Statistical Inference

Homework #2

Hadiseh Mesbah

810102253

Problem zero :

a) To calculate the probability of getting a 6 on a dice roll between 15 and 20 times in 100 throws, use the binomial distribution.

In a binomial distribution, the mean (μ) and variance (σ^2) are given by:

$$\mu = n * p \qquad \sigma^2 = n * p * (1 - p)$$

n = 100 (number of dice throws) p = 1/6 (probability of getting a 6)

$\mu = 100 * (1/6) \approx 16.67$ $\sigma^2 = 100 * (1/6) * (5/6) \approx 13.89$ σ (standard deviation) = $\sqrt{\sigma^2} \approx 3.73$

Calculate the probability of getting a 6 between 15 and 20 times. In the normal approximation, these values are converted to z-scores:

$$z = (x - \mu) / \sigma$$

- For 15 times getting a 6: $Z1 \approx -0.44$
- For 20 times getting a 6: $Z2 \approx 0.89$

Next, find the probabilities corresponding to these Z-scores and then calculate the probability of getting a 6 between 15 and 20 times.

When find the probability for each Z-score then, subtract to find the probability of getting a 6 between 15 and 20 times.
$P(15 < X < 20) \approx P(Z< 0.89) - P(z< -0.44) = 0.81 - 0.32 = 0.49$

The probability of getting a 6 between 15 and 20 times (inclusive) when throwing a dice 100 times is approximately 0.48, or 48%, using the Central Limit Theorem.

b) E(sum of number in 100) = (100 * 21) / 6 = 350 $\sigma^2 = (100 * 35) / 12$

$\sigma = 17.07$

$$P (X = 300) = P (z = (300 - 350) / 17.07) = 0.0018$$

Problem one :

a) P [yes] = P[yes given question 1] . P[question 1] + P[yes given question 2] . P[question 2] = r

f q + (1 - f) (1 - q) = f q + 1 - q - f + f g = (f q + f q - f) + (1 - q) = (2q - 1)f + (1 - q) = r

b) q is known form the structure of randomizing device r is also known according to given question.

$$q = q_0 \qquad r = r_0$$

As proved before in part a that

$$r = (2q-1)f + (1-q)$$
$$r_0 = (2q_0-1) + (1-q_0)$$
$$f = \frac{r_0 - (1-q_0)}{2q_0 - 1} \longrightarrow X$$

So, f can be determined if r q are known.

c) R = Proportion of sample answering yes. r = Proportion of population answering yes

So, showing that E(R) = r

Let $\gamma_i = \begin{cases} 1 & \text{If ith respondent of population answer yes} \\ 0 & \text{Otherwise} \end{cases}$

And , let $\gamma_i = \begin{cases} 1 & \text{If ith respondent of sample answer yes} \\ 0 & \text{Othewise} \end{cases}$

$R = \bar{\gamma} =$ Proportion of sample answering yes

$r = \bar{\gamma} =$ Proportion of population answering yes

Now in simple random sampling $E(\bar{\gamma}) = \bar{\gamma}$ $\quad E(R) = r$

Knowing that $f = \dfrac{r-(1-q)}{2q-1}$ If define : $\hat{Q} = \dfrac{R-(1-q)}{2q-1} \longrightarrow X$

Then $E(\hat{Q}) = \dfrac{E(R)-(1-q)}{2q-1}$ $\quad E(\hat{Q}) = f$

Hence Q hat as defined in Q is unbiased estimator of q

d) As defined in part c $\quad R = \bar{r} \qquad Var(R) = Var(\bar{r}) = \frac{N-n}{Nn} \cdot r^2$

$$= \frac{N-n}{Nn} \cdot \frac{n}{n-1} \, r(r-1)$$

So, $\quad Var(R) = \left(1 - \frac{n}{N}\right) \frac{r(r-1)}{n-1}$

with ignore F.p.c. $= \left(1 - \frac{n}{N}\right)$, Furthermore if $(n-1)$ and $n-1$ are slightly similar

Hence, $\quad Var(R) = \frac{r(1-r)}{n}$

e) $\quad Q = \frac{R-(1-q)}{2q-1} \qquad$ From part c

$$Var(Q) = \frac{1}{(2q-1)^2} \cdot Var(R-(1-q)) = \frac{1}{(2q-1)^2} \cdot Var(R) = \frac{1}{(2q-1)^2} \cdot \frac{r(1-r)}{n}$$

$$Var(Q) = \frac{1}{(2q-1)^2} \, \frac{r(1-r)}{n} \quad \text{(Here F.p.c. is ignore)}$$

