EXERCISE 2

QUESTION 1.

Extract a simple random sample (SRS) of 10 employees out of 100:

(i). without replacement

(ii) with replacement

Solution

(i).

```
# Select Random Sample Without Replacement

data<-1:100

data

sample(data, size = 10)
```

Output:

```
[1] 61 67 49 11  1 77 90  3 72 23

[1] 82 47 65 86 72 14 73 52 77  9

[1] 54 29 44 60 52 71 40 65 48  5
```

(ii)

```
# Select Random Sample with Replacement

data<-1:100

sample(data, size = 10, replace = T)
```

Output:

```
[1] 83   86  97  23  5    55  75   15 12  52

[1]  50  75  42  40  84  13 100  34  75   1

[1] 14    2  78  82  41  62  13   31  47  2
```

QUESTION 2. Draw a simple random sample of size 2,000 from a telephone book with 10,000 entries:

(i) Without replacement

(ii) With replacement

Solution

(i).

```
# Select Random Sample without Replacement

data<-1:10000

data

sample(data, size = 2000)

# Select Random Sample with Replacement

data<-1:10000

sample(data, size = 2000, replace = T)
```

QUESTION 3. Draw a simple random sample of size 4 (with replacement, without replacement) from the following list:

Ahmed, Ali, Mona, Amal, Emad, Nayef, Saed, Reuof, Khlood, Mohamed, Mansour, Reem, Alanoud, Hafsa

Solution

HOME WORK

QUESTION 4. A hospital has 1,125 patient records. How can one randomly select 120 records to review?

Solution

```
data<-1125
data
sample(data, size = 120)
```

Output:

```
[1]  246  716 1072  256  144  865  354  402 1105  826  423  201  393   27 1008
[16]  374  642   81 1120  550  957  772  789 1069  249  555  324 1061  757  920
[31]  251 1014  308  573  743  688  948  429  214  815 1034  158 1051  736  835
[46]  437  319  352  364  148  726 1097  346  862  581  271  837   33 1049  823
[61] 1090  890  824  682  918  464  943  556  103  543  634  841  607  905  877
[76]  729  770  155  717 1060 1113 1122  750  640  194  636 1024  440  416  594
[91]  808  408  396 1100  953  108  481  310  735   53   62  257   98  224  420
[106]  875 1027  197  413 1081   75  806  664  712   85  468 1117 1115  644   50
```

QUESTION 5. Assume you wish to randomly sample from a population of 3078 teachers across Saudi Arabia to estimate the average amount of time they spent grading homework in a specific year. Assume also the information from the past that the average and standard deviation of the time they spent have been 297 hours and 344 hours. What size sample would be need:

(a) If standard error does not exceed 18 hours.

(b) If the error of estimation no greater than e = 40 hours.

(c) If the confidence width for $\overline{Y}$ does not exceed w = 60 hours.

(d) If the relative error not larger than 12%.

Solution

Given: N=3078, Sample Mean $(\overline{Y}) = 297$, Standard Division (S)=344

(a). s.e=18

The standard deviation of $\bar{y}$, known as its standard error (S.E.), is obtained from

$$\text{S.E.}(\bar{y}) = \sqrt{V(\bar{y})} = \sqrt{\frac{(1-f)}{n}}S$$

where $f = n/N$ is the **sampling fraction** and $(1-f)$ is the finite population correction (fpc).

$$18 = \sqrt{\frac{1-f}{n}}S$$

(i) If f=0:

$$n_1 = \left(\frac{S}{18}\right)^2 = \left(\frac{344}{18}\right)^2 = 356$$

(ii) If f≠0:

$$n = \frac{n_1}{1 + {n_1}/{N}} = 326$$

R

N=3078; S=344; V=182; e=40; Z=1.96**

Part (a)

when f=n/N=0

nl=S2/V**

nl

[366]

when f is not assumed to be 0

n=nl/(1+nl/N) n

[327]

(b)

e=40

$$P\{|\overline{y} - \overline{Y}| \leq e\} = 95\%$$

But

$$P\left\{\frac{|\overline{y} - \overline{Y}|}{s.e(\overline{y})} \leq z_{\frac{\alpha}{2}}\right\} = 95\%$$

Hence

$$e = s.e(\overline{y}) \times z_{\frac{\alpha}{2}}$$

(i) If f=0:

$$n_1 = \left(\frac{S * z_{\frac{\alpha}{2}}}{e}\right)^2 = \left(\frac{344 * 1.96}{40}\right)^2 = 284$$

