Statistical Inference HW#2

Student Name:
Pouya Haji Mohammadi Gohari

SID:810102113

Date of deadline
Saturday 23rd December, 2023

Dept. of Computer Engineering

University of Tehran

Contents

1 Problem 0

1.1 Get 6, between 15 to 20 times

Cause of having large n we can use CLT[1] to find probability we can get 6, between 15 to 20 times. For each sample we can say it is a Bernoulli distribution as follow:

$$X_i \sim Benroulli(p = \frac{1}{6}) \tag{1}$$

Note that we can say the chance or probability of getting 6 for each sample is $\frac{1}{6}$ so it is a Bernoulli distribution now due to 100 samples were taken sum of the sample distributions will be Binomial distribution:

$$Y = \sum_{i=1}^{n} X_i \sim Binomial(n = 100, p = \frac{1}{6}) \tag{2}$$

Now lets obtain μ and σ from Binomial.

$$\begin{aligned} \mu_y &= E\left[\sum_{i=0}^{n} X_i\right] \\ &= \sum_{i=1}^{n} E[X_i] \\ &= np \\ \sigma_y &= \sqrt{\sum_{i=1}^{n} var[X_i]} \\ &= \sqrt{np(1-p)} \end{aligned} \tag{3}$$

Now from CLT we can show that for large sample size that:

$$\frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim Normal(0, 1) \text{ as n} \to \infty \tag{4}$$

We have obtain from equation 3 that $\mu_y = \frac{100}{6}$ and $\sigma_y = \frac{10\sqrt{5}}{6}$ so in order to find the probability of getting 6, 15 to 20 times will be:

$$\begin{aligned} P\left(15 \leq \overline{X_n} \leq 20\right) &= \Phi\left(\frac{20 - \frac{100}{6}}{\frac{10\sqrt{5}}{6}}\right) - \Phi\left(\frac{15 - \frac{100}{6}}{\frac{10\sqrt{5}}{6}}\right) \\ &= \Phi\left(\frac{2}{\sqrt{5}}\right) - \Phi\left(\frac{-1}{\sqrt{5}}\right) \\ &= 0.814 - (0.327) \\ &= 0.487 \end{aligned} \tag{5}$$

[1] Central Limit Theorem

1.2 sum of the seen numbers less than 300

In this problem let X_i be the number seen in the i^{th} so it is random due to what number we are going to see so the wanted-probability is:

$$P\left(\sum_{i=1}^{100} X_i < 300\right) \tag{6}$$

At first we are going to see the mean and standard deviation of each distribution in order to find wanted-probability:

$$\begin{aligned}
E[X_i] &= \sum_{i=1}^{6} x_i P(X_i = x_i) \\
&= \frac{1}{6}\sum_{i=1}^{6} i \\
&= \frac{1}{6}\left(\frac{6*7}{2}\right) \\
&= 3.5 \\
V[X_i] &= \sum_{i=1}^{n} x_i^2 P(X_i = x_i) - \mu^2 \\
&= \frac{1}{6}\sum_{i=1}^{n} i^2 - \mu^2 \\
&= \frac{1}{6}\left(\frac{6*7*13}{6}\right) - \mu^2 \\
&= \frac{91}{6} - \frac{49}{4} \\
&= \frac{35}{12}
\end{aligned} \tag{7}$$

lets change the equation 6 in better way to speak about CLT for this:

$$\begin{aligned}
&P\left(\sum_{i=1}^{100} X_i < 300\right) = \\
&P\left(\frac{\sum_{i=1}^{100} X_i}{100} < 3\right) = \\
&P\left(\overline{X}_{100} < 3\right)
\end{aligned} \tag{8}$$

As you can see we reach $\overline{X}_n$ and from CLT we can easily solve the problem, but calculating μ and σ is our priority:

$$\begin{aligned} E[\overline{X}_n] &= \frac{1}{n}\sum_{i=1}^{n} E[X_i] \\ &= \frac{1}{100}100E[X_i] \\ &= 3.5 \\ \sigma_{\overline{X}_n} &= \frac{1}{n}\sum_{i=1}^{n} Var[X_i] \\ &= \frac{\sigma_{X_i}}{\sqrt{n}} \\ &= \frac{\sqrt{\frac{35}{12}}}{10} \end{aligned} \tag{9}$$

Now lets get back to equation 8:

$$\begin{aligned} P\left(\overline{X}_n < 3\right) &= P\left(\frac{\overline{X}_n - \mu}{\sigma_{\overline{X}_n}} < \frac{3-\mu}{\sigma_{\overline{X}_n}}\right) \\ &= P(Z < \frac{-0.5}{\frac{\sqrt{\frac{35}{12}}}{10}}) \\ &= \Phi\left(\frac{-0.5}{\frac{\sqrt{\frac{35}{12}}}{10}}\right) \\ &= \Phi\left(-2.927700219\right) \\ &\sim 0.002 \end{aligned} \tag{10}$$

2 Problem 1

2.1 show $r = (2p-1)q + (1-p)$

Information from question:

- R: proportion of a sample answering yes
- p: probability that statement 1 is responded to
- 1-p: probability that statement 2 is responded to
- q: proportion of the population that has characteristic A.
- 1-q: proportion of the population that has not characteristic A.
- r: probability that a respondent answers yes.

From hint we show the function is true:

$$\begin{aligned} P(yes) &= P(yes|quesion1)p(question1) + p(yes|question2)p(question2) \\ r &= qp + (1-q)(1-p) \\ r &= 2pq + 1 - p - q \\ r &= q(2p-1) + 1 - p \end{aligned} \tag{11}$$

2.2 Determine q

From equation 11 we can determine q:

$$q = \frac{r + p - 1}{2p - 1} \tag{12}$$

2.3 $E[R] = r$

From definitions r is probability that a respondent answers yes and R is proportion of a sample answering yes so we can say that(let each person has probability r to answer yes):

$$P(R_i) = \begin{cases} r & \text{if answer yes meaning is } R_i = 1 \\ 1 - r & \text{if answer no meaning is } R_i = 0 \end{cases} \tag{13}$$

So we can say that R is:

$$R = \frac{\sum_{i=1}^{n} R_i}{n} \tag{14}$$

and also

$$E[R] = \frac{1}{n} \sum_{i=1}^{n} E[R_i] \tag{15}$$

At first we must calculate the $E[R_i]$ in order to prove the given equation:

$$\begin{aligned} E[R_i] &= P(yes) * 1 + P(no) * 0 \\ &= r \end{aligned} \tag{16}$$

From equations 15 and 16 we can prove it:

$$\begin{aligned} E[R] &= \frac{1}{n} \sum_{i=1}^{n} \\ &= \frac{1}{n} n E[R_i] \\ &= E[R_i] \\ &= r \end{aligned} \tag{17}$$

Now we must propose an estimate $\hat{Q}$ for q, to propose this estimator from equation 12 instead of r we choose to put R in this equation(note that p is known):

$$\begin{aligned} \hat{Q} &= \frac{R + p - 1}{2p - 1} \rightarrow \\ E[\hat{Q}] &= \frac{E[R + p - 1]}{2p - 1} \\ &= \frac{E[R] + p - 1}{2p - 1} \\ &= \frac{r + p - 1}{2p - 1} \\ &= q \end{aligned} \tag{18}$$

So this estimator is unbiased cause to $E[\hat{Q}] = q$.