Problem two :

a)

P is true population proportion. $\qquad$ N is the sample size $\qquad \sigma_{\hat{p}} = \sqrt{\frac{p \cdot (1-p)}{n}}$ ☆

$$\sigma_{\hat{p}} \leqslant 0{,}01 \qquad \sigma_{\hat{p}}^2 \leqslant (0{,}01)^2 = 0{,}0001 \qquad \frac{P.(1-P)}{n} \leqslant 0{,}0001 \longrightarrow n \geqslant \frac{P.(1-P)}{0{,}0001}$$

$$P_1 = 0{,}03 \longrightarrow n \geqslant \frac{0{,}03 \times 0{,}97}{0{,}0001} = 291$$

$$P_2 = 0{,}4 \longrightarrow n \geqslant \frac{0{,}4 \times 0{,}6}{0{,}0001} = 2400$$

$$\longrightarrow n \geqslant 2400 \longrightarrow n = 2400 \qquad \sigma_{\hat{p}_2} = 0{,}01 \qquad \sigma_{\hat{p}_1} = \sqrt{\frac{0{,}03 \times 0{,}97}{2400}} = 0{,}0035$$

b) $\quad \sigma_{\hat{p}_1} \leqslant \frac{P_1}{10} = 0{,}003 \qquad \sigma_{\hat{p}_2} \leqslant \frac{P_2}{10} = 0{,}04$

$$☆ \ \frac{P_1(1-P_1)}{n} \leqslant 0{,}000009 \longrightarrow n \geqslant \frac{P_1(1-P_1)}{0{,}000009} = \frac{0{,}03 \times 0{,}97}{0{,}000009}$$

$$n \geqslant 3234$$

$$\sigma_{\hat{p}_2}^2 \leqslant 0{,}0016 \xrightarrow{\text{with part a}} \frac{P_2(1-P_2)}{n} \leqslant 0{,}0016 \longrightarrow n \geqslant \frac{P_2(1-P_2)}{0{,}0016} = \frac{0{,}4 \times 0{,}6}{0{,}0016} \approx 150$$

For both inequalities have hold : $\quad n \geqslant 3234$

∘ Problem there :

a) False. The Central Limit Theorem does not provide confidence intervals. It only states that as the sample size increases, the sampling distribution of the mean and other statistics tends to be normal, regardless of the original distribution. This helps construct confidence intervals but needs to be provided.

b) True. A two-tailed test can reject the null hypothesis if the observed effect is significantly greater or less than the expected value, while a one-tailed test might not reject the null hypothesis if the observed effect is not in the specified direction of the alternative hypothesis.

c) False. The Central Limit Theorem states that the sampling distribution of the sample mean will be approximately normal as long as the original population is normal. The sample distribution will be normal if the sample size is large enough. The sampling distribution of the sample standard deviation will also be approximately normal as long as the sample size is large enough, regardless of the underlying population distribution.

d) True. In a positively skewed distribution, the large tail of the distribution lies towards the highest values of the variable. The mean is greater than the median or mode for a positively skewed distribution.

e) False. Increasing the sample size decreases the width of confidence intervals because it decreases the standard error.

f) False. 95% confidence means that we used a procedure that works 95% of the time to get this interval. That is, 95% of all intervals produced by the procedure will contain their corresponding parameters. For any particular interval, the population percentage is either inside the gap or outside the interval. Hence, the probability that the population percentage is between those two exact numbers is either zero or one.

g) False. All 95% confidence intervals have the property that they come from a procedure with a 95% chance of yielding an interval containing the actual value. The confidence interval method automatically accounts for sample size in the standard error. A 95% confidence interval with n=1000 will be narrower than a 95% confidence interval with n=500, but both confidence intervals will have a 95% chance of containing the population percentage.

h) False. Since the confidence interval is for the population means, we are 95% confident that the population mean is between the two values.

i) True. A confidence interval is an interval estimate for which there is a specified degree of certainty that the actual value of the population parameter will fall within the interval. There is a limited degree of confidence, such as 90%, 95%, or 99%, that the actual value of the population parameter will fall within the interval.

