FINAL EXAM
STAT 5201
Fall 2018

Due on the class Canvas web site or in Room 313 Ford Hall
on Tuesday, December 18 at 11:00 am
In the second case please deliver to the office staff
of the School of Statistics

READ BEFORE STARTING

You must work alone and may only discuss these questions with the TA or Professor Meeden. You may use the class notes, the text and any other sources of material you have access to.

Start each answer on a new page and make sure that you name is on each page

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 contacting us.

Forty-seven persons took the exam. the scores ranged from 100 to 5. There were 7 in the 90's, 10 in the 80's, 10 in the 70's, 7 in the 60's, 7 in the 50's, 3 in the 40's and 2 in the 30'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 specific but brief.

2. Information for a population with four strata is given just below.

stratum	size	Est var	cost
1	2000	450	10
2	1500	300	15
3	1000	500	20
4	1200	200	10

The second column of the table gives the sizes of the four strata, the next is our guess for the variance for each strata and the last, the cost per sampling an observation in each strata.

i) Using this information, find the optimal allocation of a sample of size 100.

ii) Suppose we want an estimate of the population mean with an variance of approximately 10. Based on the information in the table what would be the optimal allocation.

ANS i) For stratum h let N_h be its size and $w_h = N_h/N$, sd_h be our guess for the standard deviation and c_h the square root of the cost. Then in R I did the following

```
 w<-c(20/57,15/57,10/57,12/57)
> sd<-c(sqrt(450),sqrt(300),sqrt(500),sqrt(200))
> c<-c(sqrt(10),sqrt(15),sqrt(20),sqrt(10))
> n<-w*sd/c
[1] 2.3537558 1.1768779 0.8771930 0.9415023
> nn<-100*(n/sum(n))
> nn
[1] 44.00095 22.00048 16.39819 17.60038
```

ii) From classwork or the text the solution is given by $n_h = \lambda \frac{w_h sd_h}{c_h}$ where λ is the solution to the equation

$$\lambda = \frac{\sum_h w_h sd_h c_h}{10 + \sum_h w_h^2 sd_h^2/N_h}$$

This gives $\lambda = 6.77$ with the corresponding allocation of 16,8,6 and 6.

3. From a population of 20 clusters a simple random sample of 4 clusters was taken. Within in each sampled cluster a further random sample was taken. The data are in a matrix denoted by $mxclus$. The following two commands will let you see the data in R.

```
library("RCurl")
load(url("http://users.stat.umn.edu/~gmeeden/classes/5201/moredata/clusf18.RData"))
mxclus
```

The cluster sizes are only known for the clusters in the sample. Find the value of the ratio estimator for estimating the population mean and give an estimate of its variance.

ANS Using the handout on the class web page entitled **working with cluster samples** I found the estimate to be 25.21 and with an estimated variance of 19.51.

4. Consider the following table of sums of weights from a sample; each entry in the table is the sum of the 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 98.)

	Age				Sum of
	20-29	30-39	40-49	50-59	weights
Male	183	447	522	416	1568
Female	98	377	395	467	1337
Sum of weights	281	824	917	883	

Assume it is known the that the population contains 1800 men and 1200 women and 300 persons between the ages of 20-29, 900 between 30-39, 1000 between 40-49 and 800 between 50-59. Readjust the cell weights so that in the new table the marginal weights agree with the known marginal population weights.

ANS

To do my raking I used the function *altrake* which can be found on the class web page in the handout called **calibration using quadprog**. Here is my code.

```
>mx<-rbind(c(183,447,522,416),c(98,377,395,467))
>out<-altrake(mx,c(1800,1200),c(300,900,1000,800))
>out
> out
         [,1]     [,2]     [,3]    [,4]
[1,] 210.6091 537.0491 622.9238 429.418
[2,]  89.3909 362.9509 377.0762 370.582
> apply(out,1,sum)
[1] 1800 1200
> apply(out,2,sum)
[1]  300  900 1000  800
```