2.4 $Var[R]$

From equation 14 can obtain $Var[R]$:

$$\begin{aligned} Var[R] &= \sum_{i=1}^{n} \left(\frac{1}{n} Var[R_i] \right) \\ &= \frac{1}{n^2} Var[R_i] * n \\ &= \frac{Var[R_i]}{n} \\ &= \frac{(E[R_i^2] - E[R_i]^2)}{n} \\ &= \frac{1^2 * r + 0^2 * (1 - r) - r^2}{n} \\ &= \frac{r(1 - r)}{n} \end{aligned} \tag{19}$$

2.5 $Var[\hat{Q}]$

As you saw in 18 we propose an estimate for $\hat{Q}$ so according to 19 we can calculate $Var[\hat{Q}]$ as follow:(Note that p is known in population)

$$\begin{aligned} Var[\hat{Q}] &= Var[\frac{R+p-1}{2p-1}] \\ &= \frac{1}{(2p-1)^2} Var[R+p-1] \\ &= \frac{Var[R]}{(2p-1)^2} \\ &= \frac{\frac{r(1-r)}{n}}{(2p-1)^2} \\ &= \frac{r(1-r)}{n(2p-1)^2} \end{aligned} \tag{20}$$

So the $Var[\hat{Q}]$ will be $\frac{r(1-r)}{n(2p-1)^2}$.

3 Problem 2

3.1 Standard error be less than 0.01

Pursuing find sample size, defining a new random variable for each health problem would be helpful:

$$\begin{aligned} P(X_i) &= \begin{cases} p_1 & \text{if that person has first health problem} \\ 1-p_1 & \text{other wise} \end{cases} \\ P(Y_i) &= \begin{cases} p_2 & \text{if that person has second health problem} \\ 1-p_2 & \text{other wise} \end{cases} \end{aligned} \tag{21}$$

So X_i and Y_i are Bernoulli distributions with parameter p_1 and p_2 .At next step calculating expected of each in order to find these parameters:

$$\begin{aligned} E[\overline{X}_n] &= \frac{\sum_{i=1}^{n} X_i}{n} \\ &= \frac{nE[X_i]}{n} \\ &= E[X_i] \\ E[\overline{Y}_n] &= \frac{\sum_{i=1}^{n} E[Y_i]}{n} \\ &= \frac{nE[Y_i]}{n} \\ &= E[Y_i] \end{aligned} \tag{22}$$

On the other hand knowing what are these expected values are can be helpful to calculate the parameters:

$$\begin{aligned} E[\overline{X}_n] &= E[X_i] \\ &= p_1 \\ &= 0.03 \\ E[\overline{Y}_n] &= E[Y_i] \\ &= p_2 \\ &= 0.4 \end{aligned} \tag{23}$$

Lets obtain standard error and ignore the finite population correlation as question said:

$$\begin{aligned} \sigma_{first} &= \frac{\sigma_{X_i}}{\sqrt{n}} \\ &= \frac{\sqrt{p_1(1-p_1)}}{\sqrt{n}} \\ \sigma_{second} &= \frac{\sigma Y_i}{\sqrt{n}} \\ &= \frac{\sqrt{p_2(1-p_2)}}{\sqrt{n}} \end{aligned} \tag{24}$$

From the given inequality we can find minimum sample size for each:

$$\begin{aligned} \frac{\sqrt{p_1(1-p_1)}}{\sqrt{n_1}} &< 0.01 \\ \frac{p_1(1-p_1)}{n} &< 0.0001 \\ \frac{0.03 * 0.97}{0.0001} &< n_1 \\ 291 &< n \\ \frac{\sqrt{p_2(1-p_2)}}{\sqrt{n_2}} &< 0.01 \\ \frac{p_2(1-p_2)}{n_2} &< 0.0001 \\ \frac{0.4 * 0.6}{0.0001} &< n_2 \\ 2400 &< n_2 \end{aligned} \tag{25}$$

As equation 25 two values has been obtained but we must use a sample size to satisfy both inequality so choosing 2400 would be reasonable.
(Note that question said less than 0.01 not less than or equal to 0.01 so we must choose 2401 instead of 2400 but

for simplicity we chose 2401.) Lets find actual standard error with this sample size:

$$\begin{aligned}
\sigma_{first} &= \sqrt{\frac{p_1(1-p_1)}{n}} \\
&= \sqrt{\frac{0.03 * 0.97}{2400}} \\
&= \sqrt{\frac{0.0291}{2400}} \\
&= \sqrt{0.000012125} \\
&= 0.003482097 \\
\sigma_{second} &= \sqrt{\frac{p_2(1-p_2)}{n}} \\
&= \sqrt{\frac{0.4 * 0.6}{2400}} \\
&= 0.01
\end{aligned} \tag{26}$$

3.2 Standard errors less than 10% of the true value

Given information in question is as follow:

$$\begin{aligned}
\sigma_{first} &< 0.1p_1 \\
&< 0.003 \\
\sqrt{\frac{p_1(1-p_1)}{n_1}} &< 0.003 \\
\frac{p_1(1-p_1)}{n_1} &< 0.000009 \\
\frac{0.03 * 0.97}{n_1} &< 0.000009 \\
3233.333333333 &< n_1 \\
\sigma_{second} &< 0.1p_2 \\
&< 0.04 \\
\sqrt{\frac{p_2(1-p_2)}{n_2}} &< 0.04 \\
\frac{p_2(1-p_2)}{n_2} &< 0.0016 \\
\frac{0.4 * 0.6}{n_2} &< 0.0016 \\
150 &< n_2
\end{aligned} \tag{27}$$

Choosing the right sample size will cause each standard error be less than 10% of the true value for each cases.So we must choose $n = 3234$.

4 Problem 3

4.1 a

At first to have more understanding waht CLT and confidence interval says we define each of them:

- CLT: Let $X_1, \cdots, X_n$ be i.i.d[2] with mean μ and variance σ^2. Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Tehran

$$Z_n \equiv \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{\text{D}} Z \sim N(0,1)$$

- Confidence Intervals: A $100(1-\alpha)\%$ confidence interval for a population θ is a random variable calculated from the samole, which contains θ with probability $1-\alpha$.

Actually in CLT, we are proving that $\overline{X}_n$ has a distribution which is approximately Normal although in confidence interval we are going to calculate probability of a random variable which contains a population parameter like θ. So if we wrap up every thing this statement is wrong and in order to fix this:
Fix statement: In confidence interval we can use CLT theorem to find confidence interval for mean parameter like below:

$$P\left(-Z_{\frac{\alpha}{2}} \leq \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \leq Z_{\frac{\alpha}{2}}\right) = 1-\alpha \tag{28}$$

As you can see this probability is apporixmated by normal distribution with $\mu = 0, \sigma = 1$ that comes from CLT.

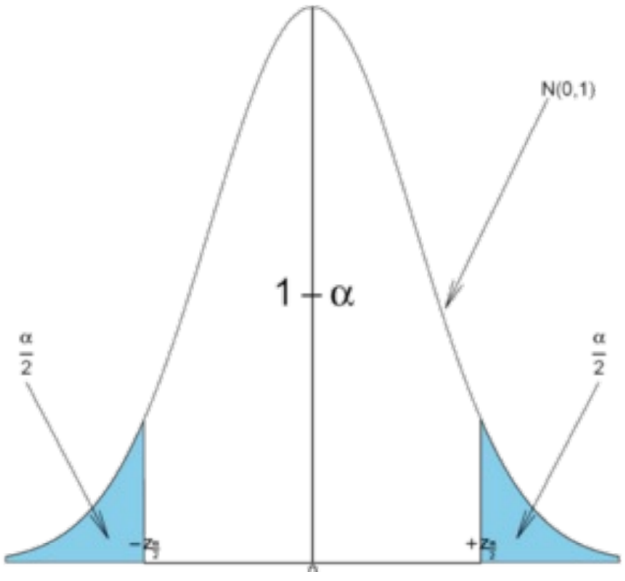

Figure 1: Confidence interval for parameter μ

[2]Independent Identical Distributions

4.2 b

This statment is true, note that it is alwyas possible that we reject 2 side tests while 1 side test dosen't reject.