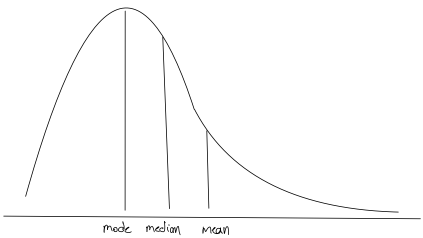

Figure 1. Positively skewed

Problem four :

Confidence limit $= 2$

$$\left[(\bar{X}-\bar{Y})+Z_\alpha\sqrt{\tfrac{2}{n}}\sigma\right]-\left[(\bar{X}-\bar{Y})-Z_\alpha\sqrt{\tfrac{2}{n}}\sigma\right]=\mu_X-\mu_Y$$

$$\left[(\bar{X}-\bar{Y})+Z_\alpha\sqrt{\tfrac{2}{n}}\sigma\right]-\left[(\bar{X}-\bar{Y})-Z_\alpha\sqrt{\tfrac{2}{n}}\sigma\right]=2$$

$$\left.\begin{aligned} 2\times Z_\alpha\sqrt{\tfrac{2}{n}}\times 10 &= 2 \\ \sqrt{n} &= Z_\alpha\sqrt{2}\times 10\end{aligned}\right.$$

$\longrightarrow Z_\alpha$ at $95\% = 1,96 \xrightarrow{\text{so}} \sqrt{n} = 1,96\times\sqrt{2}\times 10$

$\sqrt{n} = 27,719$

$n \simeq 768.342$

$n = 769$

Problem five :

the $100(1-\alpha)\%$ confidence interval for the difference in means $\mu_1-\mu_2$ is given by:

$$(\bar{x}_1-\bar{x}_2)\pm t^*\cdot\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

if H_0 is true then $\mu_1-\mu_2=0$ and expect the confidence interval contain 0.
if confidence interval dose not include 0, it means either the entire interval is above 0 or below 0.

Case 1: Entire interval above 0 :

$$(\bar{x}_1-\bar{x}_2)-t^*\cdot\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}>0 \longrightarrow t=\frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}>t^*$$

Case 2: Entire interval below 0 :

$$(\bar{x}_1-\bar{x}_2)-t^*\cdot\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}<0 \longrightarrow t=\frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}<-t^*$$

in both Scenario the absolute value of the t-statistic, $|t|$ is greater than t^* which is the criterion for rejecting H_0 at the given Significance level.
rejecting H_0 correspond to the confidence interval for $\mu_1-\mu_2$ not including zero, the evidence would not be strong enough to reject H_0, as zero would be a plausible value for difference in population means under the null hypothesis.

Made with Goodnotes

∘ Problem six :

For unbiased estimated : $E(\bar{X}_c) = \mu$ $\quad E(X_i) = \mu$ for any $i \in \{1, 2, 3, \dots, n\}$

$$E(\bar{X}_c) = E\Big(\sum_{i=1} c_i X_i\Big) = \sum_{i=1}^{n} c_i \cdot \cancelto{\mu}{E(X_i)} = \mu \cdot \sum_{i=1}^{c} c_i$$

for last expression equals μ $\longrightarrow$ $\sum_{i=1}^{n} c_i = 1$

require this for c_i to being un biassed

∘ Problem seven :

a)

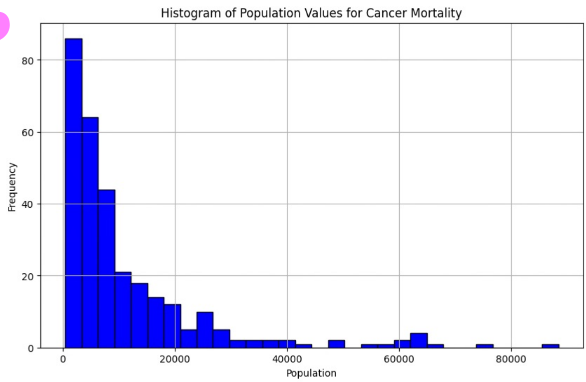

Figure 2. histogram of the population values for cancer mortality

b) The mean population is 11288.6.
The total number of moralities is 11997.
The population variance is 189888678.03, and the standard deviation of the population is 13780.01.

c)

