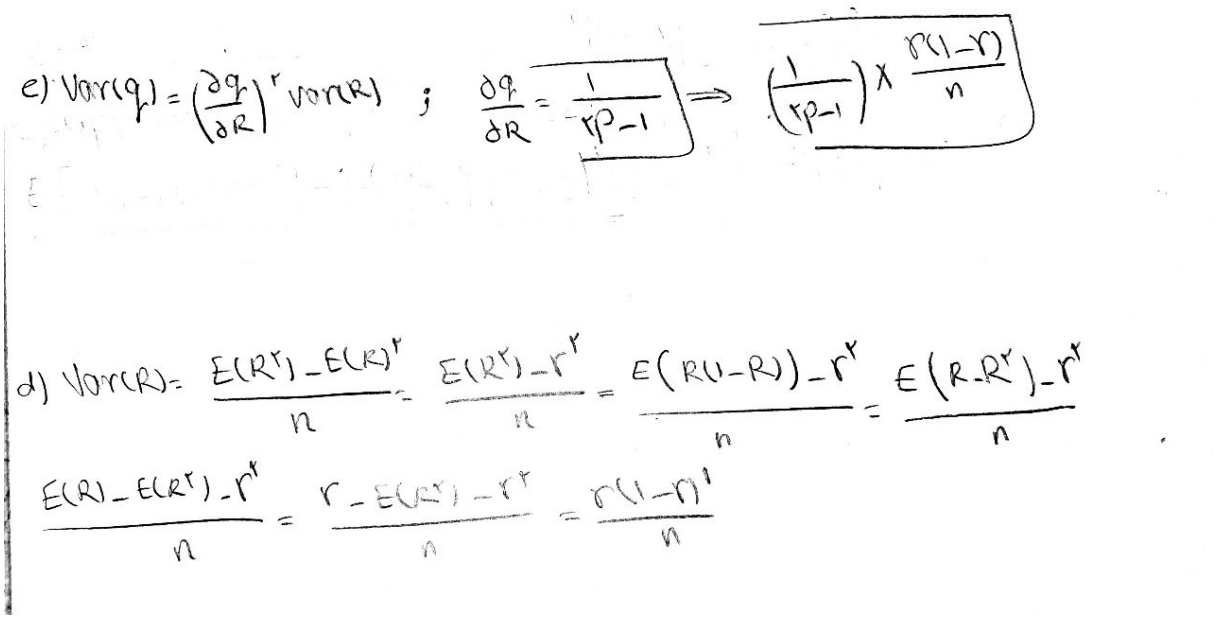


Figure 4: Problem 1) c

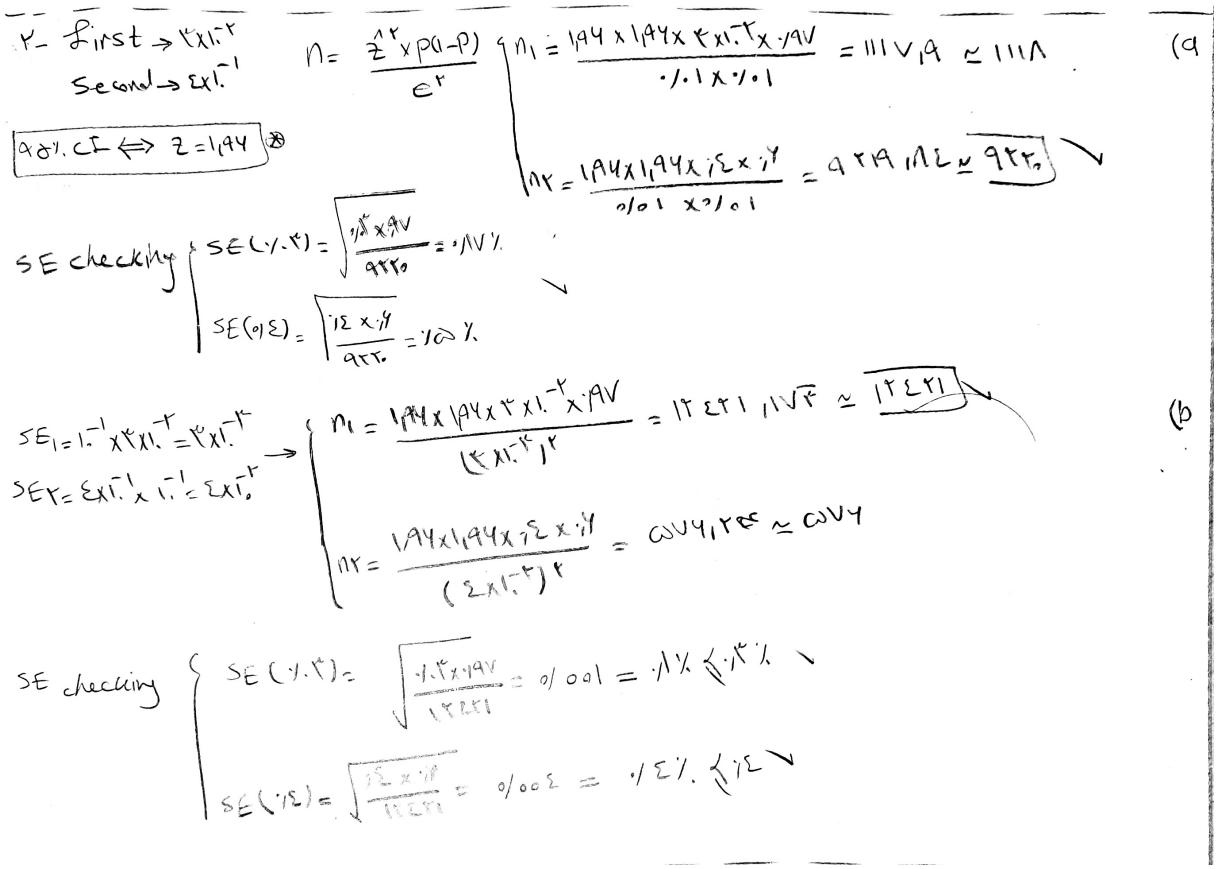


Figure 5: Problem 2

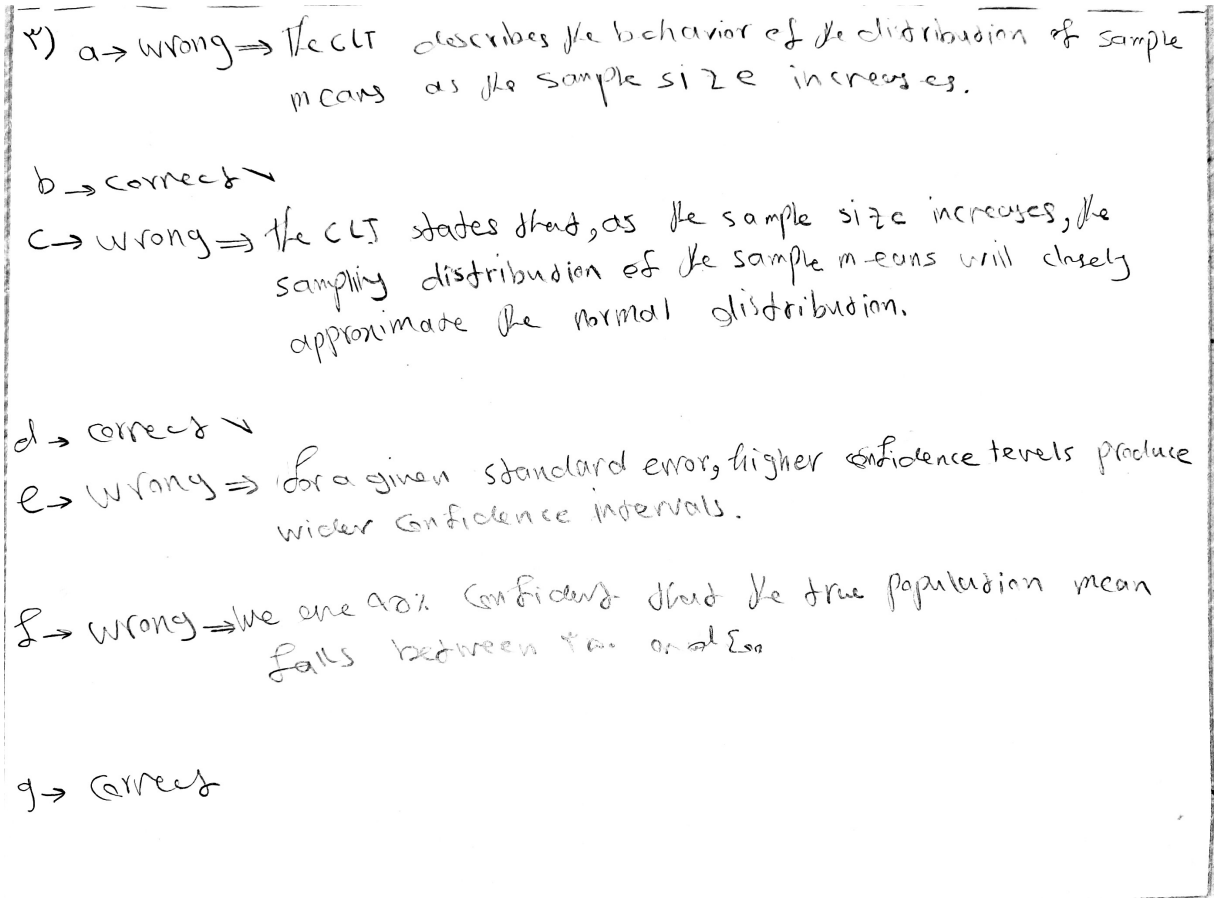


Figure 6: Problem 3) a

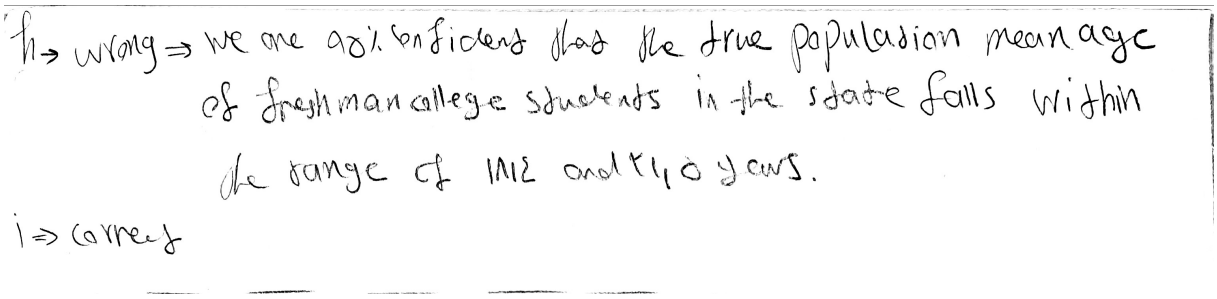


Figure 7: Problem 3) b

۴) $X - Y = W$

$W \sim (\mu_x - \mu_y,\ \sigma_x^2 + \sigma_y^2) \rightarrow Z_{\text{scoring}}(W) = \dfrac{(\bar{X}-\bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} \sim N(0,1)$

marginal error $= Z_{\alpha/2} \times \sqrt{\left(\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}\right)}$, Confidence interval is 98%, $Z = 1{,}44$

$1 = 1{,}44 \times \frac{12{,}12}{\sqrt{n}} \rightarrow n \geq 194$

