

FINAL EXAM  
STAT 5201  
Spring 2013

Due in Room 313 Ford Hall  
on Friday May 17 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 exam was taken by 54 people. There were 5 scores in the 90's, 2 in the 80's, 7 in the 70's, 11 in the 60's, 13 in the 50's, 11 in the 40's, 1 in the 30's, 2 in the 20's and 2 in the 10's.

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. The Department of Revenue of the state of Minnesota is interested in determining if a large corporation has paid the correct amount of sales tax to the state for its transactions for the past year. Since the number of transactions can be quite large one can only audit a small sample from the population of all transactions. Assume that one has a list of all the corporation's transactions identified by the amount of the transaction. Explain how you could use this information to develop a sensible sampling plan.

**ANS** You could stratify on the size of transaction and then use pps based on the amount of the transactions within strata. This allows you to sample a larger proportion of the bigger sales but still checks on some of the smaller sales.

3. The ratio estimator for a population total makes sense under the following model for the population

$$y_i = \beta x_i + z_i$$

where  $\beta$  is an unknown constant and the  $z_i$ 's are independent normal random variables with mean 0 and the variance of  $z_i$  is equal to some constant times  $x_i$ . In  $R$  generate such a population and take two samples from it under two different sampling designs using the following code.

```
set.seed(13449875)
popx<-rgamma(500,10) +20
popy<-rnorm(500,3.6*popx,.25*popx)
smp1<-sample(1:500,20,prob=popx)
smp2<-sample(1:500,20,prob=1/popx)
```

For the two samples find the ratio estimate of the population total and the usual design based estimate of their variances which assumes that the sampling design is simple random sampling without replacement. Also find the model based estimates of variance for the ratio estimator. For each sample note which one is larger and explain why this happens.

**ANS** On the class web page the Rweb handout entitled "Variance estimation for the ratio estimator" contains the function "ratiototboth" which does the necessary calculations with the function "ratiototboth". This function returns the ratio estimate of the total, its absolute error, the usual estimate of variance and the model based estimate of variance. The numbers are given just below.

```
Rweb:> round(ratiototboth(smp1,popy,popx),digits=2)
[1] 55849.46 1742.75 1178008.82 1014594.19
Rweb:> round(ratiototboth(smp2,popy,popx),digits=2)
[1] 54051.32 55.39 515595.69 529396.50
Rweb:> c(mean(popx[smp1]),mean(popx[smp2]))
[1] 31.30887 29.21815
Rweb:> c(mean(popx),sum(popy))
[1] 30.02852 54106.70909
```

From the class notes model estimate of variance is given by

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

where

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

Note the term  $\bar{x}_{nsmp}/\bar{x}_{smp}$  will be less than one when the units with larger values of  $x$  are over sampled and greater than one when they are under sampled as happens here with *smp1* and *smp2*.

4. A population of size 6,000 was broken into 3 strata of sizes 1,000, 2,000 and 3,000 respectively. Proportional allocation was used to get a sample of size 150 using simple random sampling without replacement within each stratum. The data are located at

<http://users.stat.umn.edu/~gmeeden/classes/5201/moredata/strprob13.txt>

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

i) Find the usual point estimate and 95% confidence for the population mean.

ii) Does their decision to use proportional allocation seem sensible given the results of the sample? Briefly explain.

**Ans** i) Modifying the code in the Rweb handout, “Calculating estimates for stratified random samples” one finds that the point estimate is 33.13 and the confidence interval is (32.74, 33.51).

```
Rweb:> postscript(file= "/tmp/Rout.13363.ps")
Rweb:> X <- read.table("/tmp/Rdata.13363.data", header=T)
Rweb:> attach(X)
Rweb:> names(X)
[1] "strmb" "yval"
Rweb:>
Rweb:>
Rweb:> names(X)
[1] "strmb" "yval"
Rweb:> foo<-split(yval,strmb)
Rweb:> smpsizes<-c(25,50,75)
Rweb:> stratsizes<-c(1000,2000,3000)
Rweb:> popsize<-6000
Rweb:> WW<-stratsizes/popsize
Rweb:> WW
[1] 0.1666667 0.3333333 0.5000000
Rweb:> smpmns<-sapply(foo,mean)
Rweb:> smpmns
      1      2      3
39.55200 34.93820 29.77773
Rweb:> est<-sum(WW*smpmns)
Rweb:> est
[1] 33.12693
Rweb:> smpvar<-sapply(foo,var)
Rweb:> smpvar
      1      2      3
23.4350583 4.7701416 0.8192421
Rweb:> ratiosmptopop<-smpsizes/stratsizes
```

```

Rweb:> ratiosmptpop
[1] 0.025 0.025 0.025
Rweb:> estvarofest<-sum((WW^2)*(1-ratiosmptpop)*(1/smpsizes)*smpvar)
Rweb:> SEofest<-sqrt(estvarofest)
Rweb:> SEofest
[1] 0.195923
Rweb:> lwbd<-est - 1.96*SEofest
Rweb:> upbd<-est + 1.96*SEofest
Rweb:> c(lwbd,upbd)
[1] 32.74292 33.51094

