Make sure that this examination has 14 numbered pages

University of Regina
Department of Mathematics & Statistics
Final Examination
200430
(December 17, 2004)

Statistics 257-001
Applied Sampling Techniques

Name: ________________________________ **Student Number:** ________________

Instructor: Michael Kozdron **Time:** 3 hours

Read all of the following information before starting the exam.

*You have **3** hours to complete this exam. Please read all instructions carefully, and check your answers. Show all work neatly and in order, and clearly indicate your final answers. Answers must be justified whenever possible in order to earn full credit. **Unless otherwise specified, no credit will be given for unsupported answers, even if your final answer is correct.** Several problems require written explanations in context. Only complete solutions written in the context specified by the problem will be awarded full points, and points will be deducted for incoherent, incorrect, and/or irrelevant statements.*

You may use standard notation; however, any new notations or abbreviations that you introduce must be clearly defined.

*Calculators are permitted; however, you must still show all your work. You are also permitted to have **ONE** 8.5×11 pages of handwritten notes (double-sided) for your personal use. Other than these exceptions, no other aids are allowed.*

Note that blank space is not an indication of a question's difficulty. The order of the test questions is essentially random; they are not intentionally written easiest-to-hardest.

*This test has **14** numbered pages with **12** questions totalling **200** points. The number of points per question is indicated.*

Solutions will be available on the Statistics 257 course web site after the examinations have all been graded.

DO NOT WRITE BELOW THIS LINE

Problems 1–3	______	Problem 4	______	Problem 5	______
Problem 6	______	Problem 7	______	Problem 8	______
Problem 9	______	Problem 10	______	Problem 11	______
Problem 12	______				
				TOTAL	______

1. (*6 points*) Carefully define what is meant by a *simple random sample of size n from a population of size N.*

2. (*12 points*) List three principal reasons for choosing to use stratified random sampling rather than simple random sampling.

3. (*8 points*) Carefully discuss two situations in which cluster sampling is an effective design for obtaining a specified amount of information at minimum cost.

4. (*12 points*) For the following survey situation, *and in the context of the situation described*, carefully state the target population, the frame, and the sampling units. Also discuss any possible sources of selection bias or inaccuracy of responses, if appropriate.

The Saskatchewan Provincial Travel Committee commissioned a study to identify inter-provincial (within Saskatchewan) travel patterns of Regina and Saskatoon residents, and to evaluate different sources of vacation planning information. They conducted 400 interviews with Regina residents and 400 interviews with Saskatoon residents. Telephone numbers with Regina and Sasktoon exchanges were generated randomly so that listed and unlisted telephone numbers could be reached. "Respndents were limited to heads of household and quotas were established in order to have an equal representation of male and female respndents. Additionally, income and age brackets were monitored in order to maintain the same proportions as the general population bases of Regina and Saskatoon."

5. (*18 points*) A nutritionist at the University of Saskatchewan has decided he would like to know how much pizza students can eat. Using an official list of registered full-time undergraduate students from the Registrar, he selects a simple random sample of size $n = 17$. He provides pizza at lunchtime for these 17 students, and records the total number of slices each student ate. He decides that he would like to estimate μ, the average number of pizza slices that a University of Saskatchewan student can eat at lunchtime, except that he cannot decide which estimator he should use. Let y_i denote the number of pizza slices that the i^{th} student ate, and consider the following two estimators of μ:

$$\overline{y} = \frac{y_1 + y_2 + y_3 + \cdots + y_{15} + y_{16} + y_{17}}{17}$$

and

$$\hat{y} = \frac{y_1 + 3y_2 + 3y_3 + \cdots + 3y_{15} + 3y_{16} + y_{17}}{47}.$$

(a) Show that BOTH $\overline{y}$ and $\hat{y}$ are each unbiased estimators of μ. This shows that unbiased estimators are not unique!

(b) It is also known from previous pizza experiments that $\mathrm{Var}(y_i) = 2$, $i = 1, \ldots, 17$. Compute both $\mathrm{Var}(\hat{y})$ and $\mathrm{Var}(\overline{y})$ exactly stating any assumptions that you have made, and therefore show that in this case $\mathrm{Var}(\hat{y}) \geq \mathrm{Var}(\overline{y})$.

6. (*26 points*) After hearing of the nutritionist's experiment in Problem **??**, a sociologist at the University of Saskatchewan decided to find out if a drug he invented, called *Pizza-X*, could help university students to eat more pizza. He decided to test his drug on his Sociology 101 class. Fortunately, all 100 students in the class agreed in advance to participate if selected for the experiment. Using a table of random digits, he randomly selected 10 students, and allocated 5 to the control group and 5 to the injection group. All participants were instructed not to eat after 10 a.m. and were then fed as many standard-sized cheese pizza slices during lunchtime as could be eaten. The results obtained are listed below:

control group	injection group
11	15
12	11
11	14
7	9
9	11

(a) Let $\overline{y}_1$ denote the mean number of pizza slices eaten by the control group, and let $\overline{y}_2$ denote the mean number of pizza slices eaten by the injection group. Compute $\overline{y} = \overline{y}_2 - \overline{y}_1$.

(b) Compute the estimated variance $\hat{V}(\overline{y})$.

(c) Using your answers to **(a)** and **(b)**, construct an approximate 95% confidence interval for μ, the true mean difference in number of pizza slices eaten between the control group and the injection group in Sociology 101.