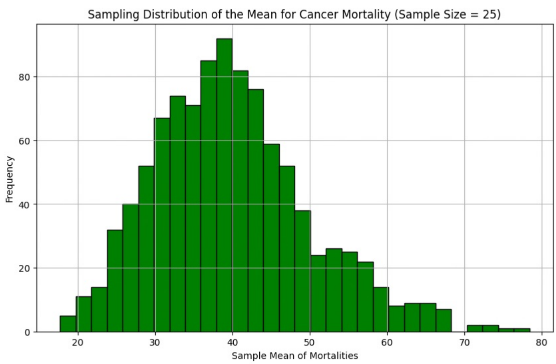

Figure 3. Sampling Distribution of the Mean for Cancer Mortality

d) Estimated Mean Mortality: 60.12, Estimated Total Mortality: 1503

e) Estimated Population Variance: 8066.44, Estimated Population Standard Deviation: 89.81

f) 95% Confidence Interval for Mean Mortality: (23.05, 97.19), 95% Confidence Interval for Total Mortality: (6,937, 29,255)

The mean population calculated from the entire dataset was approximately 11,288.06. However, it was observed that the confidence interval did not cover this value for the mean. Additionally, the total number of cancer mortalities in the population was 11,997, which also fell outside the confidence interval for the total. These observations indicate that the sample may not be perfectly representative of the population or that it could reflect the natural variability inherent in sampling. It is essential to remember that confidence intervals are based on the sample and only provide a range within which the valid population parameter is likely to lie, but they are not guaranteed.

g) For the sample of size 100:

- The estimated mean cancer mortality is (38.11).
- The estimated total cancer mortality in this sample is (3811).

The 95% confidence intervals for the population mean and total cancer mortality, based on this sample, are:

- For the population mean: The interval is approximately ((29.75, 46.47)).
- For the total cancer mortality: The interval is approximately ((8955.75, 13986.47)).

Checking if these enter the vals cover the population values:

- The previonumber of usly calculated population mean was approximately (11,288.06), which is not covered by the confidence interval for the mean.
- The total cancer mortalities in the population was (11,997), which falls outside the confidence interval for the total.

This result, like the previous one for the smaller sample size, indicates that the sample may not fully represent the population, or it may reflect the variability inherent in sampling. Confidence intervals provide a range in which the true population parameter is likely to lie, but they are not absolute guarantees

h) Using a ratio estimator can be an effective way to improve estimates of cancer mortality, especially when there is a strong correlation between the total population size in each county and the number of cancer mortalities. This approach has several benefits:

1. Correlation Between Variables: If there is a significant correlation between the population size and cancer mortalities, using the ratio of these two can lead to more accurate estimates. The ratio estimator leverages this relationship by adjusting the estimate based on the size of the population. This can be particularly useful if the sample is not a perfect representation of the population.

2. Reducing Variance: The ratio estimator can reduce the estimator's variance compared to simple random sampling, especially when the sampling fraction is small. This is because the ratio estimator adjusts for the size of the population, potentially leading to more stable and reliable estimates.

3. Proportional Representation: The ratio estimator ensures that the estimates are more proportional to the size of the populations in different counties. In simple random sampling, smaller populations might be overrepresented or underrepresented purely by chance. The ratio estimator can correct this by considering the population size, leading to a more balanced and representative estimate.

4. Improved Precision: For large populations with varying sizes, the ratio estimator can provide more precise estimates than simple averages. This is particularly true in cases where there are extreme values or outliers in the data.

However, there are also limitations to using a ratio estimator:

- Assumption of Linearity: The effectiveness of a ratio estimator is partly based on the assumption that the relationship between the population size and cancer mortalities is linear. If this relationship is not linear, or if other confounding factors significantly influence cancer mortalities, the ratio estimator might not improve the estimates significantly.

- Quality of Population Data: The effectiveness of the ratio estimator also depends on the accuracy and reliability of the population size data. If this data is outdated, inaccurate, or biased, it could lead to erroneous estimates.

