Assignment 1

Statistical analyses of cancer data set

Name: Devosmita Chatterjee
Email Id: chatterjeedevosmita267@gmail.com

Statistical analyses of cancer data set

1 Introduction

The cancer data set consists of values for breast cancer mortality from 1950 to 1960 (BreastCancer_Mortality) and the adult white female population in 1960 for 301 counties in North Carolina, South Carolina, and Georgia (AdultWhiteFemale_Population).

2 Question 65a.

We make a histogram of the population values for cancer mortality which is shown in figure 1.

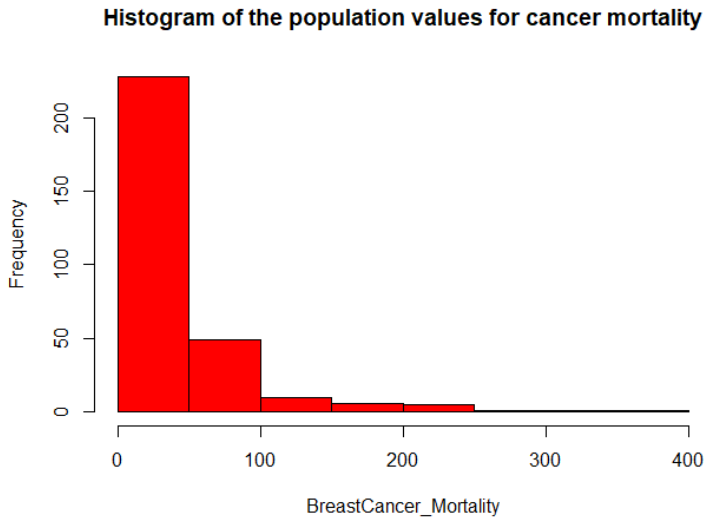

Figure 1: This figure shows the population values for cancer mortality.

3 Question 65b.

The population mean is given by

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1}$$

where x_i is the numerical value of the i^{th} member of the population and N is the size of the population. Therefore, the population mean is 39.85714.

The population total is given by

$$\tau = \sum_{i=1}^{N} x_i = N\mu. \tag{2}$$

Therefore, the total cancer mortality is 11997.

The population variance is given by

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2. \tag{3}$$

Therefore, the population variance is 2590.103.

The population standard deviation is given by

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}. \tag{4}$$

Therefore, the standard deviation is 50.89305.

4 Question 65c.

We simulate the sampling distribution of the mean of a sample of 25 observations of cancer mortality which is shown in figure 2.

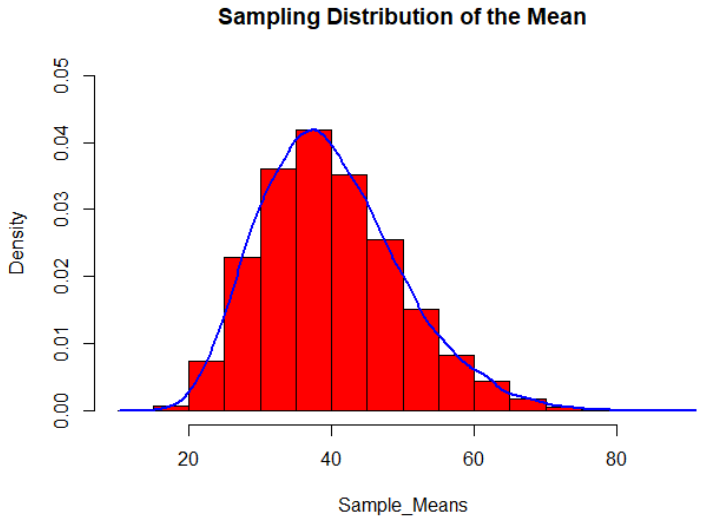

Figure 2: This figure shows the sampling distribution of the mean of a sample of 25 observations of cancer mortality.

5 Question 65d.

We draw a simple random sample of size 25 and use it to estimate the mean and total cancer mortality. The sample mean is given by

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{5}$$

where X_i is the numerical value of the i^{th} member of the sample and n is the size of the sample. Therefore, the mean is 48.64.

An estimate of the population total is given by

$$T = N\overline{X}. \tag{6}$$

Therefore, the total cancer mortality is 14640.64.

6 Question 65e.

The sample variance is given by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2. \tag{7}$$

Therefore, the population variance is 3090.157.

The sample standard deviation is given by

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}. \tag{8}$$

Therefore, the standard deviation is 55.58918.

7 Question 65f.

We form 95% confidence intervals for the population mean and total from the sample. Approximate 100(1-α)% two sided confidence intervals for the population mean μ are given by

$$I_\mu = \overline{X} \pm z_{\alpha/2} s_{\overline{X}} \tag{9}$$

where z_α denotes the normal quantile and $s_{\overline{X}} = \frac{s}{\sqrt{n}}\sqrt{1 - \frac{n}{N}}$. Therefore, 95% confidence intervals for the population mean are (27.77398, 69.50602).

Approximate 100(1-α)% two sided confidence intervals for the population total τ are given by

$$I_\tau = N\overline{X} \pm z_{\alpha/2} N s_{\overline{X}}. \tag{10}$$

Therefore, 95% confidence intervals for the population total are (8359.968, 20921.312).

