FINAL EXAM
STAT 5201
Spring 2008

Due in Room 313 Ford Hall
Friday, Friday May 19 at 4:00 PM
Please deliver to the office staff
of the School of Statistics

READ BEFORE STARTING

You must work alone and may discuss these questions only with Glen Meeden. You may use the class notes, the text and any other sources of printed material.

Put each answer on a single sheet of paper. You may use both sides. Number the question and put your name on each sheet.

You may email me with any questions. If I discover a misprint or error in a question I will post a correction on the class web page. In case you think you have found an error you should check the class home page before emailing me since it may already be fixed.

1. Find a recent survey reported in a newspaper, magazine or on the web. Briefly describe the survey. What are the target population and sampled population? What conclusions are drawn from the survey in the article. Do you think these conclusions are justified? What are the possible sources of bias in the survey? Please be brief.

2. Problem 6 on page 121 of the text.

3. Problem 21 parts a) and b) on page 128 of the text.

The answers for problems 2 and 3 are in ??

4. From a population of size $N = 2345$ a random sample of size 100 was taken and the variable of interest y was observed along with the auxiliary variable x. The data can be found at

`http://www.stat.umhtn.edu/~glen/classes/5201/moredata/sampRR1.txt`

Note you do not need our password and username to access the data. The population total for the auxiliary variable is $t_x = 175644540$ Using the data calculate the usual unbiased estimate of the population total for y along with the ratio and the regression estimates. For each find their standard error. What estimate would you recommend in this case. Briefly justify your choice.

One easy way to do the calculations for the ratio estimator is to modify the code in the **Calculation the ratio estimator for a population total** handout. Replace the function *ratiotot* with the following function.

```
ratiototsmp<-function(ysamp,xsamp,N,tx)
{
        n <- length(ysamp)
        ff<-n/N
        rhat <- sum(ysamp)/sum(xsamp)
        esttot <- rhat * tx
        dum1<-(N*N*(1-ff))/(n*(n-1))
        vartot <- dum1*sum((ysamp-rhat*xsamp)^2)
        ans<-c(esttot,vartot)
        return(ans)
}
```

ans

```
mean(ysamp)= 8828.16
> N*mean(ysamp)= 20,702,035
N^2*(1-100/N)*var(ysamp)/100=1.845622e+13
sqrt(N^2*(1-100/N)*var(ysamp)/100)= 4,296,070



ratiototsmp(ysamp,xsamp,N,tx)
[1]     19,385,792 574280947478 sqrt(574280947478)= 757,813.3
> regtotsmp(ysamp,xsamp,N,tx)
regtotsmp(ysamp,xsamp,N,tx)
[1]     1,9485,826 457097758344 sqrt(457097758344)= 676,090
```

5. From a population with 98 cluster of various sizes 20 clusters were select at random. The sample sizes within the selected clusters also vary. The y values and along with their cluster identifiers and cluster sizes can be found at

`http://www.stat.umn.edu/~glen/classes/5201/moredata/sampclus1.txt`

Using the code in the handout **Working with cluster samples** find the usual unbiased estimate of the population total and the ratio estimate for the population mean along with their standard errors.

If the total number of units in the population is know to be 1612 find the usual unbiased estimate of the population mean and its standard error.

ans Using the R function and remembering to take the square root of the variance our estimate for the pop total is 34,289.86 with standard error 2,818.5 and the ratio estimate of the population mean is 22.216 with standard error 1.133. The usual estimate of the mean is 21.27 with standard error 1.748.

6. The Beefy Burger Chain (BBC) is interested in the proportion of patrons over the age 30 making purchases in their restaurants. Restaurants vary in weekly sales volume which is closely related to the number of customers. The company keeps records of daily sales volume. We have a list of all the restaurants in the country along with their average sales volume for the past year. Design a survey which will be efficient for this purpose and briefly explain why. Show the algebraic equations for your estimate and its standard error.

ans Here restaurants are clusters or PSU's and SSU's are customers. Form strata based on regions of the country and within regions form more strata based on average daily sales volume. Select stores at random from within strata with the number of stores selected proportional to the total sales for a stratum. Then "randomly" select customers using a systematic sample like every 20th customer and ask if they are over 30 or not.

7. When a systematic sample is used it is often recommended, when getting an estimate of the variance of the sample mean, to assume that the sample was selected as a simple random sample without replacement. It was noted in class that in some cases this will be fine but in other cases this can either over estimate or under estimate the true variance of the systematic sample.

Construct three populations to demonstrate that the above can happen in practice. That is one where the assumption is approximately true, another where we get an under estimate and another where we get an over estimate.

Let N be the population size and n be the sample size and suppose $k = N/n$ is an integer. Then k is the number of possible systematic samples that can be drawn for the population. The handout **variance of a systematic sample** contains the function *varsystematic* which computes the true variance of the sample mean under the systematic sampling design and under simple random sampling without replacement.

ans Take a random sample for any distribution, for example let $y1$¡-rgamma(500,5) then the two variances should be roughly the same since the labels and the y values will be uncorrelated. Next let $y2$ just be $y1$ order from smallest to largest. Then the variance of a systematic sample will be smaller since any systematic sample will be quite representative of the population.

To get the opposite do the following. Let w¡-c(10,20,....,100) and then form a new vector, say ww, by repeating w 20 times. Then let y¡-rnorm(200,ww,1). This gives a population where no matter where you start in the first 10 units every tenth unit will be very similar but quite different from any other systematic sample of size 20.

8. In class I talked about the model based approach to the ratio estimate. Assuming the model is true an estimate of its variance for estimating the population total of y which does not depend on the sampling design is

$$\frac{N^2}{n(n-1)}(1-n/N)\sum_{i\in smp}(y_i-\hat{R}x_i)^2/x_i\frac{\bar{x}_{nsmp}}{\bar{x}_{smp}}\bar{X}$$

where

$$\hat{R} = \frac{\sum_{i \in smp} y_i}{\sum_{i \in smp} x_i}, \quad \bar{x}_{smp} = \sum_{i \in smp} x_i/n, \quad \bar{x}_{nsmp} = \sum_{i \notin smp} x_i/(N-n), \quad \bar{X} = \sum_{i=1}^{N} x_i/N,$$

and n and N are the sample and population sizes. Construct two populations where the ratio estimator is appropriate. Empirically compare the two estimators of variance by considering many samples from each of the two populations. Briefly summarize what you learned in a paragraph or two.

The handout **variance estimation for the ratio estimator** on the class web page contains some code which you could use in your simulations. For R samples the function *compratiolp* calculates the ratio estimate, its absolute error and the two estimates of variance. For each sample it also calculates the sample mean of the x values. The output is a $R \times 5$ matrix where each row, for a given sample, contains the ratio estimate, its absolute error, the usual estimate of variance, the model estimate of variance and the sample mean of the x values. The matrix has been ordered so that the sample with the smallest sample mean of the x values is in the first row. The second row contains the results for the second smallest sample mean of the x values and so on.

ans The model variance tends to be a bit longer for samples with smaller $\bar{x}_{smp}$'s and a bit shorter for samples with the larger $\bar{x}_{smp}$'s.