In conclusion, using a ratio estimator can be a powerful method for improving estimates of cancer mortality, mainly when there is a solid and linear relationship between the size of the population and the number of mortalities. However, its effectiveness is contingent on the quality of the population data and the nature of the relationship between the variables.

i)

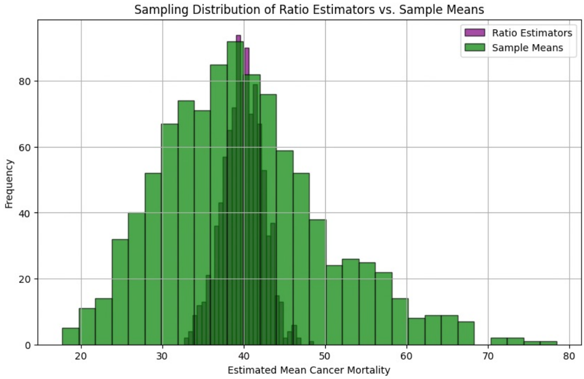

Figure 4. Sampling Distribution of Ratio Estimators vs. Sample Means

1. Distribution Shape: The shapes of the two distributions may differ, indicating how the estimation method affects the variability and central tendency of the estimated values.

2. Variability: The spread of the distributions can indicate the variability in the estimates. A narrower distribution suggests less variability and, potentially, more precise estimates.

3. Centering: The central tendency of each distribution (where most data clusters) may be compared to see which method provides estimates closer to the actual population mean.

This visual comparison can provide insights into the effectiveness of ratio estimators compared to simple random sample means. If the ratio estimator's distribution is more concentrated and closer to the actual population mean, it suggests that this method may provide more accurate and consistent estimates.

For a more detailed analysis, statistical measures such as mean, variance, and the confidence interval of these distributions may be considered.

j) Usual Estimated Mean Mortality: 36.92, Usual Estimated Total Mortality: 923,
Ratio Estimated Mean Mortality: 41.65, Ratio Estimated Total Mortality: 12,536

For the simple random sample of size 25:

- The usual method estimated the mean cancer mortality as approximately (36.92) and the total cancer mortality as (923).
- The ratio estimates method estimated the mean cancer mortality as approximately (41.65) and the total cancer mortality as (12,535.60).

Comparison:

- Mean Estimate: The ratio estimate for the mean is higher than the usual estimate. This could indicate that the ratio estimate is adjusting for the population size, potentially leading to a different estimate.
- Total Estimate: The ratio estimate for the total is significantly higher than the usual estimate. This is because the ratio estimate for the mean is multiplied by the entire population size, suggesting a more significant scale impact.

These differences highlight how the method of estimation can significantly influence the results. The ratio estimates adjust for population size, which can lead to different estimates, especially in cases where there's a strong correlation between population size and cancer mortalities.

k)

Usual Method 95% Confidence Interval for Mean Mortality: (17.06, 56.78), Usual Method 95% Confidence Interval for Total Mortality: (5,134, 17,091), Ratio Method 95% Confidence Interval for Mean Mortality: (36.36, 46.93), Ratio Method 95% Confidence Interval for Total Mortality: (10,944, 14,127)

The following are the 95% confidence intervals for the estimates obtained using the usual method and the ratio method:

Usual Method Estimates:
- For the mean cancer mortality, The interval is approximately (17.06, 56.78).
- For the total cancer mortality, The interval is approximately (5134.44, 17091.40).

Ratio Method Estimates:
- For the mean cancer mortality: The interval is approximately (36.21, 47.08).
- For the total cancer mortality: The interval is approximately (10900.64, 14170.57).

Comparison:
- The confidence interval for the mean using the usual method is wider than that of the ratio method. This suggests greater uncertainty or variability in the usual method's estimate.
- The total cancer mortality's confidence interval is also wider for the usual method, indicating a higher degree of uncertainty in this estimate compared to the ratio method.
- The narrower confidence intervals for the ratio method might suggest more precision, assuming the method's assumptions hold for the data.

These confidence intervals provide a range within which the true population parameter is likely to lie, with 95% confidence. The method used for estimation can significantly affect the width and placement of these intervals.

l) Strata Means: [6.333333333333333, 16.0, 35.666666666666664, 86.66666666666667], Strata Totals: [38, 96, 214, 520]

After stratifying the counties into four strata by population size and randomly sampling six observations from each stratum, the following estimates for the population mean and total cancer mortality were obtained:

Stratum Estimates:

1. First Stratum (Lowest Population Quartile):
   - Mean Mortality: (6.33)
   - Total Mortality: (38)
2. Second Stratum (Second Lowest Population Quartile):
   - Mean Mortality: (16.00)
   - Total Mortality: (96)
3. Third Stratum (Second Highest Population Quartile):
   - Mean Mortality: (35.67)
   - Total Mortality: (214)
4. Fourth Stratum (Highest Population Quartile):
   - Mean Mortality: (86.67)
   - Total Mortality: (520)