Figure 8: Problem 4

۵ - $h_0 \Rightarrow \mu_1 = \mu_2$ prove: h_0 is rejected only if the confidence interval for the difference of the means doesn't include zero.
$h_1 \Rightarrow \mu_1 \neq \mu_2$

$h_0 \Rightarrow \mu_1 = \mu_2 \rightarrow$ it has a meaning that there is no difference between μ_1, μ_2

$h_1 \Rightarrow \mu_1 \neq \mu_2 \rightarrow$ it has a meaning that there is a significant difference between μ_1, μ_2

* since confidence interval provides a range of values within the true difference of the population means is likely to fall, so, if confidence interval doesn't contain zero in it, it means that $\mu_1 - \mu_2 \neq 0$ or $\mu_1 \neq \mu_2 \rightarrow$ which means that there is a significant difference between μ_1, μ_2 and it results that h_0 has to be rejected.

Figure 9: Problem 5

۶- under what condition for c_i, $\bar{X}_c$ is an unbiased estimator of population mean?

unbiased checking $\rightarrow E(\bar{X}_c) = \mu$, $\bar{X}_c = \sum_{i=1}^{n} c_i X_i$

$E\left(\sum_{i=1}^{n} c_i X_i\right) = \left(\sum_{i=1}^{n} c_i E(X_i)\right) = \mu \longrightarrow \mu \sum_{i=1}^{n} c_i = \mu \rightarrow \sum_{i=1}^{n} c_i = 1$

condition $\Rightarrow$ if the summation of c_i's is equal to 1, so the $\bar{X}_c$ will be an unbiased estimation of population mean.