4.3 c

False.
CLT theorem states that the sampling distribution of a sample mean will closely resemble the normal distribution but it depends on the sample size.
For more information as sample size getting larger the variance of the $\overline{X}_n$ will getting closer to zero and generally speaking if sample size getting biger and biger then we can state that from the law of large numbers $\overline{X}_n$ can approximated to μ so the approximation of $\overline{X}_n$ is definitely depends on sample size and 2 other factors like skewness and kurtosis of distribution X_i.

4.4 d

True. In order to see skewness examples and you can refer also picture 2. As you can see generally the statement

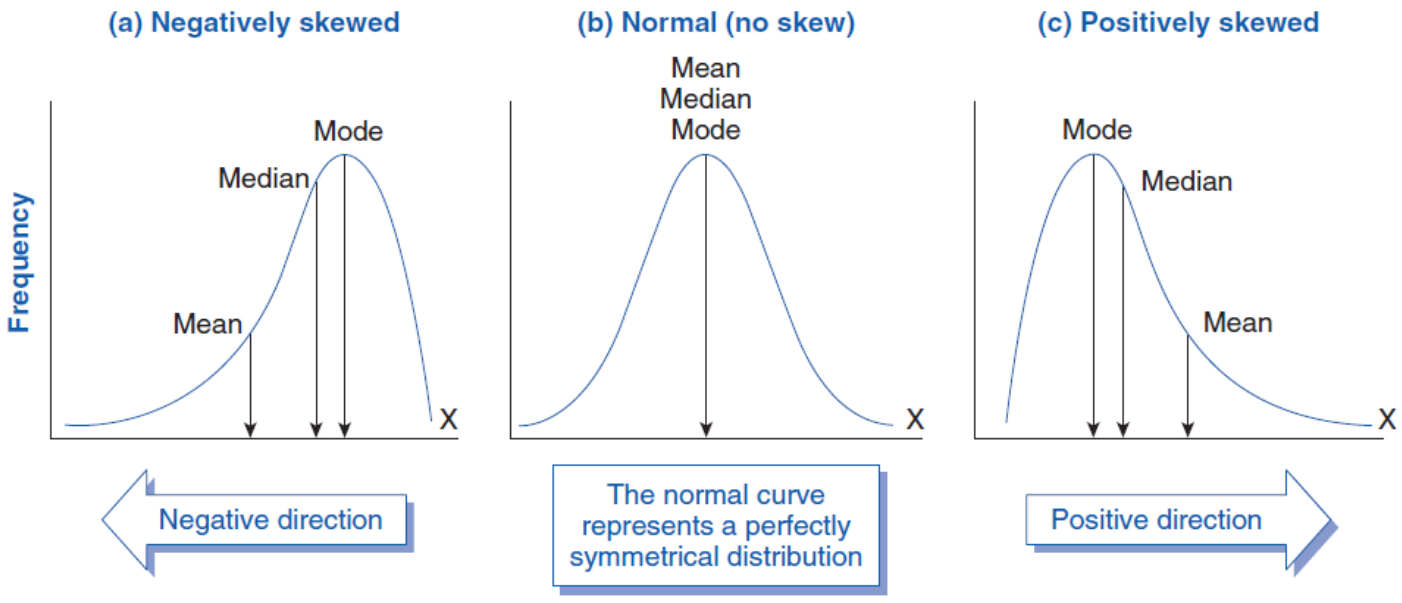

Figure 2: Skewness

is true but in some conditions the relationship between mean,median and mode and variations can be depend on the shape of distribution so usually it is true but not always.

4.5 e

For given standard error, confidence levels is denoted by $1-\alpha$ so as confidence levels get lower then α gets bigger so from 1 the more α gets bigger then confidence interval gets smaller and not wider so this statment is false and true statment is:
For a given standard error, lower confidence levels produce lower confidence intervals.

4.6 f

False, Confidence interval is for a particular sample and it is random variable due to a random sample so we can not say there is 95% probability that the population mean is between 350, 400.

4.7 g

False.
A 95% confidence interval obtained from a random sample of 1000 people has a more precise estimate of the population parameter than a 95% confidence interval obtained from a random sample of 500 people.

4.8 h

False.
When we construct confidence interval for the sample mean which is: $18.4 \leq \mu \leq 21.5$ we can not say what percentage confident and it can be any number depend on our simulation and sample size.

4.9 i

True.

5 Problem 4

For each measurements standard deviation is equal to 10. Lets define a new random variable δ as follow:

$$\delta = \overline{X} - \overline{Y} \tag{29}$$

where X, Y are group 1 and group 2 respectively, lets findout the mean and variance of δ:

$$\begin{aligned} E[\delta] &= E[\overline{X}] - E[\overline{Y}] \\ &= \mu_X - \mu_Y \\ Var[\delta] &= Var[\overline{X}] + Var[\overline{Y}] \\ &= \frac{\sigma_{\overline{X}}^2}{n} + \frac{\sigma_{\overline{Y}}^2}{n} \end{aligned} \tag{30}$$

Finally we are going to calculate the confidence interval:

$$\begin{aligned} P\left(-Z(\frac{\alpha}{2}) \leq \left(\frac{\delta - \mu_\delta}{\frac{\sigma_\delta}{\sqrt{n}}}\right) \leq Z(\frac{\alpha}{2})\right) &= 1-\alpha \\ P\left(-Z(\frac{\alpha}{2})\frac{\sigma_\delta}{\sqrt{n}} - \delta \leq -\mu_\delta \leq Z(\frac{\alpha}{2})\frac{\sigma_\delta}{\sqrt{n}} - \delta\right) &= 1-\alpha \\ P\left(-Z(\frac{\alpha}{2})\sqrt{\frac{\sigma_{\overline{X}}^2 + \sigma_{\overline{Y}}^2}{n}} - \overline{X} + \overline{Y} \leq -\mu_\delta \leq Z(\frac{\alpha}{2})\sqrt{\frac{\sigma_{\overline{X}}^2 + \sigma_{\overline{Y}}^2}{n}} - \overline{X} + \overline{Y}\right) &= 1-\alpha \\ P\left(-Z(\frac{\alpha}{2})\sqrt{\frac{\sigma_{\overline{X}}^2 + \sigma_{\overline{Y}}^2}{n}} + \overline{X} - \overline{Y} \leq \mu_\delta \leq Z(\frac{\alpha}{2})\sqrt{\frac{\sigma_{\overline{X}}^2 + \sigma_{\overline{Y}}^2}{n}} + \overline{X} - \overline{Y}\right) &= 1-\alpha \end{aligned} \tag{31}$$

So from 95% we shall say:

$$\begin{aligned} 1.96\sqrt{\frac{\sigma_{\overline{X}}^2 + \sigma_{\overline{Y}}^2}{n}} + \overline{X} - \overline{Y} &\leq 2 \\ 1.96\sqrt{\frac{200}{n}} + \overline{X} - \overline{Y} &\leq 2 \\ 1.96\sqrt{\frac{200}{n}} &\leq 2 - \overline{X} + \overline{Y} \\ \sqrt{\frac{200}{n}} &\leq \frac{2 - \overline{X} + \overline{Y}}{1.96} \\ \frac{200}{n} &\leq \frac{\left(2 - \overline{X} + \overline{Y}\right)^2}{3.8416} \\ \frac{768.32}{\left(2 - \overline{X} + \overline{Y}\right)^2} &\leq n \end{aligned} \tag{32}$$