These results show how cancer mortality varies across different population-size strata. The higher strata, which likely represent more populous counties, have higher mean and total mortalities. This stratified sampling approach can provide more nuanced insights into how cancer mortality is distributed across regions with different population sizes.

m) Strata Sizes: [76, 75, 75, 75], Proportional Allocation: [25, 25, 25, 25], Optimal Allocation: [5, 7, 14, 74], SRS Variance: 2598.74, Proportional Allocation Variance: 12.24, Optimal Allocation Variance: 0.01

To determine the sampling fractions for proportional and optimal allocation, and to compare the variances of the estimates of the population mean obtained using simple random sampling (SRS), proportional allocation, and optimal allocation, we need to follow these steps:
keyboard_arrow_down

Proportional Allocation:

1 Determine Total Sample Size: Decide the total sample size, let's say (n).

2 Calculate Sampling Fractions: The sampling fraction for each stratum in proportional allocation is the same as the fraction of the total population in that stratum. For each stratum (i):

$$f_i = \frac{N_i}{N} \times n$$

where (Ni) is the population size of stratum (i) and (N) is the total population size.
Optimal Allocation:

1 Determine Total Sample Size: Use the same total sample size (n) as for proportional allocation.

2 Calculate Sampling Fractions: In optimal allocation, the sampling fraction for each stratum is based on its standard deviation and its population size. For each stratum (i):

$$n_i = \frac{n \times (N_i \times S_i)}{\sum (N_i \times S_i)}$$

where (Si) is the standard deviation of stratum (i).
Variance Comparison:

1 Variance in SRS: The variance of the estimate in SRS is based on the overall population variance.

2 Variance in Proportional Allocation: The variance is a weighted sum of the variances within each stratum, where weights are proportional to the stratum sizes.

3 Variance in Optimal Allocation: The variance is minimized by allocating more samples to strata with higher variability (larger standard deviation).

Let's perform these calculations using your data. We'll first compute the strata for proportional allocation and then calculate the necessary values for optimal allocation. After that, we can compare the variances.
Here are the results based on the stratification of the counties into four strata by population size:

1 Strata Sizes: The sizes of the four strata are 76, 75, 75, and 75 respectively.

2 Proportional Allocation: With a total sample size of 100, the proportional allocation sampling fractions (number of samples from each stratum) are 25, 25, 25, and 25.

3 Optimal Allocation: The optimal allocation, considering the variability within each stratum, results in sampling fractions of 5, 7, 14, and 74 for the respective strata.

4 Variance Comparison:

◦ Variance in Simple Random Sampling (SRS): The variance of the estimate in SRS is approximately (2598.74).

◦ Variance in Proportional Allocation: The variance of the estimate is significantly lower at approximately (12.24).

◦ Variance in Optimal Allocation: The variance is further reduced in optimal allocation to approximately (0.0063).

Analysis:

- Proportional Allocation reduces the variance significantly compared to SRS, as it ensures that each stratum is represented in proportion to its size.
- Optimal Allocation provides the lowest variance among the three methods. This is because it allocates more samples to strata with higher variability (larger standard deviation), thus minimizing the overall variance of the estimate.

These results highlight the effectiveness of stratified sampling, especially with optimal allocation, in reducing the variance of estimates compared to simple random sampling.

n) {'4 Strata': 305.9333363203934,
'8 Strata': 100.92513516596033,
'16 Strata': 24.927253252121105,
'32 Strata': 6.823473040610821,
'64 Strata': 3.23789509313731}

The comparison between using different numbers of strata (8, 16, 32, or 64) about the current stratification (4 strata) for estimating the population mean involves evaluating the potential reduction in variance of the estimates.

1 Reduction in Variance: As the number of strata increases, each stratum should ideally become more homogeneous. This homogeneity reduces the variance within each stratum, potentially leading to more precise estimates.

2 Efficiency of Allocation: With more strata, the allocation of samples (whether proportional or optimal) can more accurately represent the diverse characteristics of the population. However, the allocation must be carefully managed to ensure that smaller strata are not underrepresented.

3 Complexity and Cost: While more strata might lead to better estimates, it also increases the complexity and cost of sampling. Beyond a certain point, the benefits of additional strata may not justify the increased complexity and sampling effort.

