Department of Statistics
Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon
Course: M.Sc. Statistics (I)
Practical Sheet: ST-106 (A)
Topic: Introduction to different statistical Software Packages.

1. Construct the following graphs of the following data

58.2	66.9	53.5	53.0	53.2	52.5	53.4	56.5	65.3	70.7
55.3	53.4	52.1	51.5	51.5	52.4	53.3	55.5	61.2	69.6
69.3	58.5	55.3	53.6	52.3	51.5	51.7	51.5	52.5	57.1
63.6	68.8	68.9	60.1	55.6	53.9	53.3	53.1	53.5	53.5
53.9	57.1	64.7	69.4	70.3	62.6	57.9	55.8	54.8	54.2
54.6	54.3	54.8	58.1	68.1	73.3	75.5	66.4	60.5	57.7

   i) Steam & leaf plot
   ii) Histogram
   iii) Dot plot
   iv) Box plot

2. Frequent airline travelers were asked to indicate the airline they believe offered the best overall services. The four choices were American Air (A),East coast Air (E).Sum coast (C) and Great Western (W).The following data were obtained

   E A E S W W E S W E E A S S W E A W W
   S E E A E E S W A S E A W A A W E S W

   Summarize the data using appropriate graph.

3. The fire department laboratory tested the flammability of frabric, using the same methods in five different labs. The measurements are the length of the burned portion of a piece of fabric held over flame for a fixed amount of time.

Sample No.	Lab 1	Lab 2	Lab3	Lab4	Lab5
1	2.9	2.7	3.3	3.3	4.1
2	3.1	3.4	3.3	3.2	4.1
3	3.1	3.6	3.5	3.4	3.7
4	3.7	3.2	3.5	2.7	3.1
5	4.2	4.1	2.8	3.3	3.5
6	3.7	3.8	3.2	2.9	2.8
7	3.9	3.8	2.8	3.2	3.5
8	3.1	4.3	3.8	2.9	3.7
9	3.0	3.4	3.5	2.6	3.5
10	2.9	3.3	3.8	2.8	3.9

   a) Compare the response of the above experiment using box plot, dot plot and interval plot.
   b) Give the above data in the form of the table.
   c) Combine the all data & draw histogram and identify the distribution of given data.
   d) Draw normal probability plot of given data.

4. The following data represent the scores obtained by 27 students in a midterm test

   79 78 78 67 76 87 85 73 66 99 84 72 66 57
   94 84 72 63 57 48 50 61 71 82 93 100 89

   a) Prepare steam & leaf display of the data
   b) Calculate the range and the inter-quartile rang.

Department of Statistics
Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon
M.Sc. (I) Statistics (Industrial Statistics)
Practical Sheet: ST-106 (C)
Topic 2: Random Sampling and Sampling distributions

---

1. Generate random sample from the finite population using (i) SRSWR (ii) SRSWOR

   (iii) Unequal probabilities for population elements.

2. Generate random sample of size 10,000 from $Poisson(\lambda = 3)$. Treat this generated sample data as characteristic values of study variables of the population of the size 10,000.
   Now, draw 5 random samples of size 100 each from this population using SRSWOR, obtain estimate of population mean based on each sample data by using sample mean as an estimator for population mean. Also find the amount of sampling error from your estimates.

3. Draw 10,000 samples of size 25 from Normal $(100, 3^2)$ and study the sampling distribution of following statistics for n=5, 10, 25.

   (i) $\bar{Y}$ (ii) $S_Y^2 = \frac{1}{n-1}\sum_{r=1}^{n}\left(Y_r - \bar{Y}\right)^2$ (iii) $Y_{(1)} = \min\left(Y_1, Y_2, \ldots Y_n\right)$ (iv) $Y_{(n)} = \max\left(Y_1, Y_2, \ldots Y_n\right)$

   (v) $R = Y_{(n)} - Y_{(1)}$

4. Consider the population of size N=5 with 'y' values as (138, 142, 145, 155,143). Write down all possible samples of size n=3 using SRSWOR for each sample obtain values of (i)

   $T_1 = \bar{Y}$ (ii) $T_2 = \frac{Y_1 + Y_3}{2}$ (iii) $T_3 = SampleMedian$

   Now verify (a) $E(T_1) = \bar{y}$ i.e. T_1 is unbiased estimator for $\bar{y}$ (population mean).

   (b) $E(T_2) \neq \bar{y}$ i.e. T_2 is biased estimator for $\bar{y}$.