5. Let $p = (p_1, \ldots, p_n)$ be a probability vector. That is each $p_i > 0$ and they sum to one. Let Δ be the set of all such possible probability vectors. Let $\alpha = (\alpha_1, \ldots, \alpha_n)$ where each $\alpha_i > 0$. We say that p has the Dirichlet distribution with parameter α if its probability density function over Δ is given by

$$f(p_1, \ldots, p_n) = \frac{\Gamma(\sum_{i=1}^n \alpha_i)}{\prod_{i=1}^n \Gamma(\alpha_i)} \prod_{i=1}^n p_i^{\alpha_i - 1}$$

Let $\alpha_0 = \sum_{i=1}^n \alpha_i$. In class it was stated that

$$E(p_i) = \alpha_i/\alpha_0 \quad \text{and} \quad Var(p_i) = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)}$$

i) Show that

$$cov(p_i, p_j) = -\frac{\alpha_i, \alpha_j}{\alpha_0^2(\alpha_0 + 1)}$$

Remember that for any $b > 0$ we have $\Gamma(b + 1) = b\Gamma(b)$

ii) Let $a_1, \ldots, a_n$ be postive real numbers and assume that p has the Dirichlet distribution with parameter α. Find an expression for $Var(\sum_{i=1}^n a_i p_i)$. You may answer the second part even if you can not do the first part.

ANS i) to keep the notation simple we will find the answer when $i = 1$ and $j = 2$.

$$\begin{aligned} E(p_1 p_2) &= \frac{\Gamma(\alpha_0)}{\prod_{i=1}^n \Gamma(\alpha_i)} \int \cdots \int p_1^{\alpha_1 + 1 -} p_2^{\alpha_2 + 1 - 1} \prod_{i=3}^n p_i^{\alpha_i - 1} \, dp_1 \cdots dp_n \\ &= \frac{\Gamma(\alpha_0)}{\prod_{i=1}^n \Gamma(\alpha_i)} \frac{\Gamma(\alpha_1 + 1)\Gamma(\alpha_2 + 1)}{\Gamma(\alpha_0 + 2)} \prod_{i=3}^n \Gamma(\alpha_i) \\ &= \frac{\alpha_1 \alpha_2}{\alpha_0(\alpha_0 + 1)} \end{aligned}$$

Next we note that

$$\begin{aligned} cov(p_1, p_2) &= E(p_1 p_2) - E(p_1)E(p_2) \\ &= \frac{\alpha_1 \alpha_2}{\alpha_0(\alpha_0 + 1)} - \frac{\alpha_1 \alpha_2}{\alpha_0^2} \\ &= -\frac{\alpha_i, \alpha_j}{\alpha_0^2(\alpha_0 + 1)} \end{aligned}$$

ii) Recalling the well know formula for the variance of a sum we will have the answer by substitution.

$$Var(\sum_{i=1}^n a_i p_i) = \sum_{i=1}^n a_i^2 Var(p_i) + 2 \sum_{i<j} a_i a_j cov(p_i, p_j)$$

6. In this problem you must preform a small simulation study to compare the ratio estimator and the regression estimator when the model generating the population is the correct one for the regression estimator. Use the following code to generate two different populations; the first with $sd = 10$ and the second with $sd = 5$. For each of the two populations take 400 simple random samples of size 10 and then of size 50. Compare both the point and interval estimators for the population total for the two estimators in the four cases. Discuss your results. Finally, construct a population such that when the sample size is 10 the average absolute error of the ratio estimator is at least four times larger that of the regression estimator.

```
set.seed(22334455)
popx<-rgamma(1000,4)+40
popy<-rnorm(1000,100 + 2*popx,sd)
```

