FINAL EXAM
STAT 5201
Spring 2010

Due in Room 313 Ford Hall
Friday May 14 at 3:45 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.

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. A random sample of size 22 was taken from a population of 500 individuals and the value of a characteristic y was observed. The data are given below.

```
4.77  4.18  6.47  3.14  4.65  3.54  8.73  1.82  3.29  4.25 10.09  2.15
3.27  3.68  5.42  7.79  6.33  5.30  3.55 10.96  9.15  7.07
```

i) Given the usual estimate of the population total for y along with its standard error.

ii) The individuals in the first row of the data are all men while those in the second row are all women. Estimate the total of y for the population of men and give the standard error of your estimate.

iii) Estimate the mean of y for the population of men and give the standard error of your estimate.

Answer

i) The estimate is

$$500\bar{y}_{smp} = 500 \times 5.44 = 2718.2$$

with standard error

$$\sqrt{500^2(1 - 22/500)V(y_{smp})/22} = 268.8$$

ii) This is a domain estimation problem. To estimate the domain total we just set every y value in the second row to 0 and repeat what we did in part i). The estimate is 1297.27 with standard error 315.9.

iii) To estimate the domain mean we condition on the fact that $n_D = 12$. The estimate is 4.76, the mean of the men in the sample. Since N_d, the number of men in the population is unknown we use n/N as an estiate of n_d/N_d. The resulting estimate of the standard error is

$$\sqrt{(1 - 22/500)(1/12)var(y_{smp \bigcap D})} = 0.709$$

3. Consider the following table of sums of weights from a sample; each entry in the table is the sum of sampling weights for persons in the sample falling in that classification (for example, the sum of the sampling weights for the number of women between the ages of 20 and 29 is 150.

	Age				Sum of
	20-29	30-39	40-49	50-59	weights
Male	100	450	350	100	1200
Female	150	400	450	200	1000
Sum of weights	300	800	900	200	

i) Assume it is known the that the population contains 1200 men and 1000 women and the 300 persons between the ages of 20-29, 800 between 30-39, 900 between 40-49 and 200 between 50-59. Readjust the cells weights so that in the new table the marginal weights agree with the known population weights.

ii) In the first part you will find that you only needed 3 passes through the table to get agreement up 2 decimal place accuracy. Give an example of a table where there are 2 rows and 3 columns and you will not achieve 2 decimal place accuracy even after 5 passes through. Note all the entries in your table should be at least as large as one.

Answer

i) Using the standard raking adjustment I found

```
First time through yields
       [,1]   [,2]   [,3]   [,4]
[1,] 146.94 494.66 475.47  83.72
[2,] 153.06 305.34 424.53 116.28

After the second time trough the first row is
over by 0.01 and the second row is under by .01

       [,1]   [,2]   [,3]   [,4]
[1,] 146.83 494.38 475.15  83.65
[2,] 153.17 305.62 424.85 116.35

The third time trough makes the rows ok too up
to 2 decimal plases
       [,1]   [,2]   [,3]   [,4]
[1,] 146.83 494.38 475.14  83.65
[2,] 153.17 305.62 424.86 116.35
```

ii) Consider the 2 by 3 table where the first row is 95, 1 and 1 and the second row is 1, 1 and 1. Use the constraints that the first row must add to 20 and the second to 80 and the constraints that the first column must add to 10, the second to 10 and the third to 80.

```
At 5 times through we get the table

  9.21  1.1  8.76
  0.79  8.9 71.24

where the first row is under by -0.93 and
the second is over by 0.93

To get 2 decimal accuracy you need to take
11 passes through to get

 9.28 1.19  9.53
 0.72 8.81 70.47
```

Clearly this is a rather silly example. But it demonstrates that if the cell weights are to far off then raking will take longer to converge.

4. A simple random sample without replacement was used to select 6 clusters from a population of 50 clusters. From each selected cluster simple random sampling without replacement was used to select a sample of four secondary sampling units.

The data are located at

`http://www.stat.umn.edu/~glen/classes/5201/moredata/clsmpf10.txt`

Note you do not need our password and username to access the data.

For these data find an estimate of the population total and the population mean along with their estimated variances.

Answer

Using the handout "Working with cluster samples" which is available on the course web site and setting $N = 50$ we find

$$\hat{t}_y = 18381.08, \quad V(\hat{t}_y)\widehat{=}5837895, \quad yratio = 19.87, \quad V(yratio)\widehat{=}1.70$$

5. A manager of a software firm was interested in estimating the total number of hours they spent on the telephone answering questions of customers who own their product. Develop a sampling plan to help her find an answer to her question under two different scenarios.

i) The manager has a list of every phone call received by their help line during the last year and how long it lasted.

ii) The company sells two products say A and B with about twice as many A's being sold as B's. Again the manager has a list of every call. But this time for a given selected call it is easy to determine if the customer's question was about product A or B but it is more work to find out how long the call lasted. Moreover the manager believes that calls about product B tend to last long than those for product A.

Answer

In part i) we would just use a srs without replacement. In part ii) we should use post-stratification. Take a large first sample and determine for each selected call whether the question was about product A or B. Then take a further subsample. The manger might have some more information to help select the allocation between type A calls and type B calls at the second stage.