   (c) $E(T_3) \neq \bar{y}$ i.e. T_3 is biased estimator for $\bar{y}$.

   Compare T_1, T_2 and T_3 in terms of their M.S.E.

Department of Statistics
Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon
Practical Sheet: ST-106 (C) M.Sc. (I) Statistics
Topic 3: Estimation of population parameters in SRS.

1. The following data gives the geographical area (in acres) under paddy for 58 villages. Draw sample of eight villages by using SRSWOR. Based on sample drawn find the estimate of 'average area per village under paddy' and estimate of its variance and 95% confidence interval for 'average area per village under paddy'.

   98, 19, 47, 137, 155, 270, 64, 69, 127, 292, 79, 81, 144, 104, 240, 273, 141, 56, 117, 201, 130, 58, 102, 170, 261, 158, 29, 102, 210, 189, 116, 46, 187, 101, 194, 93, 161, 222, 41, 127, 179, 223, 33, 114, 76, 96, 78, 88, 137, 114, 56, 108, 179, 318, 58, 58, 76, 272

2. The following are the monthly expenditures of 25 households selected randomly by using SRSWOR from a village having $N = 400$ households.

   1698, 1889, 1910, 1768, 1852, 1837, 1811, 1762, 1722, 1851, 1717, 1715, 1791, 1908, 1801, 1800, 1893, 1771, 1709, 1772, 1667, 1690, 1811, 1816, 1731.

   Obtain an unbiased estimate of monthly average expenditure per household in the village. Also, provide an estimate of 98% C.I. for the monthly average expenditure per household.

3. A SRS of 30 households was drawn from a city area containing 14,848 households. The number of persons per household in sample was as follows

   5, 6, 3, 3, 2, 3, 3, 3, 4, 4, 3, 2, 7, 4, 3,5, 4, 4, 3, 3, 4, 3, 3, 1, 2, 4, 3, 4, 2, 4

   Estimate the total number of people in the area and compute the probability that this estimate is within $\pm$ 10 % of the true value.

4. The following table shows the weekly family income (X) and weekly family expenditure on food (Y) in SRS of 30 low-income families. From this sample estimate the mean weekly expenditure on food per family and the percentage of income spent on food.

Family No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Income (X)	620	620	870	650	580	920	880	790	830	620	630	620	600	750	900
Food Cost (Y)	143	208	227	305	412	282	242	300	242	444	134	198	294	271	222
Family	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Income (X)	750	690	830	850	730	660	580	770	690	650	770	690	950	770	690
Food Cost (Y)	377	226	360	206	277	259	233	398	168	378	348	287	630	195	216

Department of Statistics
Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon
Practical Sheet: ST-106 (C) M.Sc. (I) Statistics
Topic 4: Determination of Sample size and Estimation of population proportion.

1. How large sample should be taken from a population of 2000 units so that sample estimate of population mean ($\bar{y}$) differs from its true value by a quantity less than 20 with probability 0.95? It is given that an estimate of population variance is 10,000.
   $\left(i.e.\ \text{we have}\ N = 2000,\ \ \varepsilon = 20,\ \ 1-\alpha = 0.95,\ \ S_Y^2 = 10{,}000\right)$
2. Suppose it is required to estimate the average value (population mean) of output of a group of 5000 factories in a region so that the sample estimate lies within 10% of the true value with confidence coefficient of 95%. Determine the minimum sample size required. The population coefficient of variation is known to be 60%.
   $\left(i.e.\ \text{we have}\ N = 5000,\ \ \varepsilon = 0.1*\bar{y},\ \ 1-\alpha = 0.95,\ \ CV = 60\%\ \ i.e\ \ {s_y}/{\bar{y}} = 0.6\right)$
3. Select a random sample of size 200 from *Bernoulli* (*p=0.67*). Treat it as the result of the 10th class here '1' indicates student is passed and '0' indicates student is failed. Now, based on this sample, estimate the population proportion.
4. The proportion of families having 2 or more children (*p*) in a certain city is to be estimated. There are *N*= 10,000 families, a random sample of $n = 200$ families is selected, and we find that $\hat{p} = 0.55$. Calculate 90% confidence interval for *p*.
5. A preliminary random sample of $n = 50$ students is selected from $N = 4000$ students and it is found that 30 smoke. How large a sample must be selected so that the sample estimate lies within 5% of the true value with confidence coefficient of 90%.
6. A major metropolitan newspaper selected a simple random sample of 1,600 readers from their list of 100,000 subscribers. They asked whether the paper should increase its coverage of local news. Forty percent of the sample wanted more local news. Estimate the 99% confidence interval for the proportion of readers who would like more coverage of local news?
7. In nurseries that produce young trees for sale it is advisable to estimate how many healthy young trees are likely to be on hand, since this determines policy toward the acceptance of orders. By the complete enumeration of the bed of silver maple seedlings it was found that $\bar{y} = 19, s^2 = 85.6$ and $N = 430$. With SRSWOR, how many units must be taken to estimate $\bar{y}$ within 10% apart from the chance of 1 of 20?