Figure 10: Problem 6

2 Programming Section

2.1 Problem 7

2.1.1 Histogram of the population values for cancer mortality

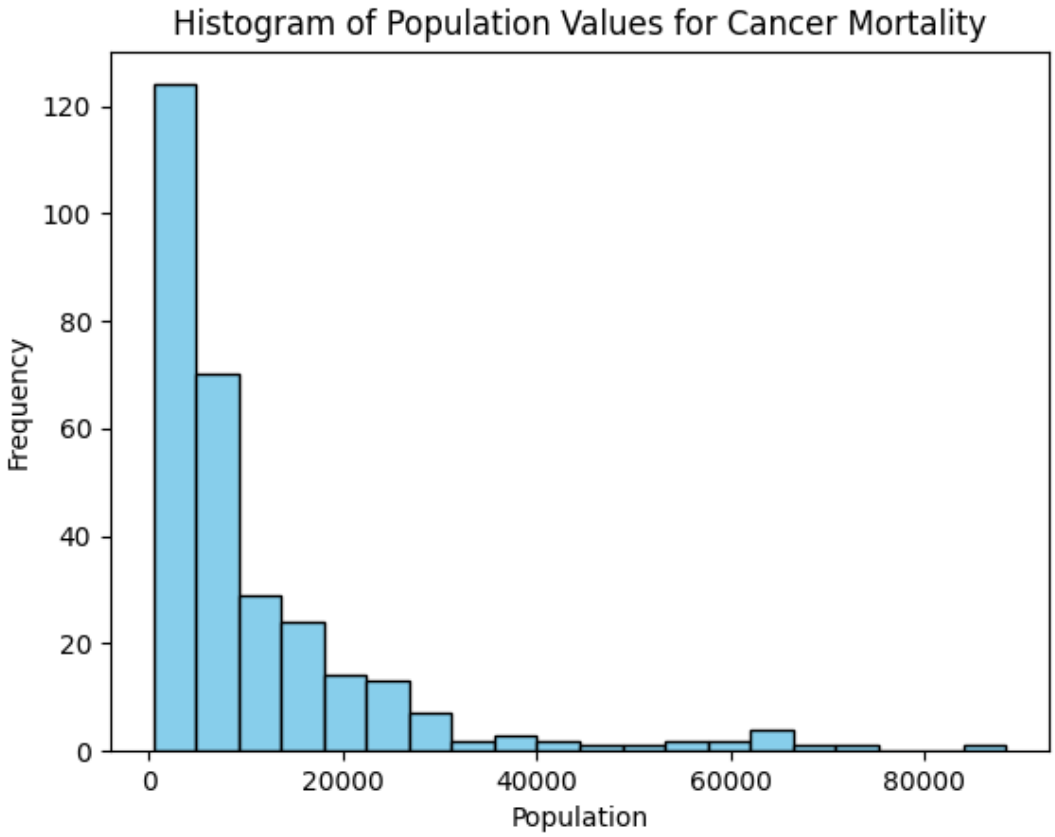


Figure 11: Caption

2.1.2 Population metrics and total cancer mortality

Parameter	Value
Total Cancer Mortality	11997
Population Mean	11288.056478405315
Population Variance	189888678.03346625
Population Standard Deviation	13780.01008829334