ANS In each row of the following we have the average values of the estimator, its absolute error, the lower bound of the 95% confidence interval, the length of the interval and the proportion of times the interval contained the true total. Note that for the first population there is not much difference between the two estimators even for the larger sample size. It is only for the second population where the regression estimator does significantly better.

```
Results when sd=10, true total=187783.3 and cov(y,x)=0.41
n=10
ansrt 187782.8 2681.436 181359.7 12846.12 0.9325
ansrg 188019.3 2626.122 182132.2 11774.15 0.9100
n=50
          [,1]     [,2]     [,3]     [,4] [,5]
ansrt 187806.6 1103.800 184962.6 5688.117 0.97
ansrg 187647.6 1069.034 185020.9 5253.344 0.94

Results when sd=5, true total= 187936.8 and cov(y,x)=0.66
n=10
          [,1]     [,2]     [,3]     [,4]  [,5]
ansrt 187880.7 1687.242 183877.0 8007.345 0.925
ansrg 188054.8 1313.061 185111.3 5887.077 0.910
n=50
          [,1]     [,2]     [,3]     [,4] [,5]
ansrt 188031.7 678.1392 186283.0 3497.445 0.96
ansrg 187868.9 534.5169 186555.6 2626.672 0.94

For the last part take popy<-rnorm(1000,200 + 2*popx,2)
keeping popx the same. In this case cor(y,x)=0.91 and
I got for n=10.
          [,1]      [,2]     [,3]      [,4]  [,5]
ansrt 287827.8 2400.8158 282058.8 11538.082 0.905
ansrg 288076.1  525.2245 286898.7  2354.831 0.910
```

Here is the code I used to get the ans.

```
set.seed(22334455)
x<-rgamma(1000,4)+40
```

```
y<-rnorm(1000,100 + 4*popx,5)

ansrtreg<-function(y,x,n,R)
    {
        N<-length(y)
        ans<-matrix(0,2,5)
        for(i in 1:R){
            smp<-sort(sample(1:N,n))
            ansrt<-ratiotot(smp,y,x)
            ansrg<-regtot(smp,y,x)
            ans<-ans+ rbind(ansrt,ansrg)
        }
        return(ans/R)
    }
```

Here *ratiotot* and *regtot* are just the functions found in the R handouts on the class web page dealing with the ratio estimator and the regression estimator.

7. Using the same popx as in problem 6 we now get popy by doing the following

```
popy<-rnorm(1000,100 + 4*popx,5)
```

For this population $cor(popy, popx) = 0.865$. Here you need to do a simulation study where you are estimating this correlation. You will take 300 samples of size $n = 60$ doing pps sampling proportional to popx.

After you get the units in a sample and their design weights renormalize these weights so that they sum to 60, the sample size. The next step is to simulate 500 complete copies of the population using the weighted polya posterior. You can do this using the function *wtpolyap* that is described in the *R* handout **using polyapost** on the class web page. For each simulated complete copy of the population compute the correlation between the two variables. Using these 500 simulate correlations, find your estimate, its relative bias, its absolute error, the length of the approximated 0.95 credible interval and whether or not this interval contains the true correlation. Finally present the average of these quantities over the 300 samples.

Note if *est* is an estimate of the true median, *md*, then the relative bias is equal to $(est - md)/md$.

Here are my results. The point estimate is working fine, the interval estimate is probably a bit short.

```
 rel bias      abs err      len of intv  freq of cov
 -0.005989091  0.029182719  0.128871517  0.910000000
```

Here is the code that I used to get the answer.

```
anscor<-function(y,x,n,Rsim,R)
    {
        N<-length(y)
        wts<-sum(x)/(n*x)
        ans<-rep(0,4)
        trucor<-cor(y,x)
        for(i in 1:R){
            smp<-sort(sample(1:N,n,prob=x))
            swts<-wts[smp]
            swts<-n*swts/sum(swts)
            simcor<-rep(0,Rsim)
            for(j in 1:Rsim){
                out<-wtpolyap(smp,swts,N-n)
                simcor[j]<-cor(y[out],x[out])
            }
            est<-mean(simcor)
            bds<-as.numeric(quantile(simcor,c(0.025,0.975)))
            if(bds[1]<=trucor & trucor<=bds[2]) cov<-1
            else cov<-0
            ans<-ans+c((est-trucor)/trucor,abs(est-trucor),bds[2]-bds[1],cov)
        }
        cat("true cor", trucor,"\n")
        return(ans/R)
    }
```