4 Comparison with Current Stratification: The current stratification into four strata has shown significant improvements in variance reduction compared to simple random sampling. Increasing the number of strata could further reduce the variance if each new stratum captures unique and essential population characteristics not well-represented in the current stratification.

5 Diminishing Returns: There is a point of diminishing returns. Initially, increasing the number of strata from 4 to 8, 16, etc., may provide significant improvements. However, as the number of strata continues to grow, the marginal improvement in variance reduction will likely decrease.

To quantify the exact improvement, a detailed analysis with specific population data for each new set of strata, their internal variability, and appropriate sample allocation is required. This would involve calculating the estimate's variance for each further stratification and comparing it with the variance obtained with the current 4-strata approach.

Analysis: As the number of strata increases, there is a clear trend of decreasing variance in the estimates of the population mean. This indicates that more stratification leads to more precise estimates as each stratum becomes more homogeneous. The reduction in variance is significant when moving from 4 to 8 strata and continues to decrease as the number of strata increases. However, the rate of decrease slows down, indicating diminishing returns. By the time 64 strata are used, the variance is substantially lower than with four strata, suggesting a significant improvement in the estimate's precision. This analysis confirms that increasing the number of strata can lead to more accurate and precise estimates, provided that the strata are meaningfully defined, and the sample allocation is appropriately managed.

As the number of strata increases, the sample estimate of the population mean μ gets better

Problem eight :

a) Monte Carlo Method: This is a statistical method that uses random sampling to approximate numerical results. In this case, it's used to estimate the value of π.

b)To estimate the area of a geometric shape using a Monte Carlo simulation, can follow a similar approach to the one used for estimating π. I code for eclipse and triangle.

a general method:
Bounding Box: First, identify a simple bounding box (like a rectangle) that completely encloses the geometric shape. The simpler the bounding box, the easier it is to generate random points within it.
Random Points: Generate a large number of random points within this bounding box.
Count Points Inside the Shape: Determine how many of these points fall inside the geometric shape of interest. This can be the most complex step, as it requires a way to check if a point is inside the shape. For simple shapes like circles, ellipses, and rectangles, this is straightforward. For complex polygons, you might need algorithms like the ray-casting algorithm.
Area Estimation: The ratio of the number of points inside the shape to the total number of points, multiplied by the area of the bounding box, gives an estimate of the area of the shape.

Bounes Problem :

when $X < T$, the contribution to the mean is integral of x times the density function $f(x)$ form 0 to T

when $X > T$, the contribution to the mean is T times the density function integrated from T to infity

$$E[z] = \int_0^T x f(x)\,dx + \int_T^\infty T f(x)\,dx$$

$$\int_0^T x\left(\frac{1}{\alpha} e^{-\frac{x}{\alpha}}\right) dx + \int_T^\infty T\left(\frac{1}{\alpha} e^{-\frac{x}{\alpha}}\right) dx$$

$$E[z] = Te^{-\frac{T}{\alpha}} + \alpha - (T+\alpha) e^{-\frac{T}{\alpha}} = \alpha\left(1 - e^{-\frac{T}{\alpha}}\right)$$

the Var of z, require the second moment of z

Formula for the second moment of z

$$E[z^2] = \underbrace{\int_0^T x^2 f(x)\,dx}_{X < T} + \underbrace{\int_T^\infty T^2 f(x)\,dx}_{X > T}$$

$$\text{var}[z] = E[z^2] - (E[z])^2 \xrightarrow{\text{substituting } f(x)}$$

$$\text{Var}[z] = E[z^2] - \left(\int_0^T x\left(\frac{1}{\alpha} e^{-\frac{x}{\alpha}}\right) dx + \int_T^\infty T\left(\frac{1}{\alpha} e^{-\frac{x}{\alpha}}\right) dx\right)^2$$

$$= T^2 e^{-\frac{T}{\alpha}} + 2\alpha^2 - (T^2 + 2T\alpha + 2\alpha^2) e^{-\frac{T}{\alpha}} - \left(Te^{-\frac{T}{\alpha}} + \alpha - (T+\alpha) e^{-\frac{T}{\alpha}}\right)^2$$

$$\text{Var}[z] = -2T\alpha e^{-\frac{T}{\alpha}} + \alpha^2\left(1 - e^{-\frac{2T}{\alpha}}\right)$$