FINAL EXAM
STAT 5201
Spring 2011

Due in Room 313 Ford Hall
Friday May 13 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.

The high score on the exam was 93. There were 5 in the 80's, 9 in the 70's, 7 in the 60's, 9 in the 50's, 3 in the 40's, 2 in the 30's and the low score was 27.

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. Suppose we wish to take a stratified random sample from a population where the following table gives the the size, a good guess for the variance and the cost per unit sampled for each stratum.

stratum	size	variance	cost
1	500	40	5
2	1000	30	3
3	2000	20	4
4	3000	15	5

Find the optimal allocation if we desire the variance of our estimator to 0.5. Repeat if we want the variance to be 0.1. In each case give the total cost of the sampling plan.

Ans This is discussed in the text in section 3.4.2. I wrote a simple program in R to do my calculations. For the 0.5 case the optimal allocation is 4, 9, 13 and 15 with a total cost of 172. For the 0.1 case the answer is 20, 44, 62 and 72 for a total cost of 837.

```
findcost<-function(N,v,c,vest)
  {
    w<-N/sum(N)
    dum1<-sum(w*sqrt(v)*sqrt(c))
    dum2<- sum((w^2)*v/N)
    lam<-dum1/(vest + dum2)
    cost<-(dum1*dum1)/(vest + dum2)
    n<-lam*w*sqrt(v)/sqrt(c)
    ans<-c(cost,sum(n),n)
    return(ans)
  }
```

In the code N is the vector of stata sizes, v is the vector of strata variances, c is the vector of strata costs and *vest* is the desired variance.

3. At a trade show all attendees who registered received a bag full of free products from the companies sponsoring the show. The list of all those who registered along with their contact information was made available to the companies. One company, included a new product which was under development in the bag. Attached to their product was a card which asked the user to rate the product and send the card back to the company. If they did the company promised to send them another sample of the product.

Two months later the company had received 100 cards from the 2,000 attendees 80 of which strongly endorsed the product. The sales manger upon seeing these results argued that the new product should be put into production since it was likely that 80% of possible customers would use the product. Do you think the manger's enthusiasm was justified? If so please explain why. If not, suggest a possible sampling plan to learn more about the attitude to possible customers for the new product.

Ans Assuming that the attendees are typical of the set of all possible users of the product giving out a free sample was not a bad idea. However the response rate is very low and the promise of another sample biases the responders towards people who like the product. To get more and

better information the company should take a random sample from the 1900 attendees who have not responded. This is possible since they have contact information for them.

4. A company owns a fleet of 900 trucks and is interested is estimating the total number of times the trucks had to be serviced during the last six months. They took a random sample of 50 trucks and found the following

	Number of visits to shop					
	0	1	2	3	4	5
# of trucks	31	7	5	4	1	2

a) Given this sample find a 95% confidence interval for the total number of times the trucks needed to be serviced.

b) An employee checked the records for each truck and determined that 275 were serviced at least once. Use this information and the sample above to find another 95% confidence interval for the total number of times the trucks needed to be serviced.

c) Which interval would you prefer. Briefly justify your answer.

ANS a) Note

$$\hat{Y} = N \times \bar{y}_{smp} = 900 \times 0.86 = 774$$

$$\begin{aligned} est\ stdev &= 900\sqrt{1 - 50/900}(\sqrt{var(y_{smp})}/\sqrt{50} \\ &= 900(0.9718253)(1.370387/7.071068) \\ &= 169.5 \end{aligned}$$

from which, assuming normality, we can find the CI.

b) Let D be the Domain of trucks who were serviced at least once. Now $N_d = 275$ is known

$$\hat{Y} = \hat{Y}_d = N_d \times \bar{y}_{dsmp} = 275 \times 2.26 = 622$$

where $\bar{y}_{dsmp}$ is the mean of the the units in the sample that belong to the Domain.

$$\begin{aligned} est\ stdev &= 275\sqrt{1 - 19/275}(\sqrt{var(y_{smpd})}/\sqrt{19} \\ &= 275(0.9648363)(1.326738/4.358899) \\ &= 80.76 \end{aligned}$$

from which again, assuming normality, we can find the CI.

c) We would expect b) to be the better estimate since it is using more information.

5. People who are in Medicare can have a personal care assistant (PCA) come to their home to help them with their health problems. A PCA can be a nurse who preforms a specific therapy for the patient. In other cases the PCA can have less medical training and perform other tasks. For example they can prepare healthy meals for the patient or help them bath. At the end of a visit the PCA fills out a time sheet indicating the number of hours they spent with the patient which is signed by both the PCA and the patient. The PCA returns the sheet to the company employing them. In addition the patient sends a copy of the time sheet to a government agency. Periodically the company turns in the time sheets for all their employs and gets reimbursed by Medicare.

The government has some reason to believe that a certain company, say C, has been charging Medicare for services that were never preformed. This is a large company who employs many PCA's

both within the Twin Cities area and also in rural areas of the state. The average number of hours per week that an Individual PCA works varies as well. Devise a sampling plan that would help the government decide if C is submitting fraudulent claims and if they are how much of a refund the government should request. Be specific in how you would estimate the amount billed in bad claims.