The intervals cover the population values when

$$\mu \in I_\mu, \tag{11}$$

$$\tau \in I_\tau. \tag{12}$$

Yes, the intervals cover the population values since 39.85714 $\in$ (27.77398, 69.50602) and 11997 $\in$ (8359.968, 20921.312).

8 Question 65g.

We draw a simple random sample of size 100.

The mean is 41.58

The total cancer mortality is 12515.58.

The population variance is 3278.125.

The standard deviation is 57.25491.

95% confidence intervals for the population mean are (32.40987, 50.75013).

95% confidence intervals for the total population are (9755.37, 15275.79).

Yes, the intervals cover the population values since 39.85714 $\in$ (32.40987, 50.75013) and 11997 $\in$ (9755.37, 15275.79)

9 Question 65l.

We stratify the counties into four strata by population size and randomly sample six observations from each stratum shown in table 1.

Table 1: This table presents the random six observations for each stratum j, j=1,$\cdots$,4.

Stratum j	**Obs.1**	**Obs.2**	**Obs.3**	**Obs.4**	**Obs.5**	**Obs.6**
Stratum 1	4	4	11	9	3	8
Stratum 2	17	14	28	11	11	10
Stratum 3	29	45	37	34	17	49
Stratum 4	75	105	246	71	103	48

The stratum proportion for each stratum j, j=1,$\cdots$,k is given by

$$\omega_j = \frac{N_j}{N}, j = 1, \cdots, k \tag{13}$$

such that $\omega_1 + \cdots + \omega_k = 1$.

An estimate of population mean is given by

$$\overline{X}_s = \omega_1 \overline{X}_1 + \cdots + \omega_k \overline{X}_k \tag{14}$$

where $\overline{X}_j$ is the estimate of sample mean of sample size n_j for each stratum j, $j = 1, \cdots, k$. Therefore, estimate of population mean is 41.09302.

An estimate of population total is given by

$$\overline{T}_s = N\overline{X}_s \tag{15}$$

Therefore, estimate of total mortality is 12369.

10 Question 65m.

We assume total sample size n=128.

The optimal allocation is given by

$$n_j = \frac{n\omega_j\sigma_j}{\overline{\sigma}}, j = 1, \cdots, k \tag{16}$$

where σ_j is the sample standard deviation of sample size n_j for each stratum j, $j = 1, \cdots, k$ and $\overline{\sigma} = \omega_1\sigma_1 + \cdots + \omega_k\sigma_k$. Therefore, the optimal allocation n_j for each stratum j, $j = 1, \cdots, 4$ is shown in table 2.

Table 2: This table presents the optimal allocation n_j for each stratum j, $j = 1, \cdots, 4$.

Stratum j	**Stratum 1**	**Stratum 2**	**Stratum 3**	**Stratum 4**
n_j	6	10	15	97

The proportional allocation is given by

$$\overline{n}_j = n\omega_j, j = 1, \cdots, k. \tag{17}$$

Therefore, the proportional allocation $\overline{n}_j$ for each stratum j, $j = 1, \cdots, 4$ is shown in table 3.

Table 3: This table presents the proportional allocation $\overline{n}_j$ for each stratum j, $j = 1, \cdots, 4$.

Stratum j	**Stratum 1**	**Stratum 2**	**Stratum 3**	**Stratum 4**
$\overline{n}_j$	32	32	32	32

The variance of the estimate of the population mean obtained using simple random sampling is given by

$$Var(\overline{X}) = \frac{\sigma^2}{n} \tag{18}$$

where $\sigma^2 = \overline{\sigma^2} + \sum_{j=1}^{k} \omega_j(\mu_j - \mu)^2$ and n is the total sample size. Therefore, the variance of the estimate of the population mean obtained using simple random sampling is 23.47303.

The variance of the estimate of the population mean obtained using proportional allocation is given by

$$Var(\overline{X}_{sp}) = \frac{\overline{\sigma^2}}{n} \tag{19}$$

where $\overline{\sigma^2} = \omega_1\sigma_1^2 + \cdots + \omega_k\sigma_k^2$. Therefore, the variance of the estimate of the population mean obtained using proportional allocation is 11.10053.

The variance of the estimate of the population mean obtained using optimal allocation is given by

$$Var(\overline{X}_{so}) = \frac{\overline{\sigma}^2}{n}. \tag{20}$$

Therefore, the variance of the estimate of the population mean obtained using optimal allocation is 4.601028.
Hence, $Var(\overline{X}) \geq Var(\overline{X}_{sp}) \geq Var(\overline{X}_{so})$.

11 Question 65n.

Comparing with part (m), the estimates of the population mean for 8, 16, 32 and 64 strata are shown in table 4.

Table 4: This table presents the estimates of the population mean for 4, 8, 16, 32 and 64 strata for fixed total sample size n=128.

Strata	$Var(\overline{X})$	$Var(\overline{X}_{sp})$	$Var(\overline{X}_{so})$
4	23.47303	11.10053	4.601028
8	19.76497	6.749478	2.20453
16	23.09793	4.496778	1.427741
32	20.69023	1.031386	0.6436891
64	18.36876	0.671849	0.3087445