(ii) If f≠0:

$$n = \frac{n_1}{1 + {}^{n_1}/_{N}} = \frac{284}{1 + {}^{284}/_{3078}} = 260$$

R

N=3078; S=344; V=182; e=40; Z=1.96**

when f=n/N=0

nl=(Z*S/e)2**

nl

[366]

 # when f is not assumed to be 0

n=nl/(1+nl/N)

n

[327]

(c). w=60

$$w = 2\, z_{\frac{\alpha}{2}}\ s.e(\overline{y})$$

$$s.e(\overline{y}) = \frac{60}{2\, z_{\frac{\alpha}{2}}}$$

(i) If f=0:

$$n_1 = \left(\frac{S * \ 2\, z_{\frac{\alpha}{2}}}{60}\right)^2 = \left(\frac{344 * 2 * \ 1.96}{60}\right)^2 = 505$$

(ii) If f≠0:

$$n = \frac{n_1}{1 + {}^{n_1}/_{N}} = \frac{505}{1 + {}^{505}/_{3078}} = 433$$

R

N=3078; S=344; V=182; e=40; Z=1.96**

when f=n/N=0

nl=(S*2*Z/60)2**

nl

[505]

 # when f is not assumed to be 0

n=nl/(1+nl/N) n

[433]

(d). Relative error not larger than (r=12%)

$$P\left\{\frac{|\overline{y}-\overline{Y}|}{\overline{Y}} \leq relative\ (r)\right\} = 95\%$$

$$e = 0.12\ \times \overline{Y} = 0.12 * 297 = 35.64$$

(iii) If f=0:

$$n_1 = \left(\frac{S *\ z_{\frac{\alpha}{2}}}{e}\right)^2 = \left(\frac{344 *\ 1.96}{35.64}\right)^2 = 358$$

(iv) If f≠0:

$$n = \frac{n_1}{1 + {}^{n_1}/_{N}} = \frac{358}{1 + {}^{358}/_{3078}} = 321$$

R

N=3078; S=344; V=182; e=40; Z=1.96**

e=0.12*297

when f=n/N=0

nl=(S*Z/e)2**

nl

[358]

 # when f is not assumed to be 0

n=nl/(1+nl/N) n

[433]

QUESTION 6. A simple random sample of 10 farms is selected from a population of 50 farms in a particular district. The numbers of cows in the sample farms are: 23; 14; 38; 11; 7; 31; 9; 18; 12; 25.

(a) Estimate the mean number of cows per farm.

(b) Estimate the variance of your estimator.

(c) Estimate the total number of cows in this population.

(d) Estimate the variance of your estimator.

(e) Give the inclusion probability for any particular farm in the population.

(f) How many possible samples are there?

(g) What is the probability of selecting the particular sample in this problem?

(h) Give an approximate 95% confidence interval for the population mean

(i) Give approximate 95% confidence interval for the population total.

Solution

(a) Estimate the mean number of cows per farm.

R

n=10; N=50

f=n/N, z=1.96

y=c(23,14,38,11,7,31,9,18,12,25)

estimate_mean=mean(y)

estimate_mean

[18.8]

(b) Estimate the variance of your estimator (i.e. $Var(\overline{y})$).

$$Var(\overline{y}) = \frac{1-f}{n} \times s^2 = \frac{1 - {}^{10}/_{50}}{10} \times (10.21763)^2 = 8.352$$

n=10; N=50

f=n/N, z=1.96

y=c(23,14,38,11,7,31,9,18,12,25)

estimate_mean=mean(y)

estimate_mean

s=sd(y)

s

variance_estimator=(1-f)*s**2/n

variance_estimator

[8.352]

(c) Estimate the total number of cows in this population.

The estimate of Y is

$$\hat{Y} = N \times \overline{y} = 50 \times 18.8 = 940$$

population total t=N*population mean

(Y_hat=N*estimate_mean)

[940]

(d) Estimate the variance of your estimator (i.e. $Var(\hat{Y})$):

$$Var(\hat{Y}) = N^2 \times Var(\overline{y}) = (50)^2 \times 8.352 = 20880$$

var(Y_hat)=N^2*var(estimate_mean)

var_Y_hat=N**2*variance_estimator

var_Y_hat

[2088]

(e) Give the inclusion probability for any particular farm in the population.

$$P\{i \in S\} = \frac{n}{N} = \frac{10}{50} = 0.20, \quad i = 1,2, \dots ,50$$

Probability_inclusion=n/N

probability_inclusion

[0.2]

(f) How many possible samples are there?

$$\binom{50}{10} = \frac{50! \times 10!}{40!} = 10272278170$$

number_of_samples=choose(N,n)

number_of_samples

[10272278170]