```

ii) Recall if  $N_h$  is the stratum size and  $\sigma_h$  is the stratum standard deviation then the optimal allocation is given by

$$n_h = n \frac{N_h \sigma_h}{\sum_j N_j \sigma_j}$$

So if we assumed that the observed strata variances where the true strata variances our allocation would be 61, 55 and 34 to the 3 strata which is far from the actual stratification.

5. 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 125.

	Age				Sum of weights
	20-29	30-39	40-49	50-59	
Male	75	400	325	300	1100
Female	125	300	275	200	900
Sum of weights	200	700	600	500	

Assume it is known the that the population contains 950 men and 1050 women and 150 persons between the ages of 20-29, 650 between 30-39, 750 between 40-49 and 450 between 50-59. Readjust the cells weights so that in the new table the marginal weights agree with the known population weights.

**Ans** Using the standard raking adjustment I found

First time through yields

```

      [,1] [,2] [,3] [,4]
[1,] 46.13 322.88 349.97 236.77
[2,] 103.87 327.12 400.03 213.23

```

The second time through yields

```

      [,1] [,2] [,3] [,4]
[1,] 45.76 321.01 347.82 235.48
[2,] 104.24 328.99 402.18 214.52

```

The third time trough makes the rows ok up to 2 decimal places

```

      [,1] [,2] [,3] [,4]
[1,] 45.76 320.99 347.79 235.46
[2,] 104.24 329.01 402.21 214.54

```

6. Consider a finite population of size  $N$ . Let  $A$  be the total number of units in the population that belong to a certain group. Suppose we take a random sample without replacement of size  $n$  and observe that  $a$  units in the sample belong to the group.

i) Give the usual 95% confidence interval for  $A$ .

ii) Now consider the problem of determining the sample size needed so that the ratio of the standard deviation of the estimator  $Np$  to  $NP$  is no larger than 0.05 when it is known that  $P \geq 0.50$ . What is the necessary sample size if it is known that  $P \geq 0.20$ . You may ignore the finite population correction factor when doing these calculations.

i) **Ans** i) Let  $P = A/N$ ,  $Q = 1 - P$ ,  $p = a/n$  and  $q = 1 - p$ . Then  $Np$  is an unbiased estimator of  $A$  with

$$\begin{aligned} \text{Var}(Np) &= N^2 \frac{PQ}{n} \frac{N-n}{N-1} \\ &\hat{=} N^2 \frac{pq}{n-1} (1-f) = \text{std err} \end{aligned}$$

and the confidence interval is just  $Np \pm 1.96\sqrt{\text{std err}}$ .

ii) From part i) we see that

$$\frac{\sqrt{\text{Var}(Np)}}{NP} = \frac{1}{\sqrt{n}} \sqrt{\frac{Q}{P}} \sqrt{\frac{N-n}{N-1}}$$

Now  $Q/P = (1 - P)/P$  is a decreasing function of  $P$  as  $P$  increases from 0 to 1. So, ignoring the finite population factor, the solution  $n$  must satisfy the equation

$$\frac{1}{\sqrt{n}} \sqrt{\frac{1 - P^*}{P^*}} = 0.05$$

where  $P^*$  is the known lower bound for the possible values of  $P$ . Hence, the necessary sample sizes are 400 and 1,600.

7. Consider a random sample of size 95 from a population of size 1,000. There are two domains of interest in the population say domain A and domain B and their sizes are unknown. The sample can be found in

<http://@users.stat.umn.edu/~gmeeden/classes/5201/moredata/domain13.txt>

The first column of the data matrix is a domain identifier. The number 1 indicates a unit from domain A, the number 2 indicates a unit from domain B and the number 3 indicates a unit which does not belong to either domain. In the sample the first 25 units are from domain A, the next 20 from domain B and the remaining 50 belong to neither domain.

i) Give the usual estimate of the population total for domain A along with its 95% confidence interval.

ii) Use Polya sampling to generate 1,000 simulated, complete copies of the population from the observed values in the sample. For each simulated complete copy find the total of the units that belong to domain A. Use these 1,000 simulated totals to find a point estimate and an approximated 95% confidence interval for the population total for domain A. The following code could be helpful in answering this question.

```
polyasim<-function(ysmp,N)
{
```

```

n<-length(ysmp)
dans<-1:n
simplpop<-ysmp
for( i in 1:(N-n)){
  dum<-sample(dans,1)
  dans<-c(dans,dum)
  simplpop<-c(simplpop,simplpop[dum])
}
return(simplpop)
}

```

iii) Use Polya sampling to find a point estimate and an approximate 95% confidence interval for the population mean of domain A minus the population mean of domain B.

