Semester 2	Applied Statistics	2015

Tutorial 11

1. A research center wants to estimate the unemployment rate of a certain district with $N = 25100$ people in the labour force. Since the unemployment rate varies across different age groups, he stratifies the population into three age groups and the results are given below:

Age group l	**Proportion in labour force** W_l	**Sample size** n_l	**Number of unemployed persons** x_l
Below 25	15%	24	6
26-40	50%	60	3
Above 40	35%	40	5

   (a) Calculate the number of people in the labour force for each of the three age groups. [3,765; 12,550; 8,785]

   (b) Estimate the overall unemployment rate in the district and provide an estimate of the standard error for this estimator. [0.10625; 0.0269]

   (c) Assume that the age groups of the respondents are unknown before the interview and hence the stratification is done *after* the interview. Revise the estimate of the standard error for the estimator in part (b). Use $s_l^2 = p_l(1 - p_l)$. [0.0272]

   (d) By comparing the sample allocations $w_l = n_l/n$ with the population proportions W_l, comment on the necessity of using post-stratification in part (c).

   (e) Suppose that this study serves as a pilot study and the costs of sampling people in different age groups are the same. Estimate the total sample size n and the strata sample sizes n_l using *Neyman allocation* such that the error of the estimation is $\delta_\mu = 0.02$ at the 95% level of confidence. [811]

2. A fashion shop estimates the percentage increase in monthly sales after the launch of a promotion campaign. A simple random sample of $n = 10$ branches is drawn from $N = 256$ branches and the monthly sales in thousand dollars and related information are given below:

Branch i	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Monthly sales after y_i	65	109	60	124	128	104	65	61	49	56
Monthly sales before x_i	52	100	60	128	104	98	48	64	96	48
Increase $d_i = y_i - x_i$	13	9	0	-4	24	6	17	-3	-47	8

$$\sum_i x_i = 798,\ \sum_i y_i = 821,\ \sum_i x_i^2 = 71,028,\ \sum_i y_i^2 = 75,765,\ \sum_i x_i y_i = 71,672,$$

$$\bar{d} = 2.3,\ s_d^2 = 377.3444$$

The shop records show that the total monthly sales for all branches before the launch of the campaign was $X = 20,500$ thousand dollars.

(a) Estimate the *percentage increase* in monthly sales after the launch of the promotion campaign and its standard error using
1. the *Ratio* estimator and
2. the estimator $\widehat{R}_o = \frac{\bar{y}}{\bar{X}}$ estimators. [1.028822, 0.076021; 1.025249, 0.117988]]

(b) Estimate the average monthly sales after the launch of the promotion campaign and its standard error using the two estimators in (a). [82.38614, 6.087601]

(c) Estimate the average monthly sales after the launch of the promotion campaign and its standard error using a new estimator defined as

$$\widehat{\overline{Y}}_d = \overline{X} + \bar{d}$$

where $\overline{X}$ is the average monthly sales of all branches before the campaign and $\bar{d}$ is the sample mean of the increases d_i in monthly sales, after the campaign. [82.37813, 6.021664]

(d) Given that the correlation coefficient $\hat{\rho}_{y,x} = 0.7854$, compare the three estimators.

3. A manufacturing company wants to estimate the loss in the number of man-hours due to sickness. A preliminary study of 10 employee records is made and the results are:

Employee i	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Loss in previous year x_i	12	24	15	30	32	26	10	15	1	14
Loss in current year y_i	13	25	15	32	26	24	12	16	2	12

$$\sum_i x_i = 179,\ \sum_i y_i = 177,\ \sum_i x_i^2 = 4067,\ \sum_i y_i^2 = 3843,\ \sum_i x_i y_i = 3927$$

The company record shows that the total number of man-hours lost because of sickness in the previous year was 16,300 and that the company has 1,000 employees.

(a) Find the *ordinary* and *ratio* estimates of the average number of man-hours lost because of sickness this year. Comment on the discrepancy between the two estimates. Standard errors of the estimates are not required. [17.7;16.12]

(b) Find the ratio estimate of the total number of man-hours lost because of sickness this year and provide the standard error of the estimate. [16117.8771; 766.2712]

(c) Find the regression estimate of the total number of man-hours lost because of sickness this year and provide the standard error of the estimate. Compare the regression estimate with the ratio estimate in part (a). Is the use of the variable 'loss in previous year' X as an auxillary variable suitable in this situation? [16293.2090; 687.8880; 1.9623]