6. Construct a population where under simple random sampling the ratio estimator has a smaller sample variance (by at least 20%) than the sample mean.

Answer

We know from class that the ratio estimator works well under the model

$$y_i = \beta x_i + \epsilon_i$$

where the ϵ_i's are independent random variables with $E(\epsilon_i) = 0$ and $V(\epsilon_i) = \sigma^2 x_i$. In R I did

```
> x<-rgamma(1000,5)+50
> y<-rnorm(1000,10*x,x)
```

I then found the variances to be 54.06 and 64.37 respectively with $cor(x, y) = 0.40$. If I replace 10 by 20 in the above I get the variances to be 57.26 and 93.20 with $cor(x, y) = 0.62$.

7. A company ships out its product in batches of 200 units. Each unit in a batch is graded as either of average quality or of excellent quality. Batches are classified into 3 types say I, II and III.

The company promises that in Type I batches about 50% of the units will be of excellent quality while this percentage is 75% in Type II batches and 90% in type III batches. On the average 40% of their batches are of Type I, another 40% are Type II and the remaining 20% are Type III. When getting ready to ship out a particular batch they noticed that its label is missing. To try to discover its type they took a random sample of size 20 units from the batch and observe that 16 of the units were of excellent quality.

i) Given the information in the sample find the posterior probabilities that it is either Type I, II or III.

ii) Use this information to simulate 500 possible copies of the units in the batch. For each simulated copy find the total number of excellent units and make an histogram of these values. Is the shape of the histogram what you expected? Briefly explain.

Answer

i) Let

- p_i = proportion of excellent units in a batch of type i.
- π_i = the prior probability that a batch is of type i

In this problem the p_i's are 0.50, 0.75 and 0.90 respectively and the π_I's are 0.40, 0.40 and 0.20 respectively. Let $y = (y_1, \ldots, y_{200})$ where $y_i = 1$ if the i unit is excellent and $y_i = 0$ when it is average. In this case we have

$$p(y) = \sum_{i=1}^{3} \pi_i \, p_i^{\sum_1^N y_j} \, (1 - p_i)^{N - \sum_j^{200}}$$

and the posterior probability that the batch is of type i is

$$p(i \mid y_{smp}) = \frac{\pi_i \, p_i^{16} \, (1 - p_i)^4}{\sum_{j=1}^{3} \pi_j \, p_j^{16} \, (1 - p_j)^4}$$

We find that the these posterior probabilities are 0.019, 0.793 and 0.188.

ii) The following R code finds the posterior probabilities and lets you simulate the total number of excellent units in the batch in question.

```
simpoptot<-function(prior,pp,y,n,N,R)
  {
    k<-length(pp)
    dd1<-rep(0,k)
    for(i in 1:k){
      dd1[i]<-(pp[i]^y)*(1-pp[i])^(n-y)
    }
    dum<-sum(prior*dd1)
    post<-(prior*dd1)/dum
    cat("post=",post,"\n")
    ty<-rep(0,R)
    for(i in 1:R){
      urn<-sample(1:k,1,prob=post)
      ty[i]<-y + rbinom(1,N-n,pp[urn])
    }
    return(ty)
   }
```

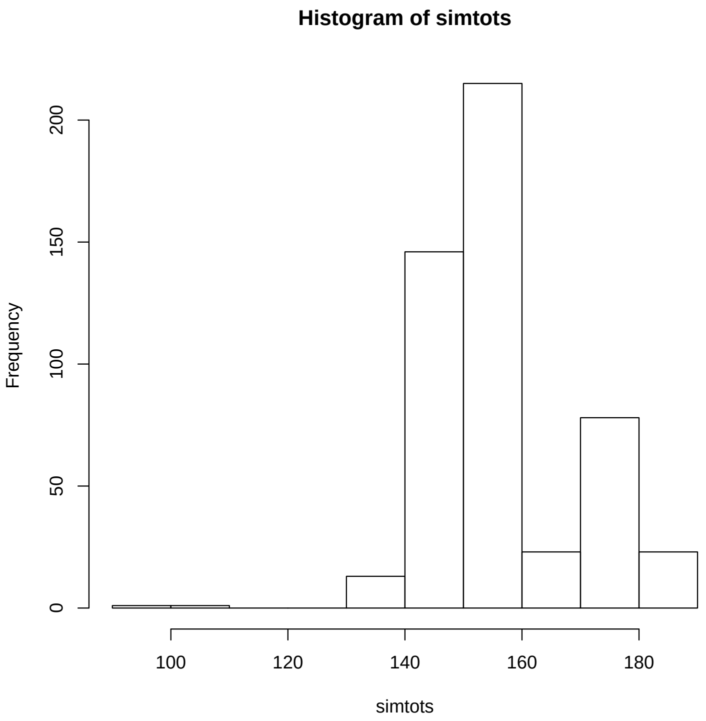

Figure 1: The histogram for problem 7

In our problem *prior* is c(0.4,0.4,0.2), *pp* is c(0.50,0.75,0.90), $y = 16$, $N = 200$ and $R = 500$.

Note most of the values in the histogram relfect the fact that the posterior probability of the batch being type II is 0.793.