FINAL EXAM
STAT 5201
Spring 2012

Due in Room 313 Ford Hall
Either Wednesday May 9 or Friday May 11 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.

There were 10 scores in the 90's, 15 in the 80's, 7 in the 70's, 12 in the 60's, 5 in the 50's, 1 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 brief.

2. A question of some interest to policymakers is the proportion of homeless people living in a large metropolitan area who are mentally ill. You have been asked to design a survey that addresses this question. In particular, it has been suggested that by taking a sample of homeless people who receive medical care at the several clinics in area who treat homeless individuals one could find an answer to this question. If you think such a survey is possible briefly outline what you would do. Otherwise explain why you think this suggestion is not feasible.

ANS

The target population is all homeless people in the area. The suggestion is to sample all homeless individuals that seek help at a clinic. Clearly with this approach there could be under-coverage and bias. Also one needs to define carefully what is meant by "being mentally ill". Even if this has been done it could be difficult to check this for individuals in the sample.

3. A simple random sample of size 100 was taken from a large population and the results can be found in the file

```
http://@users.stat.umn.edu/~gmeeden/classes/5201/moredata/rsprob12.txt
```

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

i) Using just the first 20 observations from this sample find the usual estimate of the population mean and its estimated variance.

ii) Repeat part i) but now using the entire sample.

iii) Using the entire sample find the 95% confidence interval for the population median

ANS

```
i) > mean(yval[1:20]) = 70.69411
   > var(yval[1:20])/20 =  62.89749

ii) > mean(yval) = 71.82683
   > var(yval)/100 = 10.34761

iii) dum<-1.96*sqrt(1/(4*100))
     > quantile(yval,c(0.5 - dum, 0.5+dum))
      40.2%    59.8%
      60.37080 72.17915
     40.2%    59.8%
    > median(yval) =  66.22907
```

4. The results from a srs from a stratified population are given in the following table. The columns are the strata sizes, the strata sample sizes, the sample means and the sample variances.

Stratum	N_h	n_h	$\bar{y}_h$	s_h^2
1	100	20	115.7	57.6
2	300	20	147.2	46.9
3	400	30	133.6	75.3

i) Find a 95% confidence interval for the population mean.
ii) The person taking the sampled claimed that they use optimal allocation to select the sample size based on good information about the likely sizes of the strata variances. Given their allocation can you determine their prior guess for the strata variances? Explain. Based on the results of the sample does it appear that they made good choices with their guesses for the strata variances. Assuming the observed sample variances are the true sample variances find the optimal allocation for a sample size of 70.

ANS

i) We have

$$\bar{y}_{str} = \sum_h \frac{N_h}{N}\bar{y}_h = \frac{1}{8}115.7 + \frac{3}{8}147.2 + \frac{4}{8}133.6 = 136.46$$

and

$$\begin{aligned} V(\bar{y}_{str}) \widehat{=} & \sum_h W_h^2 \frac{\sigma_h^2}{n_h}(1-f_h) \\ & = (1/8)^2\frac{57.6}{20}(1-1/5) + (3/8)^2\frac{46.9}{20}(1-1/15) + (4/8)^2\frac{75.3}{30}(1-3/40) \\ & = 0.9242 \end{aligned}$$

from which it is easy to find the confidence interval.

ii) Let σ_h^2 denote the prior guess for the stratum variance for stratum h used in the determination of the sample sizes. Recall that the optimal sample sizes must satisfy

$$\frac{n_h}{n} = \frac{N_h\sigma_h}{\sum_{i=1}^3 N_i\sigma_i}$$

Note if we replace each σ_h with $\lambda\sigma_h$ for any constant $\lambda > 0$ then the optimal allocation does not change. So we can only determine their original choices up to a proportionality constant. Since $n_h/n \propto N_h\sigma_h$ we have for $h > 1$

$$\frac{\sigma_h}{\sigma_1} = \frac{n_h N_1}{N_h n_1}$$

which equals 1/3 and 3/8 for $h = 2$ and 3. We see however that the sample values of these two ratios are 0.90 and 1.14 so the prior guesses were poor and too many observations were allocated to stratum 1.

For optimal allocation note

$$\frac{n_h}{n_1} = \frac{N_h s_h}{N_1 s_1} \quad \text{or} \quad n_h = \frac{N_h s_h}{N_1 s_1} n_1$$

So we find that $n_2 = 2.71\,n_1$ and $n_3 = 4.57\,n_1$ which yields the optimal allocation of 8, 23 and 39 for the sample sizes from the three strata.

5. A simple random sample without replacement was used to select 5 clusters from a population of 29 clusters. From each selected sample cluster simple random sampling without replacement was used to select samples of secondary units. The data are located at

```
http://@users.stat.umn.edu/~gmeeden/classes/5201/moredata/clusprob12.txt
```

Note you do not need our password and username to access these data. For these data find an estimate of the population total and the population mean along with their estimated variances.