ANS. You should consider the Twin Cities and the rural areas as two different strata. Within a stratum you can stratify on Nurse PCA's and non-Nurse PCA's. Assuming that the records for a particular PCA are together we can consider them to form a cluster. So within the strata we could sample clusters proportional to the size of the cluster. After we have collected our sample from the company we then need to go to the government to find the corresponding patient records for the selected billings. Once this information has been collected for each claim we find the difference between the amount billed by the PCA and the correct amount from the patient. Let y_i denote this amount. Then we wish to estimate the population total of the y_i's,

6. In R using the following commands

```
set.seed(11228899)
popx<-rgamma(500,25)
popy1<- rnorm(500,16*popx,1*popx)
popy2<-rnorm(500,6*popx,2*popx)
x<-popx
```

I created two populations where x is the auxiliary variable there are two versions of the y variable of interest.

a) Consider three sampling plans: SRS without replacement, pps proportional to x and pps proportional to $1/x$. For the problem of estimating the population total for popy1 consider the three estimators: the population size times the sample mean, the ratio estimator and the Horvitz-Thompson estimator. For a sample size of $n = 20$ find the true variance of the sample mean and the usual expression for the variance of the ratio estimator. For each sampling plan collect 1,000 samples of size 20. For each estimator and each sampling plan find the average value of the estimator of the population total, its average absolute error, the length of the 95% CI, the frequency with which it contained the true population total and the average estimated variance of estimator. Repeated the above when $y = popoy1$ is replaced by $y = popy1 + 1,000$ Use these simulation results to briefly explain and justify what you learned about the proper use of these three estimators in class

b. On the class web page in the list of Some Rweb handouts there is one entitled "Variance estimation for the ratio estimator". In it there is the function "ratiototboth" which for a given sample, smp, and popx and popy finds the variance of the ratio estimator under simple random sampling and the model based estimate of its variance given in class. For the two populations given by x with either popy1 or popy2 and for the two sampling plans pps proportional to x and pps proportional to $1/x$ generate 1,000 samples and find the average value of the two different estimates of the variance of the ratio estimator. Are the results what you expected? Briefly explain.

ANS The population total for y is 200680.5. The true variance of the sample mean under SRS for both y and $y + 1000$ is 83210579. Under SRS and y and x the true variance of ratio est is 7227162. The simulation results given below support the following given in class.

- Under SRS the sample mean is unbiased but will preform poorly for the other designs.
- Under SRS the sample mean and the HT estimator are the same.
- The HT estimator should work well for y under pps(x) but will preform poorly under pps($1/x$).

- The ratio estimator works well when $y \propto x$ and its performance does not seem to depend so strongly on the design although its poorest performance is under pps($1/x$).
- Both the ratio estimator and the HT estimator preform poorly in the $y + 1000$ case even though the HT is always unbiased.

```
                Results for y and SRS
est       Ave val  Ave |err|  len of CI  F of Cov   Ave var
ansybar  200556.3 7414.21    35266.93   0.92       82925934
ansratio 200624.7 2116.73    10301.72   0.94        7099095
ansht    200556.3 7414.21    35994.16   0.93       86381181
                Results for y and pps(x)
est       Ave val  Ave |err|  len of CI  F of Cov   Ave var
ansybar  208595.3 10014.78   35767.95   0.85       85155899
ansratio 200762.0  2211.55   10739.43   0.93        7719967
ansht    200608.3  2187.78   10519.53   0.92        7391970
                Results for y and pps(1/x)
est       Ave val  Ave |err|  len of CI  F of Cov   Ave var
ansybar  193157.2  9669.56   34869.41   0.82       81275586
ansratio 200584.1  2197.22   10065.47   0.90        6782293
ansht    201298.6 14247.35   71076.94   0.92      340889715

                Results for y+1000 and SRS
est       Ave val  Ave |err|  len of CI  F of Cov   Ave var
ansybar  701056.0  7229.13   35542.44   0.94       84262331
ansratio 700657.8 17192.85   82968.05   0.94      457356514
ansht    701056.0  7229.13   36275.35   0.94       87773261
                Results for y+1000 and pps(x)
est       Ave val  Ave |err|  len of CI  F of Cov   Ave var
ansybar  707994.6  9629.06   35373.60   0.86       83326439
ansratio 683811.7 21835.73   80019.48   0.86      425625851
ansht    700975.7 17622.00   85981.65   0.93      499091614
                Results for y+1000 and pps(1/x)
est       Ave val  Ave |err|  len of CI  F of Cov   Ave var
ansybar  692901.1  9803.03    34708.05  0.81       80533717
ansratio 721333.6 25035.49    84920.58  0.83      480048344
ansht    700997.9 32070.23   156574.11  0.93     1643293665
```

