Semester 2	Applied Statistics	2015

Tutorial 10

1. Enumerate all samples of size 2 that can be drawn without replacement from the set $\{0, 3, 3, 4, 5\}$ (population list).

   (a) Calculate the population mean, μ, and variance, σ^2.

   (b) Calculate the sample mean for each sample. Show that the sample mean is an unbiased estimator of μ and its variance is

   $$\frac{S^2}{n}\left(1-\frac{n}{N}\right),$$

   where $N = 5$.

2. In a survey to determine the average height, μ, of a particular population of size $N = 1000$, a sample of n individuals is drawn and the heights of the selected people averaged to estimate μ. Suppose the standard deviation of the heights in the population is $\sigma = 6$ cm.

   (a) If the sample is drawn *with replacement*, what sample size is required to ensure that with probability at least 0.9 the sample average is within 0.5 cm of μ?

   (b) If the sample is drawn *without replacement*, how does the answer to (a) change?

   (c) If we draw a sample of 40 people without replacement, approximate the probability that the average height is at least 1 cm below μ.

3. The complete data set for the 20 children in the "decayed teeth" example is:

No. of decayed teeth per child	0	1	2	3	4	5	6	7	8	9	10
No. of children	8	4	2	2	1	1	0	0	0	1	1

   Calculate the s.e.'s for the total estimators $N\bar{y}_1'$ and $N_1\bar{y}_1$ of Y. Why do Y_1 and Y coincide in this example?

4. The following table shows the stratification of 800 trading firms by their type of business and it includes the standard deviation for the annual profit in thousand dollars in each stratum.

Stratum l	Number of firm N_l	Standard deviation of annual profit S_l
1	480	211
2	320	708

(a) Suppose that a stratified simple random sample of size n is to be taken. The cost of obtaining informations from trading firms in stratum 1 is 0.08 thousand dollars per firm and the cost of obtaining informations from trading firms in stratum 2 is 0.1 thousand dollars per firm. Work out the strata sample sizes, n_1 and n_2 using *optimum allocation* in terms of $n(= n_1 + n_2)$.

(b) If there are only 9.3 thousand dollars to spend on sampling, what will be the sample size and strata sample sizes that use the optimum allocation in part (a).

(c) A stratified simple random sample using the optimum allocation with the sample size and strata sample sizes as determined in part (b) is obtained. The sample means of annual profit in thousand dollars for each stratum are shown in the following table. Estimate the mean annual profit of trading firms and provide a 95% confidence interval of the estimate. Use the population standard deviations S_l given in the previous table in your calculation.

Stratum l	Sample mean of annual profit $\overline{y}_l$
1	540
2	1730

Extra exercise

1. A study to assess the attitudes of accountants toward advertising their services involved sending questionnaires to 200 accountants selected from a list of $N = 1400$ names. A total of $n = 82$ usable questionnaires were returned. The data summary for one question is as follows:

Likelihood of advertising in the future (%)

	All respondents (% out of 82)	Those have advertised before (% out of 46)
Virtual certainty	22	35
Very likely	4	5
Somewhat likely	19	35
About 50-50	18	15
Somewhat unlikely	6	10
Very unlikely	12	0
Absolutely not	15	0
No response	4	0
Total %	100	100

Source: K. Traynor, "Accountant Advertising: Perceptions, Attitudes and Behaviors" Journal of Advertising Research 23, no.6 (1984).

(a) Estimate the population proportion having *at least* 50-50 chance of advertising in the future. Provide the standard error estimate for your answer. Construct a 95% confidence interval for the true proportion. [0.63, 0.0521, (0.528, 0.732)]

(b) Among those who have advertised in the past, estimate the population proportion having *at least* 50-50 chance of advertising again. Provide the standard error estimate for your answer. [0.90, 0.0440]

(c) Estimate the total number of accountants out of the $N = 1400$ accountants having *at least* 50-50 chance of advertising in the future. Provide the standard error estimate for your answer. [882, 72.870]

2. In a survey of the effectiveness of a one-month fitness program, the number of hours spent on exercise during last week x and the amount of weight reduced in pound after the program y for the 10 participants are given below.

Subject i	**Number of hours in exercise** x_i	**Weight reduced in lb** y_i
1	5.0	4.5
2	7.5	8.0
3	10.0	5.5
4	12.0	6.0
5	13.0	8.0
6	15.0	6.5
7	16.5	10.0
8	18.5	12.0
9	20.0	13.0
10	22.5	11.5

A sample of 4 participants is selected from them. We want to measure the average amount of weight reduced in pound for the program. For each of the following sampling schemes and estimators, compute the estimate and the variance of that estimate assuming that only the sample information is known.