(d) The company wants to estimate the rate of change of man-hours lost defined as

$$\text{Rate of change} = \frac{\text{Loss in current year} - \text{loss in last year}}{\text{Loss in last year}}$$

which is the proportion of changes to loss in last year using *ratio* estimation. Treating this study as a pilot study, determine the sample size required to estimate the rate of change for the company with the maximum error of 0.05 and at least 95% confidence. [34]

4. Prove the result in P.30 of the lecture note:

$$\text{var}(\widehat{\bar{Y}}_{st,ds}) \simeq \sum_{l=1}^{L} \frac{\widehat{W}_l^2 S_l^2}{n_l} + \frac{1}{n'} \sum_{l=1}^{L} \widehat{W}_l (\bar{Y}_l - \bar{Y})^2.$$

Extra exercise

1. In studying lung function in a group of 560 workers in a coal mine, an estimate was required of the mean value of some relevant measure Y. A simple random sample of 10 workers was chosen and their Y values, y_i, determined by an appropriate test. A note was also made of their heights, x_i. The results are:

y_i	3.0	3.5	3.3	3.1	4.1	3.2	3.7	2.9	3.9	3.4
x_i (cm)	173	183	170	175	160	157	168	180	178	163

From routine medical records the average height for the group of 560 workers is known to be $\overline{X} = 173.2$ cm.

   (a) Calculate s_x and s_y, the standard deviation of X and Y respectively. [8.72, 0.39]

   (b) Calculate the *correlation coefficient* $\widehat{\rho}$ which is a measure of linear relationship between the two variables X and Y. Compare $\widehat{\rho}$ with $\dfrac{cv(x)}{2cv(y)} = \dfrac{\overline{y}s_x}{2\overline{x}s_y}$. With the information of s_x, s_y and $\widehat{\rho}$ and together with a plot of Y against X as given below, is the ratio estimator for estimating $\overline{Y}$ a good estimator? [-0.2423]

   (c) Estimate $\overline{Y}$ from the data and calculate an approximate standard error for your estimator. [3.41, 0.1231]

2. A survey is conducted to study the total amount of home loan payment for the $N = 832$ households on home loan mortgage in a certain district. The information of the home loan payment Y (in thousand dollars), the household income X_1 (in thousand dollars) and the household size X_2 are presented in the following table:

Household i	**Home loan payment** Y	**Household income** X_1	**Household size** X_2
1	23.245	52.0	5
2	19.825	45.0	4
3	12.746	31.2	6
4	21.365	72.7	3
5	35.420	76.3	5
6	28.200	58.9	2
7	14.855	30.4	5
8	40.210	81.2	4

with summary statistics

$$\sum_i y_i = 195.866, \quad \sum_i x_{1i} = 447.7, \quad \sum_i x_{2i} = 34$$
$$\sum_i y_i^2 = 5,439.616, \quad \sum_i x_{1i}^2 = 27,796.23, \quad \sum_i x_{2i}^2 = 156$$
$$\sum_i x_{1i} y_i = 12,131.95, \quad \sum_i x_{2i} y_i = 804.711$$

(a) Estimate the total amount of home loan payment using *ordinary* estimator. Provide an estimate of the standard error for this estimator. [20370.06; 2808.244]

(b) Two auxiliary variables, the household income X_1 and the household size X_2, are used to improve the efficiency of the estimate. It is known that the average household income and the average household size for the district are $\overline{X}_1 = 54.5$ and $\overline{X}_2 = 4.1$ respectively. Calculate the *ratio* estimators of the total amount of home loan payment, using each of the auxiliary variables and provide estimates of the standard error for each estimator. What is the estimated proportion $\widehat{R}_1$ of household income spent on home loan mortgage? What is the estimated amount $\widehat{R}_2$ of home loan payment per individual? [19837.72, 1330.204; 19651.12, 4058.11]

(c) Which auxiliary variable in part (b) do you prefer? Explain briefly. For your information, the correlation coefficients between Y the home loan payment and the other auxiliary variables: namely X_1, the household income, and X_2, the household size, are, respectively, $\hat{\rho}_1 = 0.881$ and $\hat{\rho}_2 = -0.322$.

(d) Suggest one other use of the auxiliary variable in probability sampling that will lead to easy implementation and/or more efficient estimation.