Table 1: Population statistics and total cancer mortality

2.1.3 Sampling mean distribution from a set of 25 observations

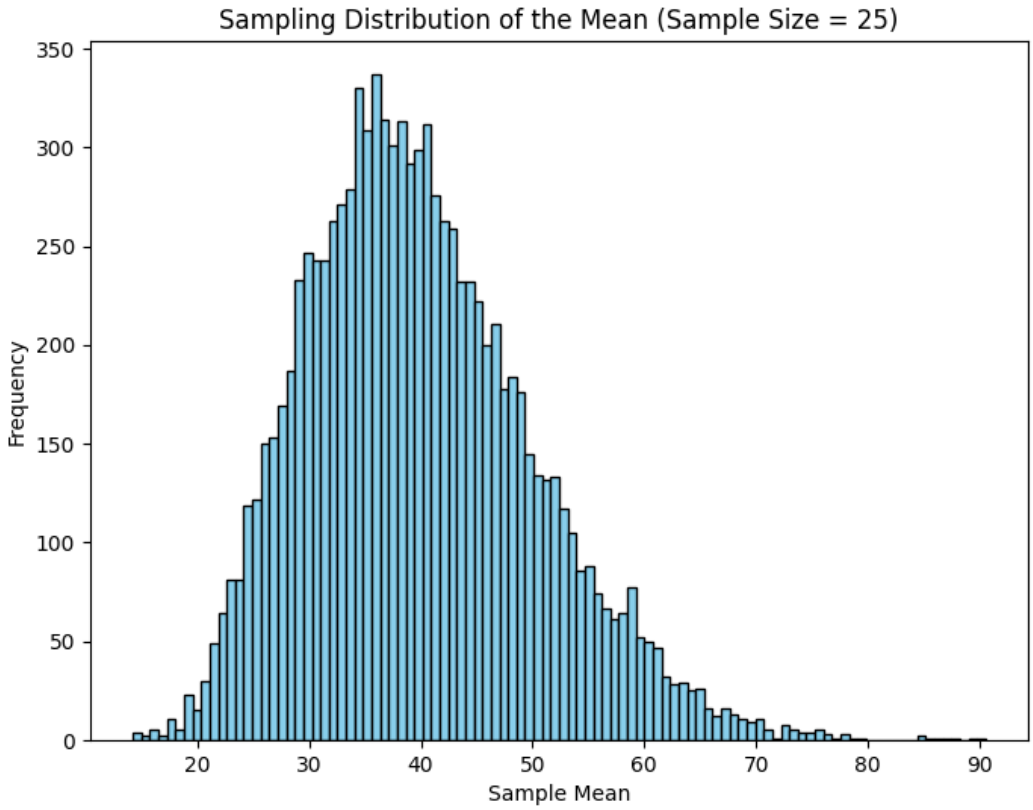


Figure 12: Histogram of sampling mean distribution from a set of 25 observations for N = 10000

2.1.4 Estimate the mean and total cancer mortality from a set of 25 observations

Statistic	Value
Sample Mean	30.92
Sample Total Mortality	773
Actual Mortalities Mean	39.86
Actual Total Mortality	11,997
Estimated Mortalities Mean	30.92
Estimated Total Mortality	9,306.92

Table 2: Estimate the mean and total cancer mortality for N = 25

2.1.5 Estimate the population variance and standard deviation from a set of 25 observations

Statistic	Value
Estimated Population Var	108141607.35
Estimated Population Var	10399.115

Table 3: Estimate the population variance and standard deviation for N = 25

2.1.6 95% confidence intervals

Parameter	Value
Mean Sample Population	8577.16
Mean Population	11288.056
Standard Error	794.266
Margin of Error	1563.039
95% Confidence Intervals	(9725.016, 12851.096)
Intervals Cover Population Values	False

Table 4: 95% confidence intervals statistic Summary for N = 25

2.1.7 Repeat parts (d) through (f) for a sample of size 100

1. part d:

Statistic	Value
Sample Mean	37.35
Sample Total Mortality	3735
Actual Mortalities Mean	39.86
Actual Total Mortality	11,997
Estimated Mortalities Mean	37.35
Estimated Total Mortality	11242.349999999999

Table 5: Estimate the mean and total cancer mortality for N = 100

2. part e:

Statistic	Value
Estimated Population Var	179159221.283
Estimated Population Var	13385.037

Table 6: Estimate the population variance and standard deviation for N = 100

3. part f:

Parameter	Value
Mean Sample Population	10849.02
Mean Population	11288.056
Standard Error	794.266
Margin of Error	1563.039
95% Confidence Intervals	(9725.016, 12851.096)
Intervals Cover Population Values	True

Table 7: 95% confidence intervals statistic Summary for N = 25

2.1.8 Assessing the Effectiveness of Ratio Estimators in Enhancing Cancer Mortality Estimates Using Total County Population Size

Yes, using a ratio estimator based on the total population of each county could be effective in improving cancer mortality estimates. By incorporating the total population information, the ratio estimator helps to account for variations in population size across different counties. This is important because areas with larger populations may naturally have more cancer cases and mortalities, even if the mortality rate per capita is relatively low. The ratio estimator helps normalize the mortality data by considering the population size, providing more accurate and comparable estimates of cancer mortality rates across counties.