(d) Based on your answer to **(c)**, is there sufficient evidence to conclude that the sociologist's drug *Pizza-X* helps his Sociology 101 students eat more pizza? Why or why not?

7. (*14 points*) A sociologist at the University of Regina is interested in the *total* number of families that rent their homes in the City of Moose Jaw. She decides to conduct a 1-in-50 systematic survey drawing upon city tax records. It is known that there are 15 200 households in Moose Jaw. The sociologist found that

$$\sum_{i=1}^{304} y_i = 76$$

where $y_i = 1$ if the family in the i^{th} household sampled rents and $y_i = 0$ if the i^{th} family does not. Using the data given, construct an approximate 95% confidence interval for τ, the total number of families who rent.

(*Hint:* If $\hat{p}$ denotes the estimated *proportion* that rent, then $\hat{\tau} = N\hat{p}$.)

8. (*20 points*) The City of Regina chief engineer is considering a zoning change for the Wascana Park subdivision to allow a new shopping complex to be built. In order to assess the opinion of residents, a cluster sample is used. The subdivision map is marked into 170 blocks, and a random sample of 15 blocks is surveyed. For each of these blocks (labelled i, for $i = 1, \ldots 15$), the number of adult residents (m_i) and the number of adult residents opposing the zoning change (y_i) are recorded. The data are summarized below:

$$\sum_{i=1}^{15} m_i = 546, \quad \sum_{i=1}^{15} m_i^2 = 9981, \quad \sum_{i=1}^{15} m_i y_i = 4035, \quad \sum_{i=1}^{15} y_i = 182, \quad \sum_{i=1}^{15} y_i^2 = 1819.$$

(a) Compute $\overline{y}$ and $\hat{V}(\overline{y})$, and use them to construct an approximate 95% confidence interval for the proportion of Wascana Park adult residents opposed to the zoning change.

(b) Suppose that the chief engineer wanted a bound on the error of estimation of μ to be $B = 0.04$. What sample size should the chief use?

9. (*24 points*) A psychologist at the University of Regina is interested in educational reform and wants to obtain information on the amount of time spent on non-academic activities such as sports and video games by elementary school children. She uses stratified sampling as indicated in the table below, collecting information on y, the percentage of time spent on non-academic activities.

STRATA	strata size	sample size	sample mean	sample variance
Separate Schools	24	6	13	9
Public Schools	54	12	26	16

(a) Find an approximate 95% confidence interval for the population mean μ.

(*continued*)

Suppose that the psychologist had wanted a bound on the error of estimation for μ of $B = 1.5$. What sample size should she have used and how should it have been allocated for each of the allocation methods given in **(b)** and **(c)** below?

(b) *proportional allocation:*

(c) *Neyman allocation:*

10. (*24 points*) Each person in a population of adults is interviewed, an identification number is assigned to each one, and the sex and age of each individual is recorded. The population size is 1000 of which 550 are female and 450 are male. The population mean age is 45.

From the list of identification numbers a simple random sample of size 50 is obtained and the following measurements are recorded: x_i is the age (in years) and y_i is the systolic blood pressure (in mg/100ml) of the i^{th} subject.

The resulting data are

$$\sum_{i=1}^{50} x_i = 2248, \quad \sum_{i=1}^{50} x_i^2 = 104384, \quad \sum_{i=1}^{50} y_i = 6744, \quad \sum_{i=1}^{50} y_i^2 = 928436, \quad \sum_{i=1}^{50} x_i y_i = 305125.$$

Find an approximate 95% confidence interval for the population mean of y using the estimation methods given in **(a)** and **(b)** below.

(a) *ratio estimation:*

(*continued*)

(b) *regression estimation:*

(c) Compute the relative efficiency of these two estimators. Is there sufficient evidence to favour one over the other? Why or why not?

11. (*18 points*) A senior administrator at the University of Regina wishes to estimate the proportion of its students that have used cocaine, a sensitive subject. Students were classified into one of two strata–undergraduate and graduate–and were randomly sampled within the stratum. Since there was some concern that students might be unwilling to disclose their use of cocaine to a university official, the following *random response method* was used. The university official constructs a deck of 30 cards. On 26 of them are marked **N** for *never used cocaine* and 4 of them are marked **C** for *have used cocaine at least once*. Each sampled student was asked to draw a card from the deck and to respond *yes* if the letter agrees with the group that student belongs to. The results are as follows:

STRATA	total number of students	number sampled	number answering yes
undergraduate	8972	900	123
graduate	1548	150	27

(a) Construct a 95% confidence interval for the proportion of undergraduates who have used cocaine at least once.

(b) Construct a 95% confidence interval for the proportion of graduates who have used cocaine at least once.

(c) Based on your answers to **(a)** and **(b)**, is there a statistically significant difference in the use of cocaine between undergraduates and graduates? Is this surprising? Why or why not?

12. (*18 points*) Suppose that the City of Regina wants to estimate how many homeowners band their elm trees to prevent *Dutch Elm Disease*. A sample is to be selected using one of the following: simple random sampling, systematic sampling, cluster sampling. The method of data collection could be one of: personal interview, telephone survey, mailed questionnaire, direct observation. Choose both a sampling scheme and a method of data collection, and discuss how you might actually conduct the survey using your two choices.

(*In answering this question, you might want to draw on many of the ideas discussed throughout the course. If there are any benefits or limitations to your plan, you should mention them as well.*)

(*The end. Happy Holidays.*)