Another interpretation is difference between min and max of CI is 2 so we can say that:

$$Z(\frac{\alpha}{2})\sqrt{\frac{\sigma_{\overline{X}}^2+\sigma_{\overline{Y}}^2}{n}}+\overline{X}-\overline{Y}+Z(\frac{\alpha}{2})\sqrt{\frac{\sigma_{\overline{X}}^2+\sigma_{\overline{Y}}^2}{n}}-\overline{X}+\overline{Y}\leq 2 \tag{33}$$

Where $Z(\frac{\alpha}{2})$ for 95% CI is 1.96 so:

$$\begin{aligned} 2*1.96\sqrt{\frac{200}{n}} &\leq 2 \\ \sqrt{\frac{200}{n}} &\leq \frac{1}{1.96} \\ \frac{200}{n} &\leq \left(\frac{1}{1.96}\right)^2 \\ 768.32 &\leq n \end{aligned} \tag{34}$$

which n must be 769 in each group to CI for 95% confidence interval for the difference in the mean outcomes between the two groups is less than or equal to 2 units.

6 Problem 5

As we know in two-sample t test, we have t-distribution as follow:

$$\frac{\overline{x}_1 - \overline{x}_2 - \delta}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2} \tag{35}$$

where

$$s_p^2 = \frac{(n_1-1)\,s_1^2 + (n_2-1)\,s_2^2}{n_1+n_2-2} \tag{36}$$

So we can obtain confidence interval for δ from t-distribution:

$$P\left(-t_{\frac{\alpha}{2},n_1+n_2-2} \leq \frac{\overline{x}_1 - \overline{x}_2 - \delta}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \leq t_{\frac{\alpha}{2},n_1+n_2-2}\right) = 1-\alpha$$

$$P\left(-t_{\frac{\alpha}{2},n_1+n_2-2} * s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \leq \overline{x}_1 - \overline{x}_2 - \delta \leq t_{\frac{\alpha}{2},n_1+n_2-2} * s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right) = 1-\alpha$$

$$P\left(-t_{\frac{\alpha}{2},n_1+n_2-2} * s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} + \overline{x}_1 - \overline{x}_2 \leq \delta \leq t_{\frac{\alpha}{2},n_1+n_2-2} * s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} + \overline{x}_1 - \overline{x}_2\right) \tag{37}$$

So the confidence interval for δ or $\mu_{x_1} - \mu_{x_2}$ is:

$$(\overline{x}_1 - \overline{x}_2) \pm t_{\frac{\alpha}{2},n_1+n_2-2} * s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \tag{38}$$

Now lets see what happen when confidence interval for $\mu_{x_1} - \mu_{x_2}$ does not include zero.Two different condition can be assumed to not having zero in our confidence interval:

- $(\overline{x}_1 - \overline{x}_2) + t_{\frac{\alpha}{2},n_1+n_2-2} * s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} < 0$
- $(\overline{x}_1 - \overline{x}_2) - t_{\frac{\alpha}{2},n_1+n_2-2} * s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} > 0$

At first condition if $(\overline{x}_1 - \overline{x}_2) + t_{\frac{\alpha}{2},n_1+n_2-2} * s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ be less than 0 then we know that maximum of $\mu_{x_1} - \mu_{x_2}$ is less than zero, more precisely in mathematical form:

$$\begin{aligned}
&\mu_{x_1} - \mu_{x_2} \leq (\overline{x}_1 - \overline{x}_2) + t_{\frac{\alpha}{2},n_1+n_2-2} * s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} < 0 \\
&\mu_{x_1} - \mu_{x_2} < 0 \\
&\mu_{x_1} < \mu_{x_2}
\end{aligned} \tag{39}$$

So if first condition will be appeared, it would be obvious that from equation 39 we can say $\mu_{x_1} < \mu_{x_2}$ and it would reject the null hypothesis. At second condition if $(\overline{x}_1 - \overline{x}_2) - t_{\frac{\alpha}{2},n_1+n_2-2} * s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ be bigger than 0

then the minimum of $\mu_{x_1} - \mu_{x_2}$ is bigger than zero, more precisely in mathematical form:

$$\begin{aligned} 0 < (\overline{x}_1 - \overline{x}_2) + t_{\frac{\alpha}{2}, n_1+n_2-2} * s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_{x_1} - \mu_{x_2} \\ 0 < \mu_{x_1} - \mu_{x_2} \\ \mu_{x_2} < \mu_{x_1} \end{aligned} \tag{40}$$

So if second condition will be appeared, it would be obvious that from equation 40 we can say $\mu_{x_2} < \mu_{x_1}$ and it would reject the null hypothesis.
From wrapping all things up, if our confidence interval does not include zero then null hypothesis H_0 will be rejected.

7 Problem 6

As we know the estimation for parameter $\hat{\theta}$ is unbiased when:

$$E[\hat{\theta}] = \theta \tag{41}$$

So if $\overline{X}_c$ is unbiased when:

$$\begin{aligned} E[\overline{X}_c] &= \mu \\ E\left[\sum_{i=1}^{n} c_i X_i\right] &= \mu \\ \sum_{i=1}^{n} c_i E[X_i] &= \mu \end{aligned} \tag{42}$$

where each sample will represent the whole population and so $E[X_i] = \mu$ then we can use this equation in order to solve the question above:

$$\begin{aligned} \sum_{i=1}^{n} c_i E[X_i] &= \mu \\ \mu \sum_{i=1}^{n} c_i &= \mu \\ \sum_{i=1}^{n} c_i &= 1 \end{aligned} \tag{43}$$

So if the sum of c_i's will be equal to one the $\overline{X}_c$ is an unbiased estimator of the population mean.

8 Problem 7

8.1 make a histogram

At first we are going to make a histogram for population and cancer mortalitiy and we make another histogram based on ratio of the mortalitiy in each population