**Ans** i) For this problem the standard approach is to replace each  $y_i$  for  $i \notin A$  with the value 0 and then compute the usual estimates. Denote this new sample by  $y'_i$ .

$$\text{pt est} = \frac{1000}{95} \sum_{i=1}^{95} y'_i = 8646.3$$

The standard error of this estimate is 1430.0 so the interval estimate is  $8646.3 \pm 1.96(1430.0)$  which equals (58435.5, 114490)

ii) Using the suggested code I found that the point estimate under Polya sampling was 8634.8 and the approximated interval estimate was (6129.2, 11713.2) which are very similar to part i).

iii) For this part you need to keep track of both domains and a slight variation on the suggested code will do that. The point estimate of the difference was -11.76 and the approximate 95% confidence interval was (-13.56,-9.83).

8. In this problem we will consider a very special kind of judgment sampling which assumes that it is possible to efficiently and cheaply compare a pair of units in the population while determining the actual value of a unit is much more difficult. We will consider the special case where sampled units come in pairs and are ranked by an "expert". In other words given a pair of sampled units the expert can decide which has the larger  $y$  value. We do not need to assume that the expert is always correct. Here we will assume that the ranking is done using the values of an auxiliary variable  $x$  which is correlated with  $y$ , the characteristic of interest. In half of the ranked pairs we will then observe the  $y$  value of the unit which was ranked smaller and in the other half of the ranked pairs we observe the  $y$  value of the unit which was ranked larger.

Given a sample of labels, say  $smp$  of length  $2n$  and an auxiliary variable  $popx$  this returns a such a judgment sample of length  $n$  where  $n/2$  units were the smaller unit in a ranked pair and  $n/2$  units where the larger of the ranked pair.

```

findsmp<-function(smp,popx)
{
  subsmp<-numeric()
  n<-length(smp)/2
  odd<-seq(1,n-1,length=n/2)
  even<-seq(2,n,length=n/2)
  for(i in odd){
    if(popx[smp[i]] < popx[smp[i+n]]) subsmp<-c(subsmp,smp[i])
    else subsmp<-c(subsmp,smp[i+n])
  }
}

```

```

}
for(j in even){
  if(popx[smp[j]] > popx[smp[j+n]]) subsmp<-c(subsmp,smp[j])
  else subsmp<-c(subsmp,smp[j+n])
}
return(subsmp)
}

```

i) Generate a population of size 500 where the correlation between  $x$  and  $y$  is at least 0.50. (Use `set.seed` so I can generate the population if I wish.) Then generate a random sample of size 40 and use the above code to get a judgment sample of size 20. Then compute the estimate of the population mean using the full sample, the judgment sample and the sample which consists of the first 20 units in the full random sample. Next find the absolute error of each estimate. Repeat this for 500 samples and present the average value and average absolute error for the 3 estimates.

ii) Based on the results in part i) briefly discuss how you think this estimator might behave more generally.

**Ans i)** Using the code

```

findestlp<-function(n,popy,popx,R)
{
  N<-length(popy)
  mny<-mean(popy)
  ans<-matrix(0,3,2)
  for(i in 1:R){
    smpful<-sample(1:N,n)
    smpcmp<-findsmp(smpful,popx)
    estful<-mean(popy[smpful])
    errful<-abs(estful-mny)
    estcmp<-mean(popy[smpcmp])
    errcmp<-abs(estcmp-mny)
    esthlf<-mean(popy[smpful[1:(n/2)]])
    errhlf<-abs(esthlf-mny)
    ans[,1]<-ans[,1] + c(estful,estcmp,esthlf)
    ans[,2]<-ans[,2] + c(errful,errcmp,errhlf)
  }
  ans<-round(ans/R,digits=2)
  return(ans)
}

```

I found

```

> set.seed(554477)
> x<-rgamma(500,5) + 40
> y<-rnorm(500, 3*x,10)
> cor(x,y)
[1] 0.5391985
> mean(y)
[1] 135.3591
> findestlp(40,y,x,500)

```

	[,1]	[,2]
[1,]	135.35	1.45
[2,]	135.31	1.96
[3,]	135.52	2.05

ii) As the correlation between  $x$  and  $y$  increases away from 0.50 we would expect that a judgment sample should produce a better estimator than one based on a simple random sample of the same size because it makes use of some additional information. On the average we would expect it to produce more balanced samples than are produced under simple random sampling. Hence we would expect it to be unbiased and on the average to have smaller error. How much better depends on how good the expert is.