2.1.9 Sampling mean distribution from a set of 25 observations for ratio estimators

Part (c) focuses on the mean of the mortality values themselves, while part (i) centers around the ratio of the mean mortality to the mean population size. Information about the population size is incorporated by the ratio estimator, potentially resulting in distinct insights or more efficient estimators, depending on the relationship between mortality and population size. The histograms produced in (c) and (i) mirror these distinct perspectives.

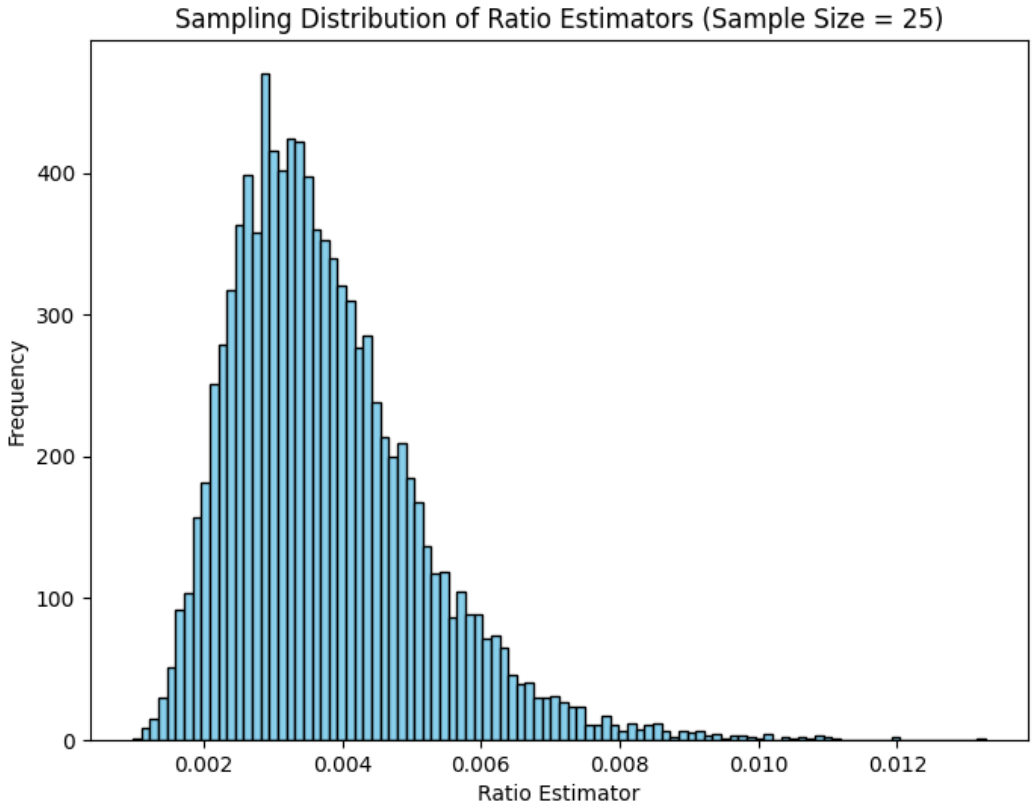


Figure 13: Caption

2.1.10 Stratify the counties into four strata and estimate the Population mean and total mortality

For this section, the data is initially divided into four groups based on population size. Subsequently, the mean population is calculated for each group, and the total mean population is estimated by summing and dividing it into four.

Stratum	Mean Population	Total Mortality
1	1463.33	182.18
2	4241.00	280.93
3	8807.83	915.04
4	26702.33	2452.15
Total Mean Population: 10303.62		

2.1.11 County Stratification, Sampling Fractions, and Variance Comparison in Population Estimation

For this part, the calculation of Sampling Fractions and Variance is performed through different methods.

1. Sampling Fractions:

   Calculates the sampling fractions for two allocation methods:

   - **Proportional Allocation:** Allocates samples proportionally to each stratum based on population size.
   - **Optimal Allocation:** Assumes known variances and allocates samples optimally based on inverse stratum variances.