(a) A *simple random sample* of size 4 is drawn and that subjects 3,5,8,9 are included in the sample. The estimator used is $\widehat{\overline{Y}}_1 = \overline{y}$ where $\overline{y}$ is the sample mean of y for the 4 selected subjects. [9.625; 1.8344]

(b) A combination of purposive and simple random sampling is used such that subject 3 is *always included* and subjects 5, 8 and 9 are *randomly selected* from the remaining population. The estimator used is $\widehat{\overline{Y}}_2 = \frac{1}{4}y_3 + \frac{3}{4}\overline{y}'$ where y_3 is the y value for subject 3 and $\overline{y}'$ is the sample mean of y for the 3 selected subjects.

Compare the variance of this estimator $\widehat{\overline{Y}}_2$ with the variance of $\widehat{\overline{Y}}_1$. Explain the difference. [9.625; 0.875]

(c) The population of 10 participants are divided into 2 groups: those with the number of hours spent on exercise x less than 14 in the 'light exercise group' and otherwise, in the 'heavy exercise group'. Subjects 3 and 5 are selected from the 'light exercise group' and subjects 8 and 9 are selected from the 'heavy exercise group'. The estimator used is $\widehat{\overline{Y}}_3 = \frac{1}{3}\overline{y}_l + \frac{2}{3}\overline{y}_h$ where $\overline{y}_l$ and $\overline{y}_h$ are respectively the sample means for the light and heavy exercise groups. Compute the estimate $\widehat{\overline{Y}}_3$ and compare the variance of this estimator with the variance of $\widehat{\overline{Y}}_1$. Explain the difference. [10.583; 0.1709]

3. A survey is conducted to investigate certain attitude and behavior for the $N = 9,200$ undergraduate students in a university. A simple random sample of 15 male students and 15 female students are selected from respectively $N_1 = 4,400$ male students and $N_2 = 4,800$ female students. The amounts y_1 and y_2 spent on clothing by male and female students respectively and the indicators x_1 and x_2 of whether clothing is important for male and female students respectively are given below:

	Male students		**Female students**	
Student i	**Amount spent on clothing** y_{1i}	**Clothing is important?** x_{1i}	**Amount spent on clothing** y_{2i}	**Clothing is important?** x_{2i}
1	31	0	56	1
2	54	1	47	1
3	15	1	63	0
4	22	0	48	1
5	34	0	56	0
6	26	0	76	1
7	47	1	49	1
8	61	1	35	0
9	35	0	42	0
10	27	0	55	1
11	18	0	61	1
12	42	0	86	0
13			45	1
14			54	1
15			51	0
	$\sum_i y_{1i} = 412$ $\sum_i y_{1i}^2 = 16390$		$\sum_i y_{2i} = 824$ $\sum_i y_{2i}^2 = 47604$	

* Three male students do not response because of non-contact and refusal and hence $n_1 = 12$ and $n_2 = 15$.

(a) Estimate the total number of *female* students $N_1 P_1$ who agree that clothing is important and provide an estimate of the standard error for this estimator. [2880, 627.485]

(b) Estimate the average amount spent on clothing by male students $\overline{Y}_1$ and female students $\overline{Y}_2$ respectively and provide an estimate of the standard error for each of these estimators. [34.33, 4.118; 54.93, 3.337]

(c) Is there a gender difference in the average amount spent on clothing? Assume that the sample sizes for each group of students are large enough.

(d) We want to estimate the overall average amount spent on clothing $\overline{Y}$ using stratified random sample estimator. Calculate the estimate and the estimated standard error for this estimator. [45.081, 2.627]

4. A household survey is designed to estimate the proportion P of families who own their flat. It is know that the value of P should lie between 20% to 30%.

   (a) With simple random sampling, how large the value of sample size n necessary to estimate P with a margin of error not exceeding 3%? [897]

   (b) Without any knowledge of the possible values of P, can you provide a conservative estimate of n necessary to estimate P with a margin of error not exceeding 3%? [1068]

   (c) If instead of the margin of error given, it is the margin of standard error to be 1.5%, what is the new conservative estimate for the sample size n? (*Hint: The relationship between margin of error and margin of standard error follows* $d = 1.96 \cdot se(\widehat{P})$ *for 0.95 level of confidence.*) [1112]

   (d) Suppose the researcher wishes to estimate $(P' - P)$ with a margin of error not exceeding 3% where P' is the proportion of household who rent their flat. What sample size do you suggest if the researcher thinks that P' lies between 30% to 40% and that the two proportions are independently distributed on the households? What will be the conservative estimate of sample size? (*Hint:* $S^2_{p'-p} = S^2_p + S^2_{p'}$ *or* $var(\widehat{P}' - \widehat{P}) = var(\widehat{P}') + var(\widehat{P})$) [1923, 2134]

Neglect fpc in all your calculations and regard the population size to be very large.