ANS

Using the handout "Working with cluster samples" which is available on the course web site and setting $N = 29$ we find

$$\hat{t}_y = 31,841.13, \quad V(\hat{y}_y)\hat{=}17606605, \quad yratio = 41.28, \quad V(yratio)\hat{=}2.80$$

6. Construct a stratified population with three strata where the strata sizes are 200, 300 and 500. For your population and a fixed sample size n find the true variance of the usual stratified estimator when proportional allocation is used. Compare this variance to the true variance of the usual sample mean for a srs of the same size. The problem is to construct a population where the stratified estimator has a larger variance than the variance of sample mean. Describe how you constructed your population and computed the two variances. When generating you population use the set.seed() command in R so that I can reconstruct your population.

ANS

Let V_{srs} and V_{prop} denote the variances for the sample mean and the stratified estimator. I showed in class that

$$V_{srs} = V_{prop} + \frac{1-f}{n(N-1)}\left\{\sum_h N_h(\bar{Y}_h - \bar{Y})^2 - \frac{1}{N}\sum_h (N - N_h)\sigma_h^2\right\}$$

So you want a population where the expression in $\{\}$ is negative. This will happen when each stratum is very much like any other stratum. So to find such a population I took a random sample of size 1,000 from a normal distribution with mean equal to 200 and a standard deviation of 30. (I used set.seed(99876).) For this population and $n = 50$ I found that $V_{srs} = 17.10$ and $\{\} = -0.0188$.

Another way is to make stratum 1, where the fewest observations will be taken, have the largest variance. For example in R let

```
pop<-c(rnorm(100,5,.1),rnorm(100,95,.1),rnorm(800,10,.1))
```

For this population with $n = 50$ I found that $V_{srs} = 13.23$ and $\{\} = 4.84$.

7. Consider 11 urns labeled $0, 1, \ldots, 10$ respectively. Each urn contains 10 balls where urn i contains i balls labeled 1 and $10 - i$ balls labeled 0. Let $\pi_i > 0$ where $\pi_0 + \pi_1 + \cdots + \pi_{10} = 1$. Suppose a urn is chosen at random according the distribution defined by the π_i's and then four balls are selected at random from the selected urn using **simple random sampling without replacement**. If 3 of the 4 balls selected were 1 and 1 was 0 find:

i) The posterior probability that urn j was chosen for $j = 3, 4, \ldots, 9$.

ii) The probability that the next ball chosen at random from the selected urn has the value 1.

ANS

i) Let E be the event that that 3 of the four balls selected were 1 and the other was 0. Then

$$P(E) = \sum_{k=3}^{9} \pi_k \frac{\binom{k}{3}\binom{10-k}{1}}{\binom{10}{4}}$$

Let A_j denote the event that the urn with j 1's was selected. Then

$$P(A_j \mid E) = \pi_j \frac{\binom{j}{3}\binom{10-j}{1}}{\binom{10}{4}} / P(E)$$

ii) Let B be the event that the next ball selected is a 1. Then

$$\begin{aligned}
P(B \mid E) &= \sum_j P(B \cap A_j \mid E) \\
&= \sum_j P(E \cap A_j \cap E)/P(E) \\
&= \sum_j \Big(P(E)P(A_j \mid E)P(B \mid A_j, E) \Big)/P(E) \\
&= \sum_j P(B \mid A_j)P(A_j \mid E) \\
&= \sum_{j=3}^{9} \frac{j-3}{6} P(A_j \mid E)
\end{aligned}$$

8. Let *ysmp* denote a sample of size n from a population of size N. The following simple R function

```
polyasmp<-function(ysmp,N){
  n<-length(ysmp)
  ans<-ysmp
  for(i in 1:(N-n)){
    ans<-c(ans,sample(ans,1))
  }
  return(ans)
}
```

lets one use polya sampling to generate complete copies of the population using the values in a sample.

For this problem you need to construct a population of 300 values from which you will take 100 simple random samples without replacement of size 21. For each sample you will need to find the sample median. For each sample you will also need to simulate 500 complete copies of the population using the above function. For each simulated complete copy of the population find its median and then the average of these 500 simulated medians to get a second estimate of the population median based on polya sampling. To compare these two methods of estimation find the absolute value of the difference between each estimate and the true population. Finally, report the average value of each method of estimation and the average of the absolute errors for the two methods where the average is taken over the 100 samples you observed.

ANS

In R I used set.seed(112233) and then generated my pop as follows: pop¡-5*rgamma(300,4) Now median(pop) = 18.17 and for my 100 samples of size 21 I found that the average value of the sample median was 18.16 with average absolute error of 1.58. For the estimator based on polya sampling from the sample to generate complete simulated copies of the population the average value of the estimator was 18.23 with an average absolute error of 1.46.

Below is the R code I used to get my answer.

```
onesmp<-function(pop,n,R)
  {
    trmd<-median(pop)
    smp<-sample(pop,n)
    est<-median(smp)
    frqans<-c(est,abs(est-trmd))
    N<-length(pop)
    dans<-rep(0,R)
    for(i in 1:R){
      dans[i]<-median(polyasmp(smp,N-n))
    }
    est<-mean(dans)
    ppans<-c(est,abs(est-trmd))
    ans<-c(frqans,ppans)
    return(ans)
  }

onesmplp<-function(pop,n,R,R1)
  {
    ans<-rep(0,4)
    for(i in 1:R1){
      ans<-ans + onesmp(pop,n,R)
    }
    ans<-round(ans/R1,digits=2)
    return(ans)
  }
```