b) From the theory we would expected the model based estimate of variance to be smaller under pps(x) than under pps(1/x). This is true but the difference is quite small. Note that the average of the model based estimate of variance is much less sensitive to the design than the average of the usual estimate of variance. This suggests that the usual estimate of variance is biased downwards for samples with $\bar{x}_{smp} < \bar{x}$ and biased upwards for samples with $\bar{x}_{smp} < \bar{x}$.

```
       Results for popy1 and pps(x) followed by pps(1/x)
est          Ave |err|  Std ave var   Model ave var
200637.58    2143.27    7750240.87    7058512.36
200528.39    2065.72    6721953.59    7070468.28
       Results for popy2 and pps(x) followed by pps(1/x)
est          Ave |err|  Std ave var   Model ave var
74207.82     4602.82 35582840.70 32305331.91
73461.98     4664.33 30756811.65 32435700.03
```

Here is my R code for solving part b.

```
ratiototboth<-function(smp,popy,popx)
{
        n <- length(smp)
        N<-length(popx)
        ff<-n/N
        ysamp<-popy[smp]
        xsamp<-popx[smp]
        xnsamp<-popx[-smp]
        tx<-sum(popx)
        trtot<-sum(popy)
        rhat <- sum(ysamp)/sum(xsamp)
        esttot <- rhat * tx
        err<-abs(esttot - trtot)
        dum1<-(N*N*(1-ff))/(n*(n-1))
        usualvartot <- dum1*sum((ysamp-rhat*xsamp)^2)
        usualans<-c(esttot,err,usualvartot)
        dum2<-sum(((ysamp -rhat*xsamp)^2/xsamp))*((mean(xnsamp)
              *mean(popx))/mean(xsamp))
        modelvartot<-dum1*dum2
        ans<-c(esttot,err,usualvartot,modelvartot)
         return(ans)
}

 compare1lp<-function(popy,popx,n,design,R)
{
    ans<-rep(0,4)
     N<-length(popy)
    for(i in 1:R){
       smp<-sample(1:N,n,prob=design)
       ans<-ans+ ratiototboth(smp,popy,popx)
     }
     ans<-round(ans/R,digits=2)
     return(ans)
}
```

7. Consider the following experiment involving two urns I and II from which we will be doing srs with replacement. Each urn contains balls labeled 0 and 1. Let $P(1 \mid I)$ and $P(1 \mid II)$ be the probability that a ball selected from the respective urns is one. Let $\Theta = \{\theta_1, \ldots, \theta_k\}$ be a set of probabilities, i.e. for each i we have $0 < \theta_i < 1$. The experiment proceeds as follows.

- A value of θ_i is selected at random from Θ using the probability function $f(\cdot)$.
- Given θ_i urn I is selected with probability θ_i and urn II is selected with probability $1 - \theta_i$ and then a ball is chosen at random from the selected urn. Its value is noted and then it is returned to the urn.
- With the same θ_i the previous step is repeated getting a second ball from the newly selected urn.

For $i = 1$ and 2 let $u_i = I$ or II depending on which urn is selected. Let $w_1 = 1$ and $w_2 = 1$ denote the events that the first and second balls selected were both a one. With this notation we can now write

$$P(\theta_i, u_1 = I, w_1 = 1, u_2 = II, w_2 = 1) = f(\theta_i)\theta_i P(1 \mid I)(1 - \theta_i)P(1 \mid II)$$

a) Find an expression for $P(w_1 = 1)$.
b) Show that

$$P(w_2 = 1 \mid w_1 = 1) = \sum_{i=1}^{k} P(\theta_i \mid w_1 = 1)P(w_2 = 1 \mid \theta_i)$$

Ans a) Note

$$\begin{aligned} P(w_1 = 1) &= \sum_{i-1}^{k} f(\theta_i)\Big\{\theta_i P(1 \mid I) + (1 - \theta_i)P(1 \mid II)\Big\} \\ &= \sum_{i-1}^{k} f(\theta_i)P(w_1 = 1 \mid \theta_i) \end{aligned}$$

b) Note $P(w_1 = 1, w_2 = 1 \mid \theta_i) = P(w_1 = 1 \mid \theta_i)P(w_2 = 1 \mid \theta_i)$ and so we have

$$\begin{aligned} P(w_2 = 1 \mid w_1 = 1) &= P(w_1 = 1, w_2 = 1)/P(w_1 = 1) \\ &= \left\{ \sum_{i-1}^{k} f(\theta_i)P(w_1 = 1.w_2 = 1 \mid \theta_i) \right\} /P(w_1 = 1) \\ &= \left\{ \sum_{i-1}^{k} f(\theta_i)P(w_1 = 1 \mid \theta_i)P(w_2 = 1 \mid \theta_i) \right\} /P(w_1 = 1) \\ &= \sum_{i-1}^{k} \frac{f(\theta_i)p(w_1 = 1 \mid \theta_i)}{P(w_1 = 1)} P(w_2 = 1 \mid \theta_i) \\ &= \sum_{i-1}^{k} P(\theta_i \mid w_1 = 1)P(w_2 = 1 \mid \theta_i) \end{aligned}$$