FINAL EXAM
STAT 5201
Spring 2009

Due in Room 313 Ford Hall
Friday May 15 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 well known "Fast Food" chain wants to estimate the potential size of the market for a new product. The plan is to try it out at a sample of n restaurants ($N = 970$) where they will employ heavy local advertising. At the end of 4 weeks they will measure sales volume (in \$) for one week. We are interested in estimating the total potential size of this new market using a SRS without replacement of size n. Let

$$t(y) = t = \sum_{i=1}^{N} y_i \quad \text{and} \quad \hat{t}(y_{smp}) = \hat{t} = N\bar{y}_{smp}$$

be the population total and its usual estimate. You may assume that $\hat{t}$ is approximately normally distributed.

(a.) Vice-president A believes that a good upper bound for the population variance of the potential volume is \$3,000,000. He wants the length of a 95% confidence interval for the total potential volume to be no more than \$1,000,000. Propose a sample size for him.

(b.) Vice-president B believes that the coefficient of variation for the population of potential volume is approximately 0.15. she wants

$$P\left(\frac{|\hat{t} - t|}{t} \leq 0.10)\right) = 0.95$$

Propose a sample size for her.

(c.) There is also Vice-president C who is convinced that they will see that the volume of the new product will be a nearly constant percentage of each restaurant's current total sales volume which is well-known to them. Propose a survey plan to V.P. C and explain why the company should be interested.

Answer

a) Let σ^2 be the population variance. Then to have

$$P\left(|\hat{t} - t| \leq 500,000) = P(|N(0,1)| \leq \frac{500,000}{976\sqrt{(1 - n/976)(\sigma^2/n)}}\right) = 0.95$$

we must have

$$\frac{500,000}{976\sqrt{(1 - n/976)(\sigma^2/n)}} = 1.96 \quad or \quad n = 970/(1 + \frac{500,000^2}{(1.96^2)\,970\,\sigma^2})$$

But replacing σ^2 by an upper bound will be conservative and this yields $n = 41.52$ so a sample of size 42 should do the job.

b) To have

$$P(\frac{|\hat{t}-t|}{t} \leq 0.10)) = P\left(|N(0,1)| \leq \frac{0.1t}{N\sqrt{(1-n/N)(\sigma^2/n)}}\right)$$
$$= P\left(|N(0,1)| \leq \frac{0.1((t/N)/\sigma)}{\sqrt{(1-n/N)(1/n)}}\right)$$
$$P\left(|N(0,1)| \leq \frac{0.1/0.15}{\sqrt{(1-n/N)(1/n)}}\right)$$
$$= 0.95$$

we must have

$$\frac{(2/3)}{/\sqrt{1/n - 1/970}} = 1.96 \quad \text{or} \quad n = 8.64$$

So a sample of size 9 will do the job.

c) If they are correct then the ratio estimator should work better.

3. Optimal allocation was used to assign sample sizes to the two strata described below

stratum	N_h	σ_h^2	n_h	c_h
1	2000	60	116	\$9
2	800	180	84	?

(a.) Determine c_2 to the nearest dollar.

(b.) What is the standard error of $\bar{y}_{st}$ for this sampling plan? SRS without replacement was used in each stratum.

(c.) Reallocate the sampling budget using proportional allocation and calculate the standard error of $\bar{y}_{prop}$.

Answer

a) In terms of the total sample size n_h in stratum h under optimal allocation we must have

$$\frac{n_h}{n} = \frac{N_h\sigma_h/\sqrt{c_h}}{\sum_k N_k\sigma_k/\sqrt{c_k}}$$

where n is the total sample size. Since ever term except c_2 is known in the above equation it is easy to solve for it. We find that $c_2 = 8.24$.

b) Remembering that

$$V(\bar{y}_{str}) = \sum_h W_h^2 \frac{\sigma_h^2}{n_h}(1 - n_h/N_h)$$

where $W_h = N_h/N$ we find $V(\bar{y}_{str}) = 0.249 + 0.157 = 0.405$ and so the answer is $\sqrt{0.405} = 0.637$.

c) For proportional allocation $n_h = n(N_h/N)$ which is this case becomes 143 and 57 with variance equal 0.438. So answer is $\sqrt{0.438} = 0.662$.

4. A foundation mailed out questionnaires to randomly drawn owner occupied residential addresses in St. Paul. Their sampling frame was prepared by the Ramsey County Assessor's Office.

One question asked was if the responding household favored using real estate taxes to support sports stadium construction. By the cut-off date for returning the questionnaires they had received 600 replies and 75% favored the proposal.

(a.) Using these data construct an unbiased estimate of the proportion of owner occupied residential households in St. Paul which favor real estate taxes for stadium construction. Find the standard error of your estimate.

(b.) A clerk at the foundation was examining the 600 returned mail responses and noticed that 83.3% of the respondents were under age 40, whereas the city census showed that only 30% of residential real estate owners were under 40. Those respondents under 40 favored the proposal by 84% while those over 40 were only 30% in favor. Use this information to create a new estimate of the proportion of owner occupied residential households in St. Paul which favor real estate taxes for stadium construction.