Department of Statistics
Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon
Practical Sheet: ST-106 (C) M.Sc. (I) Statistics
Topic 5: Stratified random sampling & estimation of population parameters

1. Consider a population with *N=250* and four strata with sizes $(N_1, N_2, N_3, N_4) = (50,30,100,70)$. Draw stratified random sample of size of *n=25* with respective sample sizes for four strata as $(n_1, n_2, n_3, n_4) = (5,3,12,5)$.
2. From a population with *N=5000* and *3* strata with sizes $N_1 = 1000, N_2 = 1500, \quad N_3 = 2500$. In the following table sample y-values are given

Stratum	N_h	n_h	Sample y-values
1	1000	10	5, 4, 1, 2, 4, 3, 7, 5, 0, 3
2	1500	15	11, 9, 13, 9, 4, 5, 5, 14, 8, 9, 11, 7, 10, 14, 7
3	2500	20	22, 16, 18, 18, 23, 20, 20, 19, 20, 16, 21, 20, 26, 20, 26, 18, 21, 17, 22, 21

Estimate (i) population mean (ii) S.E. of your estimate of population mean by using this sample.

3. The table shows the sizes of 64 cities (in 1000's) in 1920 in United States of America. The cities are arranged in two strata, the first containing the 16 cities and second contains remaining 48 cities. Find the standard error of the 'estimated total size' by (1) drawing the sample of 20 cities from 64 (16+48) cities using simple random sampling (2) drawing the sample using stratified random sampling with stratum sizes $(n_1 = 5 , n_2 = 15)$.

Stratum 1	799,773,778,734,588,577,507,507,457,438,415,401,387,381,324,315
Stratum 2	314,298,296,258,256,243,238,237,235,235,216,208,201,192,180,179, 172,172,163,162,161,159,153,144,138,138,138,138,136,132,130,126, 121,120,119,118,118,116,116,113,113,110,110,108,106,104,101,100

4. A primary school has four classes class I, II, III and IV with number of students 40, 35, 35 and 30 respectively. A stratified random sample with allocation vector (n_1, n_2, n_3, n_4) $(8,5,7,5)$ was drawn and the weights of students selected in sample are noted as below.

Class	Weights of students selected (in kg)	Class	Weights of students selected (in kg)
I	15.5, 14.5, 17.5, 20.5, 21.5, 20, 16, 22	III	24.5, 29.5, 27.5, 24.5, 28, 23, 28.5
II	23.5, 20.5, 24.5, 26.5, 26	IV	25.5, 30.5, 31.5, 28.5, 33

i) Obtain the estimate of mean weight of students in the school.

ii) Obtain 95% C.I. for mean weight of students in the school.

Department of Statistics
Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon
Practical Sheet: ST-106 (C) M.Sc. (I) Statistics
Topic 6: 'Allocations' & 'Estimation of population parameters' in stratified sampling

1. A Loksabha constituency consists of three tehsils with populations sizes $N_1 = 50000, N_2 = 35000$ and $N_3 = 25000$. A politician Mr. X wants to know the proportion $(\hat{p})$ of people who may vote in favour of him in election. The statistician has taken SRS from three tehsils with sample sizes $n_1 = 2000, n_2 = 1500$ and $n_3 = 1500$ respectively. Based on opinion survey conducted to estimate (p), it has been observed that 1050,790 and 450 peoples respectively from three tehsils have voted in favour of Mr. X. Using stratified random sampling method, estimate (p) and S.E. of your estimate of (p).
2. The data shown in table gives the distribution of number of factories, according to number of workers in factories. The stratum mean and standard deviation of output and cost of sampling are given table. Calculate the sampling variance of 'estimated average output from sample of 600 factories' if

   (i) The factories are selected by SRSWOR.