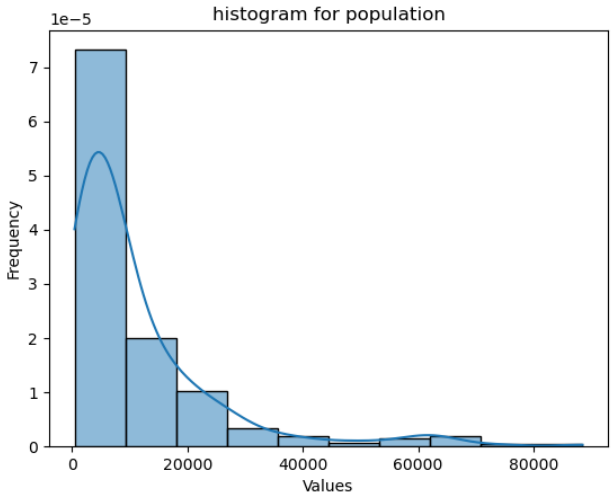

Figure 3: population histogram

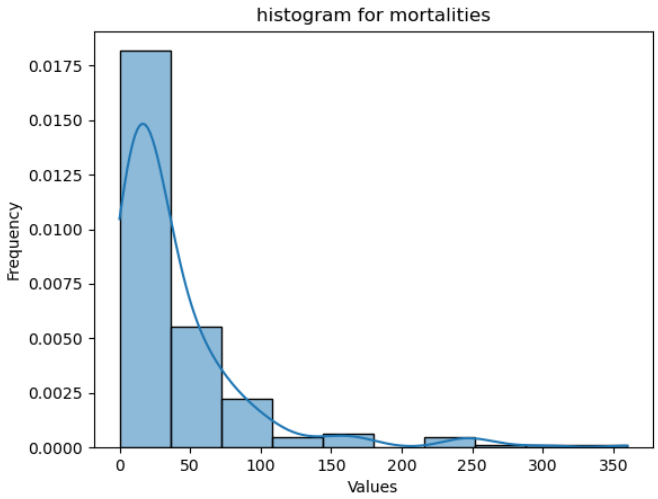

Figure 4: cancer mortalitiy histogram

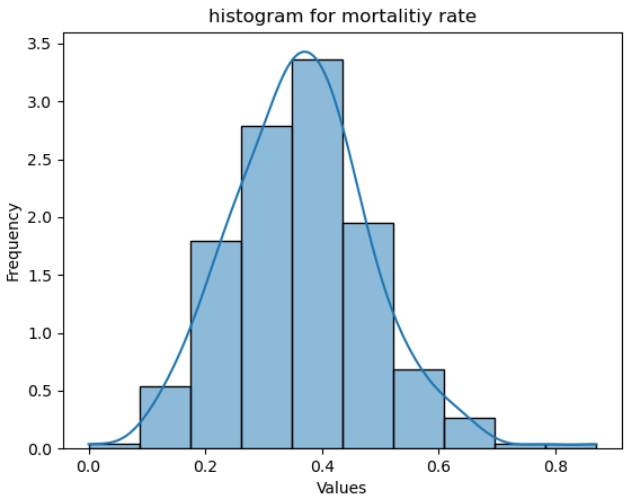

Figure 5: ratio histogram

8.2 mean, variance and std population and total cancer mortalitiy

We form a table below in order to see the parameters of the population and mortalitiy which are in 1 and 2 tables respectively:

	mean	variance	standard deviation
population	11288.056478405315	189888678.0334662	13780.010088293338

Table 1: parameters for population

	mean	variance	standard deviation	total mortalitiy
cancer	39.857142857142854	2598.7361904761915	50.97780095763441	11997

Table 2: paramters for cancer

8.3 simulate sampling for 25 observations

In order to simulate this we consider 1000 simulations while in each simulate we get 25 sample and calculate mean of each and make a histogram for 1000 means (note that from CLT we must expect that histogram would be like normal distibution):

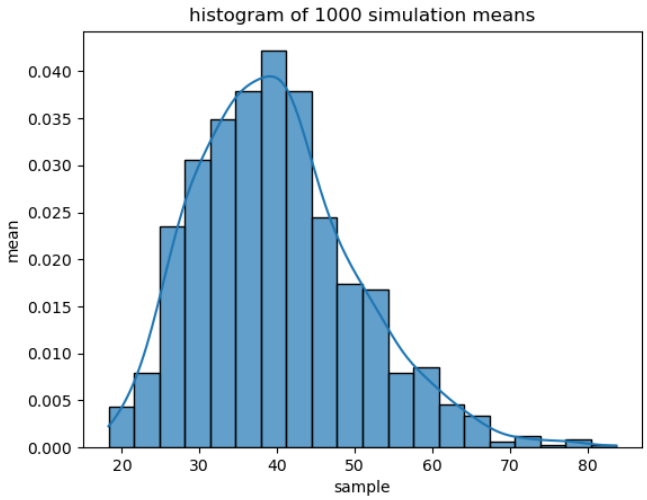

Figure 6: histogram of 1000 means for each sample

As you can see in figure 6 we expected a normal distribution and also we can see the mean of this histogram approximately is mean of cancer mortality as expected from CLT.

8.4 estimate mean and total mortalitiy from 25 sample

As we know the sample mean and the variance mean is:

$$\begin{aligned} \overline{Y}_n &= \frac{1}{n}\sum_{i=1}^{n} Y_i \\ S_n^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \overline{Y}_n\right)^2 \end{aligned} \tag{44}$$

And it was discussed that these mean and varinace of sample are unbiased estimators of μ and σ^2 respectively,So we can use these facts to estimate the mean and total cancer mortalitiy as follows:(for a 25 sample size)

$$\begin{aligned} \overline{Y}_n &= \frac{1}{25}\sum_{i=1}^{25} Y_i \\ Total &= N * \overline{Y}_n \\ &= 301 * \overline{Y}_n \end{aligned} \tag{45}$$

So let's get 25 random sample and calculate it's mean and multiply it by $N = 301$ in order to estimate total mortalitiy cancer. $36.4, 10956.4$ are the estimated mean and total cancer mortalitiy respectively.
Attention: If you run multiple times you will get different numbers due to sample is random.

8.5 estimate population variance and std from the sample

As we saw in equation 44 we can use this to estimate population variance as follow:

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 \tag{46}$$

Notation: X is for population and Y is for the cancer mortalitiy.
We calculate the variance of the sample and multiply to the following coefficient to be unbiased:

$$\begin{aligned} S_n^2 &= \frac{n}{n-1} * variance(sample) \\ S_n = std &= \sqrt{\frac{n}{n-1} * variance(sample)} \end{aligned} \tag{47}$$

The estimated variance of population and standard deviation is:

$$(\text{estimated variance}, \text{estimated std}) = (115957317.88194442, 10768.347964378956)$$

8.6 Form 95% confidence interval for sample in d

We will construct three confidence interval for mean of cancer mortalitiy, mean of population and total cancer mortalitiy as follow:(note that our sample size is less than 30 so we consider t-distibution for confidence interval)

$$\begin{aligned} &\left(\text{mean of cancer} - t_{\alpha/2,\ 24} * \frac{\text{std of sample}}{\sqrt{25}}, \text{mean of cancer} + t_{\alpha/2,\ 24} * \frac{\text{std of sample}}{\sqrt{25}}\right) \\ &\left(\text{mean of population} - t_{\alpha/2,\ 24} * \frac{\text{std of sample}}{\sqrt{25}}, \text{mean of population} + t_{\alpha/2,\ 24} * \frac{\text{std of sample}}{\sqrt{25}}\right) \\ &\left(\text{total cancer} - t_{\alpha/2,\ 24} * \frac{\text{std of sample}}{\sqrt{25}} * 301, \text{total cancer} + t_{\alpha/2,\ 24} * \frac{\text{std of sample}}{\sqrt{25}} * 301\right) \end{aligned} \tag{48}$$

Just careful about std of sample in each CI cause we get 25 samples but it has 2 differnt values for sample one for cancer and another for population in counties. and std for each CI is coresponding sample for that data.

$$\begin{aligned} &(5987.844425041649, 14877.75557495835) \\ &(20.215421810307294, 52.58457818969271) \\ &(6084.841964902495, 15827.958035097505) \end{aligned} \tag{49}$$

And as you see in the part b we calculate each parameters so all of them have population parameters.

8.7 repeat previous parts for 100 sample size

The difference is just using Noraml distibution for constructing the confidence intervals.
Esimated mean cancer for a sample : 35.45
Estimated total cancer for a sample : 10670.45
Estimated population variance is: 145921671.2191614
Estimated population standard deviation is : 12079.804270730607

CI for mean of population for sample: $(7988.9483629368015, 7988.9483629368015)$
CI for mean of cancer for sample : $(27.31274730763195, 43.587252692368054)$
CI for total cancer for sample : $(8221.136939597216, 13119.763060402785)$
As you can see CI for each one has population parameters.