(g) What is the probability of selecting the particular sample in this problem?

$$P\{sample\} = \frac{1}{10272278170} = 9.734939\ e^{-11}$$

probability of selecting the particular sample =1/choose(N,n)

[9.734939 e-11]

(h) Give an approximate 95% confidence interval for the population mean

$$C.I.(\overline{Y}) = \left(\overline{y} - 1.96 \times \sqrt{Var(\overline{y})}, \quad \overline{y} + 1.96 \times \sqrt{Var(\overline{y})}\right)$$
$$= (13.13563;\ 24.46437)$$

##CI=sample_mean(+or-) *z* sqrt (variance_estimator)

CI1=estimate_mean+c(-1,1)*z*sqrt(variance_estimator)

CI1

[13.13563 24.46437]

(i) Give approximate 95% confidence interval for the population total.

$$C.I.(Y) = \left(\hat{Y} - 1.96 \times \sqrt{Var(\hat{Y})}, \quad \hat{Y} + 1.96 \times \sqrt{Var(\hat{Y})}\right)$$
$$= (656.7817;\ 1223.2183)$$

CI2=Y_hat+c(-1,1)*z*sqrt(var_Y_hat))

[656.7817; 1223.2183]

QUESTION 7. Consider a population of N = 8 units. The value yi are given in the following table.

i	1	2	3	4	5	6	7	8
y_i	1	2	4	4	7	7	7	8

(a) Determine the population mean $\overline{Y}$ and population variance S^2.

(b) Select all the 56 samples of size 3 from the population of 8 units without replacement.

(c) Obtain all the sample values y_1; y_2; y_3.

(d) Obtain all the sample means y.

(e) Obtain all the sample standard deviation s.

(f) Verify that the sample mean is an unbiased estimator of the population mean.

(g) Verify that the sample variance is an unbiased estimator of the population variance

(h) By using the value of the population standard deviation S, determine all the 95% confidence intervals for the population mean.

- Deduce the proportion of the confidence intervals enclosing the actual population mean $\overline{Y}$.
- By using the values of the sample standard deviation s, determine all the 95% confidence intervals for the population mean.
- Deduce the proportion of the confidence intervals enclosing the actual population mean $\overline{Y}$.

Solution

N=8

Y=c(1, 2, 4, 4, 7, 7, 7, 8)

(a) Determine the population mean $\overline{Y}$ and population variance S^2.

pop_mean=mean(Y)

pop_mean

[5]

pop_variance=var(Y)

pop_variance

[6.857]

(b) Select all the 56 samples of size 3 from the population of 8 units without replacement.

```
install.packages("combinat")
library("combinat")
```

n=3

pop=1:8

samples=combn(pop,n)

samples

```
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]    1    1    1    1    1    1    1    1    1     1
[2,]    2    2    2    2    2    2    3    3    3     3
```

```
[3,]    3    4    5    6    7    8    4    5    6    7
     [,11] [,12] [,13] [,14] [,15] [,16] [,17] [,18] [,19]
[1,]    1    1    1    1    1    1    1    1    1
[2,]    3    4    4    4    4    5    5    5    6
[3,]    8    5    6    7    8    6    7    8    7
     [,20] [,21] [,22] [,23] [,24] [,25] [,26] [,27] [,28]
[1,]    1    1    2    2    2    2    2    2    2
[2,]    6    7    3    3    3    3    3    4    4
[3,]    8    8    4    5    6    7    8    5    6
     [,29] [,30] [,31] [,32] [,33] [,34] [,35] [,36] [,37]
[1,]    2    2    2    2    2    2    2    2    3
[2,]    4    4    5    5    5    6    6    7    4
[3,]    7    8    6    7    8    7    8    8    5
     [,38] [,39] [,40] [,41] [,42] [,43] [,44] [,45] [,46]
[1,]    3    3    3    3    3    3    3    3    3
[2,]    4    4    4    5    5    5    6    6    7
[3,]    6    7    8    6    7    8    7    8    8
     [,47] [,48] [,49] [,50] [,51] [,52] [,53] [,54] [,55]
[1,]    4    4    4    4    4    4    5    5    5
[2,]    5    5    5    6    6    7    6    6    7
[3,]    6    7    8    7    8    8    7    8    8
     [,56]
[1,]    6
[2,]    7
[3,]    8
```