   (ii) The factories are selected by SRSWOR from each stratum with Proportional allocation and Neyman's optimum allocation, using constant cost for each stratum.

   Compare variances of your estimator under SRSWOR design and different allocations

   (iii) If the total cost of survey is Rs.6000 (C) with an overhead cost of Rs.1500 (C_0), find an optimum allocation that will minimize the variance. Find the corresponding value of the variance and compare the same with that for the proportional allocation for the same sample size.

Class: No. of workers	No. of factories	Output per factory (000 Rs.)	Std. Deviation (000 Rs)	Sampling Cost (Rs.)
1-49	18260	100	80	6
50-99	4315	250	200	9
100-249	2233	500	600	10
250-999	1057	1760	1900	12
1000-above	567	2250	2500	15

3. Table gives the stratification of all farms in a country by farm size and the acreage of *corn* in the sampled farms. Obtain the estimate of average acres of corn per farm in the population and its variance. Also, obtain the variances of estimates of population mean under proportional allocation and optimum allocation for the same total sample size of 21 farms. Hence, obtain the efficiency of given allocation and proportional allocation with respect to optimum allocation.

Farm size (acres)	No. of farms(N_h)	Acreage of corn in sampled farms (10 acres)
0-40	12	3.5, 2.7, 1.1, 1.9
41-80	24	2.4, 6.9, 3.7, 2.8, 3.4, 2.6
81-120	19	1, 1.3, 2.9, 1.6, 3.4
121-160	13	1, 0, 2.3, 0.5
161-200	9	5.8, 4.9

Department of Statistics
Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon
Practical Sheet: ST-106 (C) M.Sc. (I) Statistics
Topic 7: Estimation of population parameters using Ratio method of estimation

1. The number of laborers x (in thousands) and the quantity of raw materials y (in lakhs of bales) are given below for 20 jute mills. Draw a sample of 5 units by SRSWOR. Estimate the total amount of raw materials consumed by the 20 mills by

   (i) Sample mean estimator (ii) Ratio estimator.

   Compare the estimators $N\overline{Y}$ and $N\overline{Y}_R$ by using their true M.S.E.s and estimated M.S.E.s

Jute Mill	x	y	Jute Mill	x	y	Jute Mill	x	y	Jute Mill	x	y
1	368	31	6	529	61	11	512	31	16	376	38
2	384	33	7	703	68	12	503	29	17	412	42
3	361	37	8	396	42	13	472	38	18	345	45
4	347	39	9	473	41	14	429	41	19	297	32
5	403	43	10	509	49	15	387	40	20	633	50

2. From the population of $N = 100$ villages 10 villages were selected by SRSWOR design and following data were recorded. If $\overline{x} = 1572$ estimate total population (y_{tot}) of $N = 100$ villages with its S.E. by ratio method of estimation.

Village No.	1	2	3	4	5	6	7	8	9	10
Population of villages (1991) (X)	1707	1373	1630	1373	1376	1452	1160	1511	1628	1492
Population of villages (2001) (Y)	2116	1685	2130	1775	1795	1972	1650	1888	1947	2007

3. Table shows the number of inhabitants (in 10000's) in all 30 cities from a state in 1991(x) and 2001(y). Draw 200 samples of 12 cities by SRSWOR. Estimate the total population of whole state by 'sample mean estimators' and 'ratio estimators' for each sample. Now compare ratio estimate with the estimate based on the sample mean by using Boxplot and Histogram.

x_i	y_i	x_i	y_i	x_i	y_i
76	80	93	104	50	64
138	143	172	183	44	58
67	67	78	106	77	89
29	50	66	86	64	63
381	464	60	57	64	77
23	48	46	65	56	142
37	63	2	50	40	60
120	115	507	634	40	64
61	69	179	260	38	52
387	459	121	113	136	139

Department of Statistics
Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon
Practical Sheet: ST-106 (C) M.Sc. (I) Statistics
Topic 8: Estimation of population parameters by Ratio and Regression method

1. Following table give the yield of paddy and area in acres of 5 villages, selected by simple random sampling without replacement from 12 villages in certain region. We have mean area per village ($\overline{x} = 988.75$)

Village No.	Area(in acres)(X)	Yield of Paddy(Y)
1	1054	10316
2	973	7025
3	1089	10512
4	1054	8963
5	894	8783

Estimate the yield of paddy per village ($\overline{y}$) by ratio method and regression method. Also, obtain the S.E. of your estimates. Estimate efficiency of regression method over ratio method.