8.8 Effectiveness of ratio estimator

Yes indeed we can use ratio estimator cause to these resons:

- If there is strong positive correlation between 2 variables then ratio estimator can be effective.
- Ratio estimator can reduce the variance of the estimate and if variance of the our estimator reduced so it will be effective.

Now lets see the correlation between these two varibales:

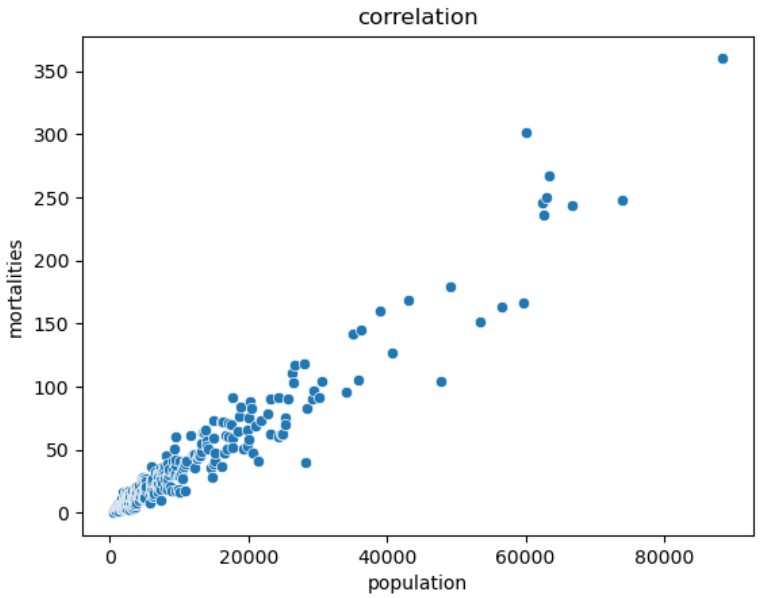

Figure 7: correlation with scatter plot

As we can see in figure 7 our variables are correlated so ratio estimator can be indeed effective.

8.9 simulation for ratio estimator

Before code this we must assure that we use good ratio estimator, our ratio of interest is :

$$r = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} \tag{50}$$

And the estimator ratio would be for a sample consisting pairs (X_i, Y_i) the estimator for r is:

$$\begin{aligned} R &= \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i} \\ &= \frac{\overline{Y}_n}{\overline{X}_n} \end{aligned} \tag{51}$$

so the ratio estimator of μ_y would be:

$$\begin{aligned} \overline{Y}_R &= \mu_x \frac{\overline{Y}}{\overline{X}} \\ &= \mu_x R \end{aligned} \tag{52}$$

Now let's simulate 1000 times getting samples of size 25 and see the difference between ratio estimator and part c estimator.

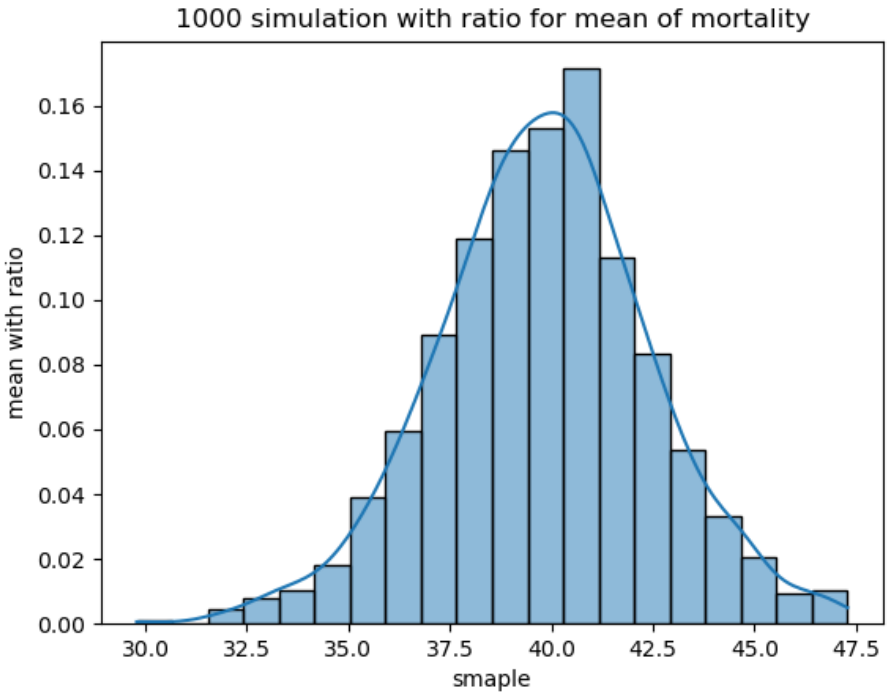

Figure 8: histogram for mean of cancer for each sample with ratio

If we differentiate between figure 8 and 6 we can see the ratio estimator has lower variation to simple random sampling in part c.(As we expected we said that ratio estimator can reduce variance of estimator in previous part and it is indeed effective)

8.10 Draw a sample ...

For total we should use following equation:

$$\begin{aligned} T &= N\frac{\overline{Y}}{\overline{X}}\mu_{population} \\ &= NR\mu_{population} \end{aligned} \tag{53}$$

Now lets differentiate between ratio estimator and part d:
Esimated mean cancer for a sample : 41.68
Estimated total cancer for a sample : 12545.68
Esimated mean ratio cancer for sampe sample : 37.610973135139155
Estimated total cancer ratio for sampe sample : 11320.902913676886
As you can see ratio estimator has closer value to our population parameters and it will be concluded that ratio estimator is better.

8.11 Form confidence intervals

For construction confidence intervals we should know how to construct confidence interval for ratio estimator. Suppose that our estimate ratio is $R = \frac{\overline{Y}}{\overline{X}}$ so we have:

$$\begin{aligned} S_x^2 &= \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)}{n-1} \\ S_y^2 &= \frac{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)}{n-1} \end{aligned} \tag{54}$$

now we should calculate the S_R from page 233 from John's Rice book we shall say:

$$S_R^2 = \frac{1}{n}\frac{1}{\overline{X}^2}\left(R^2 S_x^2 + S_y^2 - 2RS_{xy}\right) \tag{55}$$

where S_{xy} is:

$$S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) \tag{56}$$

cause we have 25 sample then it should be t-distibution so we shall say:
confidence interval for r is $R \pm t_{\frac{\alpha}{2},n-1}S_R$ and confidence interval for μ_y would be $R * \mu_x \pm t_{\frac{\alpha}{2},n-1}S_R\mu_x$
CI for mean cancer using ratio estimator: $(33.71008561929527, 47.30026914335908)$
CI for total cancer mortality using ratio estimator: $(10146.735771407877, 14237.381012151083)$
CI for mean population using ratio estimator: $(9244.088949699955, 12970.833128222122)$
This will give us less variation than simple ranodom sampling cause ratio estimator hass lower variance so it is a better estimation and it is effective.