(c) Obtain all the sample values y_1; y_2; y_3.

```
sample_values=combn(Y,3)
```

sample_values

```
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]    1    1    1    1    1    1    1    1    1     1
[2,]    2    2    2    2    2    2    4    4    4     4
[3,]    4    4    7    7    7    8    4    7    7     7
     [,11] [,12] [,13] [,14] [,15] [,16] [,17] [,18] [,19]
[1,]     1     1     1     1     1     1     1     1     1
[2,]     4     4     4     4     4     7     7     7     7
[3,]     8     7     7     7     8     7     7     8     7
     [,20] [,21] [,22] [,23] [,24] [,25] [,26] [,27] [,28]
[1,]     1     1     2     2     2     2     2     2     2
[2,]     7     7     4     4     4     4     4     4     4
[3,]     8     8     4     7     7     7     8     7     7
     [,29] [,30] [,31] [,32] [,33] [,34] [,35] [,36] [,37]
[1,]     2     2     2     2     2     2     2     2     4
[2,]     4     4     7     7     7     7     7     7     4
[3,]     7     8     7     7     8     7     8     8     7
     [,38] [,39] [,40] [,41] [,42] [,43] [,44] [,45] [,46]
[1,]     4     4     4     4     4     4     4     4     4
[2,]     4     4     4     7     7     7     7     7     7
[3,]     7     7     8     7     7     8     7     8     8
     [,47] [,48] [,49] [,50] [,51] [,52] [,53] [,54] [,55]
[1,]     4     4     4     4     4     4     7     7     7
[2,]     7     7     7     7     7     7     7     7     7
[3,]     7     7     8     7     8     8     7     8     8
     [,56]
[1,]     7
[2,]     7
[3,]     8
```

(d) Obtain all the sample means y.

```
sample_means=apply(sample_values,2,mean)
sample_means
[1] 2.333333 2.333333 3.333333 3.333333 3.333333 3.666667
 [7] 3.000000 4.000000 4.000000 4.000000 4.333333 4.000000
[13] 4.000000 4.000000 4.333333 5.000000 5.000000 5.333333
[19] 5.000000 5.333333 5.333333 3.333333 4.333333 4.333333
[25] 4.333333 4.666667 4.333333 4.333333 4.333333 4.666667
[31] 5.333333 5.333333 5.666667 5.333333 5.666667 5.666667
[37] 5.000000 5.000000 5.000000 5.333333 6.000000 6.000000
[43] 6.333333 6.000000 6.333333 6.333333 6.000000 6.000000
[49] 6.333333 6.000000 6.333333 6.333333 7.000000 7.333333
[55] 7.333333 7.333333
```

(e) Obtain all the sample standard deviation s.

```
sample_sd=apply(sample_values,2,sd)
sample_sd
[1] 1.5275252 1.5275252 3.2145503 3.2145503 3.2145503
 [6] 3.7859389 1.7320508 3.0000000 3.0000000 3.0000000
[11] 3.5118846 3.0000000 3.0000000 3.0000000 3.5118846
[16] 3.4641016 3.4641016 3.7859389 3.4641016 3.7859389
[21] 3.7859389 1.1547005 2.5166115 2.5166115 2.5166115
[26] 3.0550505 2.5166115 2.5166115 2.5166115 3.0550505
[31] 2.8867513 2.8867513 3.2145503 2.8867513 3.2145503
[36] 3.2145503 1.7320508 1.7320508 1.7320508 2.3094011
[41] 1.7320508 1.7320508 2.0816660 1.7320508 2.0816660
[46] 2.0816660 1.7320508 1.7320508 2.0816660 1.7320508
[51] 2.0816660 2.0816660 0.0000000 0.5773503 0.5773503
```

[56] 0.5773503

(f) Verify that the sample mean is an unbiased estimator of the population mean.

mean(sample_means)

[5]

(g) Verify that the sample variance is an unbiased estimator of the population variance

mean(sample_sd**2)

[6.857143]

(h) By using the value of the population standard deviation S, determine all the 95% confidence intervals for the population mean.

- Deduce the proportion of the confidence intervals enclosing the actual population mean $\overline{Y}$.
- By using the values of the sample standard deviation s, determine all the 95% confidence intervals for the population mean.
- Deduce the proportion of the confidence intervals enclosing the actual population mean $\overline{Y}$.