3. A researcher studied the average household income (in units of $10,000) of all households in a certain district. He divided households into $L = 3$ strata according to the type of dwelling: public, rented private and owned private housing. From a total of 10,000 households in the district, he obtained a simple random sample (SRS) of 50 households in each stratum. Then the average household income $\overline{y}_l$ and the proportion p_l of households which have household incomes under $30,000 for each stratum were calculated as below:

Stratum l	**Stratum weight** W_l	**Stratum mean** $\overline{y}_l$	**Stratum variance** s_l^2	**Stratum proportion** p_l
Public	0.55	2.7	4	0.54
Rented private	0.10	3.5	11	0.39
Owned private	0.35	7.2	85	0.11

(a) Find the population size of each stratum. [5500; 1000; 3500]

(b) Suppose that the researcher wanted to estimate the average household income for households in the district but he knew very little about *Survey Sampling*. He just used the mean of the three stratum sample means to be the estimate of overall mean. In other words, he took

$$\overline{y}_{st1} = \frac{1}{L}\sum_l \overline{y}_l = \frac{1}{3}(\overline{y}_1 + \overline{y}_2 + \overline{y}_3)$$

as the estimate. Calculate this estimate and provide an estimate of variance for this estimator. [4.467; 0.2182]

(c) His colleague suggested him two more estimators for the average household income in the district. The first one regards the sample as one SRS and the estimator is the sample mean

$$\overline{y}_{st2} = \frac{1}{n}\sum_{l}\sum_{i} y_{li}$$

where y_{li} is the household income for household i in stratum l. The second one assumeed a proportional allocation and used the proportions in sample sizes as weight. The estimator is

$$\overline{y}_{st3} = \sum_{l} \frac{n_l}{n}\overline{y}_l.$$

Calculate these two estimates and compare them with $\overline{y}_{st1}$. What is the variance of these 2 estimators? [4.467; 4.467; 0.24135; 0.2182]

(d) Find the proper estimate for the mean household income under stratified simple random sampling and provide an estimate of variance for this estimator. [4.355; 0.2313]

(e) Estimate the difference in proportion of household which have household incomes under $30,000 between households residing in public housing and households under Home Ownership Scheme and provide a 95% confidence interval for the estimate. Ignore the finite population correction factors. Is the difference significant at 0.05 level? [0.15; (-0.045, 0.345)]

(f) Treating this study as a pilot study and that the sample variance is an estimate of population variance in each stratum, allocate a sample of size 300 using *Neyman allocation*. [71; 21; 208]

4. A council plans to open a new market in a certain district. The information on the number of persons (x_i), the family income in thousand dollars (y_i) and the family expenditure on food in thousand dollars (z_i) in a simple random sample of 30 families selected from 660 families are obtained and shown in the following table.

Family	**Size**	**Income**	**Food**	**Family**	**Size**	**Income**	**Food**
i	x_i	y_i	z_i	i	x_i	y_i	z_i
1	2	62	14.3	16	5	75	37.7
2	3	62	20.8	17	3	69	22.6
3	3	87	22.7	18	4	83	36.0
4	5	65	30.5	19	2	85	20.6
5	4	58	41.2	20	4	73	27.7
6	7	92	28.2	21	2	66	25.9
7	2	88	24.2	22	5	58	23.3
8	4	79	30.0	23	3	77	39.8
9	2	83	24.2	24	4	69	16.8
10	5	62	44.4	25	7	65	37.8
11	3	63	13.4	26	3	77	34.8
12	6	62	19.8	27	3	69	28.7
13	4	60	29.4	28	6	95	63.0
14	4	75	27.1	29	2	77	19.5
15	2	90	22.2	30	2	69	21.6

$$\sum_i x_i = 111 \qquad \sum_i y_i = 2195 \quad \sum_i z_i = 848.2$$
$$\sum_i x_i^2 = 477 \qquad \sum_i y_i^2 = 164305 \quad \sum_i z_i^2 = 27060.26$$
$$\sum_i x_i y_i = 8081 \qquad \sum_i x_i z_i = 3364.3 \quad \sum_i y_i z_i = 62771.4$$

(a) Estimate the average number of persons per family and provide the variance of the estimate. [3.7, 0.0727]

(b) By using the information that the average income for families in that district is $\overline{Y} = 72$ thousand dollars, provide the ratio estimate for the average expenditure on food in thousand per family and the variance of this estimator. [27.8225, 3.3816]

(c) If the average expenditure on food in thousand per individual is desired, what will be the estimate using the usual ratio estimator $\widehat{R} = \frac{\overline{z}}{\overline{x}}$? Can you propose another estimator using the results in part (a) and (b)? Variance estimates of both estimators are *not* required. How would you compare these two estimators for the average expenditure on food per individual? Which one is preferred? [7.64, 7.52]