8.12 stratify counties into four strata

At first we are going show the properties of stratified estimates. Suppose that we have L strata in all.Total population size is N where each strata has N_l population size which means:

$$N = N_1 + N_2 + \cdots + N_L \tag{57}$$

Assume each strata has mean and variance of μ_l and σ_l and let x_{il} denote ith population value in the lth stratum and $W_l = N_l/N$ as fraction of the population:

$$\mu = \frac{1}{N}\sum_{l=1}^{L}\sum_{i=1}^{N_l} x_{il} \tag{58}$$

The stratified estimate $\overline{X}_s$ of the population mean is unbiased means:

$$E[\overline{X}_s] = \sum_{l=1}^{L}\frac{N_l}{N}E[\overline{X}_l] = \mu \tag{59}$$

and variance of $\overline{X}_s$ is approximately (not considering finite correction):

$$Var\left(\overline{X}_s\right) = \sum_{l=1}^{L}\frac{W_l^2\sigma_l^2}{n_l} \tag{60}$$

At first we are going split our data in 4 strata by their percentiles and see population parameters in each: Now lets

l	N_l	$\mu_{population}$	μ_{cancer}	$\sigma_{population}$	σ_{cancer}
1	76	1956.8947	8.0526	656.0088	4.1373
2	75	4453.7733	15.68	998.0886	6.8263
3	75	9312.1066	33.9066	2111.0324	33.9066
4	75	29553.8666	102.2133	16883.7824	68.5906

get 6 sample from each strata and calculate the mean of them and eventually find $\overline{Y}_s$ from Y_l's:

$$\begin{aligned}\overline{Y}_s &= \sum_{l=1}^{4} W_l\overline{Y}_l \\ &= 37.31893687707641\end{aligned} \tag{61}$$

$$\begin{aligned}\overline{X}_s &= \sum_{l=1}^{4} W_l\overline{X}_l \\ &= 9734.288482834994\end{aligned} \tag{62}$$

As you can see the mean of cancer mortalitiy is 37.31893687707641 and mean of population is 9734.288482834994 so the total cancer mortalitiy is

$$\begin{aligned}total &= N\overline{Y}_s \\ &= 11233.0\end{aligned} \tag{63}$$

For more work we can make simulation and see the histogram of the means of cancer for each sample in figure 9. As we expected histogram of figure 9 probability of real mean of cancer has the highest value cuase to CLT.

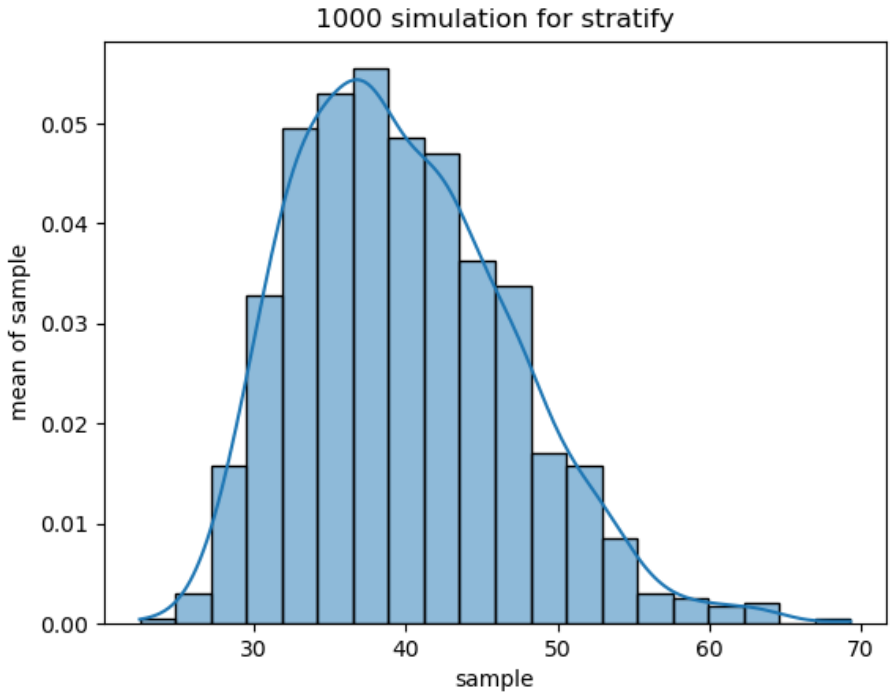

Figure 9: Simulation for mean of cancer for stratify

8.13 Methods of allocation

For optimal allocation we must minimize $Var(\overline{X}_s)$, in order to minimize this we must find n_l for each strata by the following equation:

$$n_l = n\frac{W_l\sigma_l}{\sum_{k=1}^{L} W_k\sigma_k} \tag{64}$$

Now lets calculate them with code and write them:

$$\begin{aligned} n_1 &= 1 \\ n_2 &= 2 \\ n_3 &= 3 \\ n_4 &= 18 \end{aligned} \tag{65}$$

From each strata we find the sample we should get.
For proportional allocation is just derive from previous part which from every strata we gets 6 samples.
In order to compare the variances of the population mean(hear we consider cancer mean not real population mean!) we should use theorem C page 235 of Rice's book which is:

$$Var(\overline{Y}_{sp}) - Var[\overline{Y}_{so}] = \frac{1}{n}\sum_{l=1}^{L} W_l\left(\sigma_l - \hat{\sigma}\right)^2 \tag{66}$$

where:

$$\hat{\sigma} = \sum_{l=1}^{L} W_l \sigma_l \tag{67}$$

We obtain $\hat{\sigma}$ from code which is:

$$\hat{\sigma} = 22.975801577250905 \tag{68}$$

Now lets calculate the $Var[\overline{Y}_{sp}] - Var[\overline{Y}_{so}]$:

$$\begin{aligned} Var[\overline{Y}_{sp}] - Var[\overline{Y}_{so}] &= \frac{1}{n} \sum_{l=1}^{L} W_l \left(\sigma_l - 22.975801577250905\right)^2 \\ &= \frac{1}{24} \sum_{l=1}^{4} W_l (\sigma_l - 22.975801577250905) \\ &= 29.16114051366075 \end{aligned} \tag{69}$$

For differentiate between variance of the mean of a simple random sampling and the variance of the mean of a stratified random sample based on a proportional allocation is:(theorem D in page 237 Rice's book)

$$\begin{aligned} Var[\overline{Y}] - Var[\overline{Y}_{sp}] &= \frac{1}{n} \sum_{l=1}^{L} W_l (\mu_l - \mu)^2 \\ &= \frac{1}{24} \sum_{l=1}^{4} W_l (\mu_l - 39.857142857142854)^2 \\ &= 57.4465412978737 \end{aligned} \tag{70}$$

So as we expected optimal allocation has lower variance so it should be better compare to proportional allocation and respectively proportional allocation has better variance than simple random sampling so its more effective. If i want to compare the simple random sampling to optimal allocation we should sum two equations 69 and 70 we can calculate what's was wanted:

$$\begin{aligned} (Var[\overline{Y}_{sp}] - Var[\overline{Y}_{so}]) + (Var[\overline{Y}] - Var[\overline{Y}_{sp}]) &= 29.16114051366075 + 57.4465412978737 \\ Var[\overline{Y}] - Var[\overline{Y}_{so}] &= 86.60768181153446 \end{aligned} \tag{71}$$

As we can see optimal allocation is more effective than simple random sampling.