> S=sd(Y); f=n/N

> CI1=sample_means-1.96*sqrt((1-f)/n)*S

> CI2=sample_means+1.96*sqrt((1-f)/n)*S

> cbind(CI1,CI2)

	CI1	CI2
[1,]	-0.009314741	4.675981
[2,]	-0.009314741	4.675981

[3,] 0.990685259 5.675981

[4,] 0.990685259 5.675981

[5,] 0.990685259 5.675981

[6,] 1.324018592 6.009315

[7,] 0.657351926 5.342648

[8,] 1.657351926 6.342648

[9,] 1.657351926 6.342648

[10,] 1.657351926 6.342648

[11,] 1.990685259 6.675981

[12,] 1.657351926 6.342648

[13,] 1.657351926 6.342648

[14,] 1.657351926 6.342648

[15,] 1.990685259 6.675981

[16,] 2.657351926 7.342648

[17,] 2.657351926 7.342648

[18,] 2.990685259 7.675981

[19,] 2.657351926 7.342648

[20,] 2.990685259 7.675981

[21,] 2.990685259 7.675981

[22,] 0.990685259 5.675981

[23,] 1.990685259 6.675981

[24,] 1.990685259 6.675981

[25,] 1.990685259 6.675981

[26,] 2.324018592 7.009315

[27,] 1.990685259 6.675981

[28,] 1.990685259 6.675981

[29,] 1.990685259 6.675981

[30,] 2.324018592 7.009315

[31,] 2.990685259 7.675981

[32,] 2.990685259 7.675981

[33,] 3.324018592 8.009315

[34,] 2.990685259 7.675981

[35,] 3.324018592 8.009315

[36,] 3.324018592 8.009315

[37,] 2.657351926 7.342648

[38,] 2.657351926 7.342648

[39,] 2.657351926 7.342648

[40,] 2.990685259 7.675981

[41,] 3.657351926 8.342648

[42,] 3.657351926 8.342648

[43,] 3.990685259 8.675981

[44,] 3.657351926 8.342648

[45,] 3.990685259 8.675981

[46,] 3.990685259 8.675981

[47,] 3.657351926 8.342648

[48,] 3.657351926 8.342648

[49,] 3.990685259 8.675981

[50,] 3.657351926 8.342648

[51,] 3.990685259 8.675981

[52,] 3.990685259 8.675981

[53,] 4.657351926 9.342648

[54,] 4.990685259 9.675981

[55,] 4.990685259 9.675981

[56,] 4.990685259 9.675981

> sum(CI1<=5 & CI2>=5)/56*100

[96.42857]

(CI1=sample_means-1.96*sqrt((1-f)/n)*sample_sd)

(CI2=sample_means+1.96*sqrt((1-f)/n)*sample_sd)

cbind(CI1,CI2)

sum(CI1<=5 & CI2>=5)/56*100

[85.7]

QUESTION 8. To determine the sample size n, it can be required that $\overline{y}$ should not differ from $\overline{Y}$ by more than a specified error *e*. This requirement can be expressed as the probability statement:

$$P\left[\left|\bar{y}-\bar{Y}\right|<e\right]=(1-\alpha)$$

Show that the value of *n* is given by

$$e = z_{\alpha/2}\sqrt{\frac{1-f}{n}}\times s$$

$$e = z_{\alpha/2}\sqrt{1-\frac{n}{N}}\frac{S}{\sqrt{n}}$$

Solution

$$P\{|\overline{y}-\overline{Y}| < e\} = 1-\alpha$$

But

$$\frac{|\overline{y}-\overline{Y}|}{s.e(\overline{y})} \sim N(0,1)$$

Hence

$$P\left\{\frac{|\overline{y}-\overline{Y}|}{s.e(\overline{y})} < \frac{e}{s.e(\overline{y})}\right\} = 1-\alpha$$

So

$$z_{\alpha/2} = \frac{e}{s.e(\overline{y})}$$

$$e = z_{\alpha/2} \times s.e(\overline{y})$$

$$e = z_{\alpha/2} \times \sqrt{var(\overline{y})}$$

$$e = z_{\alpha/2} \times \sqrt{\frac{1-f}{n} \times s^2}$$

QUESTION 9. Which of the following SRS designs will give the most precision for estimating a population mean $\overline{Y}$? Assume that each population has the same value of the population variance s^2.

(1). An SRS of size 400 from a population of size 4000

(2). An SRS of size 30 from a population of size 300

(3). An SRS of size 3000 from a population of size 300; 000; 000.

Solution

(1). n=400; N=4000

$$var(\overline{y}) = \frac{1-f}{n} \times s^2 = \left(\frac{1-{}^{400}/_{4000}}{400}\right) \times s^2 = (0.00225) \times s^2$$

(2). n=30; N=300

$$var(\overline{y}) = \frac{1-f}{n} \times s^2 = \left(\frac{1 - {}^{30}/_{300}}{30}\right) \times s^2 = (0.003) \times s^2$$

(3). n=3000; N=300,000,000

$$var(\overline{y}) = (0.00033) \times s^2$$

So the most precision in case 3.