2. Variance Comparison:

   Compares the variances of population mean estimates obtained using three sampling methods:

   - **Simple Random Sampling:** Randomly selects samples without considering strata.
   - **Proportional Allocation:** Allocates samples proportionally to stratum sizes.
   - **Optimal Allocation:** Allocates samples optimally based on inverse stratum variances.

Stratum	**Proportional Allocation**	**Optimal Allocation**
0	0.043772	2.323703×10^{-6}
1	0.098311	1.003834×10^{-6}
2	0.205553	2.243935×10^{-7}
3	0.652364	3.508007×10^{-9}

Table 8: Sampling Fractions for Proportional and Optimal Allocation

Method	**Variance**
Simple Random Sampling	1.89888678×10^{8}
Proportional Allocation	1.21515088×10^{8}
Optimal Allocation	$3.55543781 \times 10^{-6}$

Table 9: Variances for Different Sampling Methods

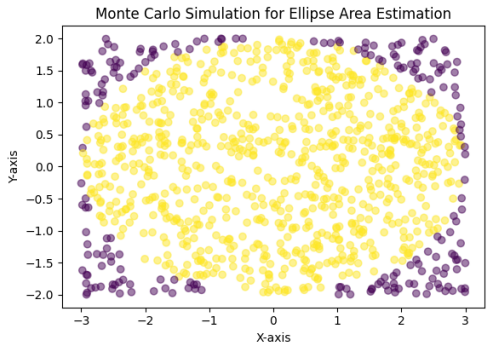
2.2 Problem 8

1. This function is written to estimate the value of π using the Monte Carlo method. The Monte Carlo method is used to calculate certain integrals or differential equations that don't have a closed-form solution.

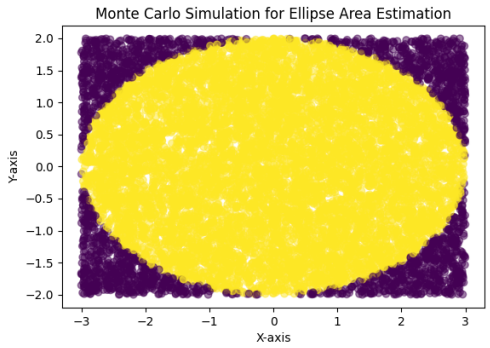
N	Estimator Result
10	2.8
100	3.12
1000	3.112
10000	3.148
100000	3.14664
1000000	3.141812
10000000	3.1420716

Table 10: Estimator Results for Different Values of N

2. The decision was made to calculate the area of an ellipse for the Bonus part using the Monte Carlo method. The approach is found to be very similar to a circle estimator, but now involves major and minor diameters. Therefore, a uniform parameter is assigned for the major diameter, and the other is designated for the minor diameter.



(a) Plot for N=100



(b) Plot for N=1000

Figure 14: Ellipse area estimator

2.3 Bonus Problem

The mean of Z can be calculated using the law of total expectation:

$$E[Z] = E[Z \mid X \leq T] \times P(X \leq T) + E[Z \mid X > T] \times P(X > T) \tag{1}$$

$$E[Z \mid X \leq T] = \mathrm{E[X \mid X \leq T]} = \alpha \left(1 - e^{-\frac{T}{\alpha}}\right) \tag{2}$$

$$P(X \leq T) = 1 - e^{-\frac{T}{\alpha}} \tag{3}$$

$$E[Z \mid X > T] = T \tag{4}$$

$$P(X > T) = T \tag{5}$$

Therefore:

$$E[Z] = \alpha \left(1 - e^{-\frac{T}{\alpha}}\right)^2 + T \times e^{-\frac{T}{\alpha}} \tag{6}$$

The variance of Z can be calculated using the formula:

$$\mathrm{Var}(Z) = E\left(Z^2\right) - [E(Z)]^2 \tag{7}$$

$$E\left(Z^2\right) = 2\alpha^2 - 3\alpha^2 e^{-\frac{T}{\alpha}} + 2\alpha^2 e^{-\frac{T}{\alpha}} + T^2 e^{-\frac{T}{\alpha}} \tag{8}$$

Being simplified and expanded upon:

$$\mathrm{Var}(Z) = \alpha^2 e^{-\frac{T}{\alpha}} - 2\alpha T e^{-\frac{T}{\alpha}} + T^2 e^{-\frac{T}{\alpha}} \tag{9}$$