8.14 stratify into $8, 16, 32, 64$

We split data by percentiles into given stratas and do the pervious part procedure to see the differences between optimal allocation and simple random sampling and proportional allocation

8.14.1 strata = 8

$$\begin{aligned} Var[\overline{Y}_{sp}] - Var[\overline{Y}_{so}] &= 22.10892648783374 \\ Var[\overline{Y}] - Var[\overline{Y}_{sp}] &= 75.44654328534814 \\ Var[\overline{Y}] - Var[\overline{Y}_{so}] &= 97.55546977318188 \end{aligned} \tag{72}$$

8.14.2 strata = 16

$$\begin{aligned} Var[\overline{Y}_{sp}] - Var[\overline{Y}_{so}] &= 10.52823186551177 \\ Var[\overline{Y}] - Var[\overline{Y}_{sp}] &= 92.30090817436724 \\ Var[\overline{Y}] - Var[\overline{Y}_{so}] &= 102.82914003987901 \end{aligned} \tag{73}$$

8.14.3 strata = 32

$$\begin{aligned} Var[\overline{Y}_{sp}] - Var[\overline{Y}_{so}] &= 4.639634275678016 \\ Var[\overline{Y}] - Var[\overline{Y}_{sp}] &= 100.00614706330032 \\ Var[\overline{Y}] - Var[\overline{Y}_{so}] &= 104.64578133897834 \end{aligned} \tag{74}$$

8.14.4 strata = 64

$$\begin{aligned} Var[\overline{Y}_{sp}] - Var[\overline{Y}_{so}] &= 4.651401689568797 \\ Var[\overline{Y}] - Var[\overline{Y}_{sp}] &= 101.30249960449295 \\ Var[\overline{Y}] - Var[\overline{Y}_{so}] &= 105.95390129406177 \end{aligned} \tag{75}$$

From this stratas we con conclude that as strata gets bigger optimal allocation and proportional allocation gets closer to each other but still optimal allocation is better. But as strata gets bigger there will be significant distance between stratify mehtods and and simple random sampling.

9 Problem 8

9.1 What given code does?

This function is getting 2 sample of size n and between $[0, 1)$ and see if the following equation was true then that particular point is in that area (which is circle).

$$X^2 + Y^2 < 1 \tag{76}$$

If we wrap this up like we have 2 uniform random variable $\sim U(0, 1)$ and see if points would be in circle with radious 1 or not.Then total points within the circle divided by whole points will give us estimate of $\frac{\pi}{4}$ cause it is a quarter of a circle with radious 1 so for estimate π we must multiply it by 4. This method called Monte Carlo in order to calculate the area of circle with radious 1.
Assume that we want to calculate the integral bellow:

$$\int_{x_0}^{x_1} \int_{y_0}^{y_1} f(x, y)dxdy \tag{77}$$

From monte carlo we can say that we have Uniform 2D random variable as follow:

$$X_i = p(x, y) = \frac{1}{(x_1 - x_0)} \frac{1}{(y_1 - y_0)} \tag{78}$$

So from 2D estimator with monte carlo we have:

$$F_n = \frac{(x_1 - x_0)(y_1 - y_0)}{n} \sum_{i=1}^{n} f(X_i) \tag{79}$$

As n grows this F_n would be closer to answer of this integral.
In here we calculate the area of circle with monte carlo method and as n grows The estimation would be closer to area of circle with radious 1.Which the answer for given n's are as follow:

$$\text{estimated area} = \begin{cases} 2.8 & n = 10 \\ 3.13924 & n = 100000 \\ 3.141484 & n = 10000000 \end{cases} \tag{80}$$

As we can see estimated area is much closer to real area of circle with radious of 1 as n grows.

9.2 Estimate ellipse area (bonus)

We use monte carlo in order to calculate the ellipse area, we know that b is semi-minor axis and a is semi-major axis and the area ellipse with a and b is:

$$\text{ellipse area} = \pi ab \tag{81}$$

So in order to estimate ellipse area using monte carlo we define two random variable and get samples with $runif(n, 0, a)$, $runif(n, 0, b)$ and see if the following function is true:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} <= 1 \tag{82}$$

If points satisfy ineqality above then this points are in the ellipse area,note that we are calculating quarter of ellipse so we must multiply it by 4 as follow:

$$(4 * (sum((x^2/a^2 + y^2/b^2) <= 1)/n) * a * b) \tag{83}$$

we must multiply it by a and b cuase we must calculate ellipse for the given n's and $a = 2, b = 1$:

$$\text{real ellipse area is : } 6.28319 \tag{84}$$

$$\text{estimated area for ellipse} = \begin{cases} 4.8 & n = 10 \\ 6.28488 & n = 100000 \\ 6.282842 & n = 10000000 \end{cases} \tag{85}$$

10 Bonus Problem

At first we are going to calculate probability when $X \leq$ T:

$$\begin{aligned} P\left(X \leq \mathrm{T}\right) &= \int_0^T \frac{1}{\alpha} e^{-x/\alpha} dx \\ &= 1 - e^{-T/\alpha} \end{aligned} \tag{86}$$

So the probability of X be bigger than T would be:

$$\begin{aligned} P\left(X > \mathrm{T}\right) &= 1 - P\left(X \leq T\right) \\ &= 1 - \left(1 - e^{-T/\alpha}\right) \\ &= e^{-T/\alpha} \end{aligned} \tag{87}$$

Now let's obtain mean of Z:

$$\begin{aligned} E[Z] &= \frac{1}{\alpha} \int_0^T x e^{-x/\alpha} dx + \frac{1}{\alpha} \int_T^\infty T e^{-x/\alpha} dx \\ &= \frac{1}{\alpha} \left[-\alpha x e^{-x/\alpha} \Big|_0^T + \alpha \int_0^T e^{-x/\alpha} dx \right] - \alpha \frac{T}{\alpha} e^{-x/\alpha} \Big|_T^\infty \\ &= \frac{1}{\alpha} \left[-\alpha T e^{-T/\alpha} - \alpha^2 \, e^{-x/\alpha} \Big|_0^T \right] + T e^{-T/\alpha} \\ &= \frac{1}{\alpha} \left[-\alpha T e^{-T/\alpha} + \alpha^2 - \alpha^2 e^{-T/\alpha} \right] + T e^{-T/\alpha} \\ &= \alpha - \alpha e^{-T/\alpha} \end{aligned} \tag{88}$$

Before caluclating variance of Z we should calculate $E[Z^2]$:

$$\begin{aligned} E[Z^2] &= \frac{1}{\alpha} \int_0^T x^2 e^{-x/\alpha} dx + \frac{1}{\alpha} \int_T^\infty T^2 e^{-x/\alpha} dx \\ &= \frac{1}{\alpha} \left[-\alpha x^2 e^{-x/\alpha} \Big|_0^T + 2\alpha \int_0^T x e^{-x/\alpha} dx \right] + \frac{1}{\alpha} \alpha T^2 e^{-T/\alpha} \\ &= \frac{1}{\alpha} \left[-\alpha x^2 e^{-x/\alpha} \Big|_0^T - \, 2\alpha^2 x e^{-x/\alpha} \Big|_0^T + 2\alpha^2 \int_0^T e^{-x/\alpha} dx \right] + T^2 e^{-T/\alpha} \\ &= \frac{1}{\alpha} \left[-\alpha T^2 e^{-T/\alpha} - 2\alpha^2 T e^{-T/\alpha} + 2\alpha^3 (1 - e^{-T/\alpha}) \right] + T^2 e^{-T/\alpha} \\ &= -2\alpha T e^{-T/\alpha} + 2\alpha^2 (1 - e^{-T/\alpha}) \end{aligned} \tag{89}$$

Now from equations 88 and 89 we can obtain $Var[Z]$:

$$\begin{aligned} Var[Z] &= E[Z^2] - E[Z]^2 \\ &= -2\alpha T e^{-T/\alpha} + 2\alpha^2 (1 - e^{-T/\alpha}) - \alpha^2 + 2\alpha^2 e^{-T/\alpha} - \alpha^2 e^{-2T/\alpha} \\ &= \alpha^2 - 2\alpha T e^{-T/\alpha} - \alpha^2 e^{-2T/\alpha} \end{aligned} \tag{90}$$