2. An experienced farmer makes an eye estimate of the weight of peaches x_i on each tree in an orchard of $N = 200$ trees. He finds $x_{tot} = 11600\ lb$. The peaches are picked and weighed on a simple random sample of 10 trees, with the following results

		Tree Number										
		1	2	3	4	5	6	7	8	9	10	Total
Act. Wt.	Y_i	61	42	50	58	67	45	39	57	71	53	543
Est. wt.	X_i	59	47	52	60	67	48	44	58	76	58	569

Compute the estimate by using ratio and regression method, also find their SEs. Estimate efficiency of regression method over ratio method.

3. Following table gives the number of standards of pepper in 15 clusters of 4 yields each, selected by simple random sampling without replacement out of 300 fields.
   (i) Estimate average number of standards along with its standard error.
   (ii) Estimate relative efficiency of cluster sampling design over SRSWOR design

Cluster	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
field_1	22	53	43	50	73	65	71	24	21	36	72	68	59	43	76
field_2	18	47	29	47	62	71	75	49	72	43	49	64	72	35	58
field_3	27	38	37	41	58	69	31	43	47	51	56	76	67	71	47
field_4	28	29	47	51	47	59	21	75	72	39	69	57	78	40	34

Department of Statistics
Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon
Practical Sheet: ST-106 (C) M.Sc. (I) Statistics
Topic 9: Random Sampling & Estimation of population parameters under PPS designs

1. Consider a population of size $N = 10$, $U = \{1:10\}$ with size vector $x = \{3,1,11,6,4,8,3,7,4,3\}$. Draw a random sample with $n = 3$ by using i) PPSWR ii) PPSWOR.
2. Consider the population $U = \{1:14\}$ with size vector $x = \{9,7,5,6,3,7,3,5,2,7,4,6,4,9\}$
   (i) Using PPSWR design sample of $n = 4$ draws was drawn and the sample is $s = \{2,8,12,8\}$ with its *y*-values $\{29,13,30,13\}$. Obtain an estimate of population mean and its S.E.
   (ii) Using PPSWOR design sample of $n = 3$ draws was drawn and the sample is $s = \{2,4,13\}$ with its $y - values = \{29,25,18\}$. Estimate the population total and mean.
3. Consider the population $U = \{1:20\}$ with size vector and y-values as follows:
   $\underline{x} = \{15,16,9,11,8,9,5,12,11,13,7,9,11,9,11,12,8,12,8,7\}$
   $y = \{61,67,38,46,33,39,22,50,46,52,29,39,47,38,45,48,34,50,34,30\}$
   Draw a random sample of size n = 4 using Midzuno sampling design. Based on your sample drawn, estimate the populations mean using H-T estimator. Obtain variance of H-T estimator for population mean. Also, estimate the variance of H-T estimator for population mean.

Tips:

1. PPS: Probability proportionate to sizes.
2. Midzuno: 1st unit is chosen from population with probabilities proportionate to sizes and other by SRSWOR.
3. PPSWR estimation of parameters: $z_i = y_i/Np_i$
4. $\pi_i = p_i\left(\frac{N-n}{N-1}\right) + \left(\frac{n-1}{N-1}\right)$ $\pi_{ij} = \left(\frac{n-1}{N-1}\right)\left\{p_i + p_j + (1 - p_i - p_j)\left(\frac{n-2}{N-2}\right)\right\}$

are single and double inclusion probabilities respectively for Midzuno sampling.

5. $\overline{Y}_{HT} = \frac{1}{N}\sum_{i=1}^{N}\frac{y_i T_i}{\pi_i}$ $V(\overline{Y}_{HT}) = \frac{1}{N^2}\left(\sum\left(\frac{1-\pi_i}{\pi_i}\right)y_i^2 + \sum_{j=1}^{N}\sum_{\substack{i=1\\ i\neq j}}^{N}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right)y_i y_j\right)$

$$\widehat{V(\overline{Y}_{HT})} = \frac{1}{N^2}\left(\sum\left(\frac{1-\pi_i}{\pi_i}\right)y_i^2\frac{T_i}{\pi_i} + \sum_{j=1}^{N}\sum_{\substack{i=1\\ i\neq j}}^{N}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right)y_i y_j\frac{T_i T_j}{\pi_{ij}}\right)$$