(c.) Since by the cut-off date they had received only 600 of the 4000 mailed questionnaires the sponsors of the survey were concerned about response bias. They decided to use interviewers to contact a random sample of 300 out of 3400 non-respondents. In this sample 40% favored the proposal. Use this additional information to improve your estimate in part b.

Answer

a) The estimate is 0.75 with a standard error of $\sqrt{(0.4 \times 0.6)/600} = 0.018$.

b) We want to post-stratify using age to define two strata; owners ¡ 40 and those who are older. The new estimate is

$$(3/10)(0.84) + (7/10)(0.30) = 0.462$$

c) We can think of the non-responders as forming a separate strata which forms $3400/4000 = 0.85$ part of the population. So the new estimate is

$$(0.15)(0.462)) + (0.85)(0.40) = 0.409$$

5. Consider a Medical device which must be surgically implanted into the human body. A manufacturer is introducing a new device and wishes to develop a sampling plan which among other things would allow them to estimate the probability that the device fails after t days.

The company will have a registry of every device implanted. This registry will include the date of the implant, where it was done and contact information for the patient at time of implant.

The implants must be preformed in hospitals and more than 50% of the implants will be done in hospitals which will preform at most 5 operations in a given year. After a year the distribution of the number of hospitals which will have done i implants might look something like this.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	20
155	203	128	95	79	84	84	72	62	48	31	31	18	21	14	10	6	5	1

For example 95 hospitals have done 4 implants.

Cost considerations make it preferable to sample hospitals that do more implants. Also the company researchers prefer getting more information than just whether or not the device has failed. They believe that larger hospitals are more forth coming and they want to see more of them in the sample.

(a.) Develop a sampling plan for the first year which tries to address these conflicting needs.

(b.) Briefly explain how the results of the first year might affect your sampling design for later years.

Answer

(a.) Hospitals should be considered as clusters. In the first year you could select clusters using pps proportional to size. Then in the larger clusters (say hostipals with > 10 implants) you could only sample 10 of the patients.

(b.) If the results from the first year indicate that there is not much difference between patients who received their implants at small hospitals and those who received then at large hospitals then future samples could focus more on the larger hospitals.

6. In the model based approach the auxiliary variable x is considered to be fixed and the values of y are realizations from some joint distribution. A standard model, which leads to the ratio estimator, is

$$y_i = \beta\, x_i + z_i$$

where the z_i's are independent random variables with $E(z_i) = 0$ and $V(z_i) = x_i\sigma^2$.

On the class web page there is a link entitled "Model based properties of the ratio estimator" which you should use to answer this question. In it I generated a particular set of values for x which you should take as fixed. To generate possible sets of y values I set $\beta = 3$ and $\sigma = 8$ which for part a) of the question you may also take as fixed.

For a fixed sample, *smp*, we are interested in the average behavior of the $\delta_{ratio}(y_{smp})$ where we are averaging over the possible values of y_{smp} under our model. In particular, for a fixed sample, we are interested in:

- the average bias of the estimator, or the average of

  $$\delta_{ratio}(y_{smp}) - \mu(y)$$

  where $\mu(y)$ is the mean of y and

- the average of estimated model variance of the estimator, or the average of

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

  where

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

In the code I generate $K = 1000$ sets of possible y and found the above two averages for the fixed sample

```
> smp
 [1]   9  29  49  69  89 109 129 149 169 189
```

(a.) For this model how much do these two computed averages depend on my choice of the sample? (Hint: you should rerun the code with different choices for the fixed sample.)

(b.) How strongly do your answers to part a) depend on my choices for β and σ^2?

Answer

a) From class we know that the first average should always be approximately 0 since the ratio estimator is model unbiased. The second average does depend on the values in *smp* and is smallest when it contains the n units with the largest x values.

b) Again only the second average should depend on these choices.

7. Even though the Horvitz-Thompson estimator is always design unbiased if the weights vary a lot it can have quite a large variance. In practice very small weights are increased and very large weights are decreased. On the class web page there is a link entitled "Adjusting the Horvitz-Thompson Estimator" which allows one to make simple adjustments in the HT weights. Let $lowbd < upbd$ be two specified numbers. Then given a sample any weight which is below $lowdb$ is set equal to this value and any weight which is above $upbd$ is set equal to this value. Although this can destroy the unbiasedness of the HT estimator it can make its mean squared error smaller.

In the code I have created a simple population where the HT estimator would be appropriate when the design is pps proportional to x. The code allows one to draw K samples and compare the HT and adjusted HT for particular choices of $lowbd$ and $upbd$, In the code my choice of $lowbd = 12.9$ and $upbd = 48.8$ does not led to any improvement. but other choices do. The problem is to use the code to find values of $lowbd$ and $upbd$ for which the adjusted HT is at least 10% better than the HT.

Answer

Not clear what is the best choice but $lowbd = quantile(wts, 0.1)$ and $upbd = quantile(wts, 0.9)$ works.