36-303: Sampling, Surveys and Society
Exam 2
Tue Apr 13, 2010

- You have 80 minutes for this exam.
- The exam is closed-book, closed notes.
- A calculator is allowed.
- **Two formula sheets are provided for your convenience.**
- Please write all your answers on the exam itself; your work must be your own.

Question	Points Possible	Points Earned
1	20	
2	20	
3	18	
4	24	
5	18	
Total	100	

Name: _____

Signature: _____

Some Useful Formulas From the Statistics of Survey Sampling, I

Equally-Likely Outcomes & Counting

- If K outcomes $O_1, \ldots, O_K$ are equally likely, then the probability of any one of them is $1/K$.
- Consider taking a sample of n objects from a population of N objects.
  - Sampling with replacement, there are N^n possible samples of size n; the probability of any one of them is $1/N^n$.
  - Sampling without replacement, there are $\binom{N}{n} = \frac{N!}{n!(N-n)!}$ possible samples of size n [where $N! = N \cdot (N-1) \cdot (N-2) \cdots 3 \cdot 2 \cdot 1$], so the probability of any one of them is $1 \Big/ \binom{N}{n}$.

Discrete Random Variables

Let X and Y be random variables with sample spaces $\{x_1, \ldots, x_K\}$ and $\{y_1, \ldots, y_K\}$ and distributions

$$P[X = x_i, Y = y_j] = p_{ij} \quad , \quad P[X = x_i] = p_{i\cdot} = \sum_{j=1}^{K} p_{ij} \quad , \quad P[Y = y_j] = p_{\cdot j} = \sum_{i=1}^{K} p_{ij}$$

Then, for example

$$E[X] = \sum_{i=1}^{K} x_i p_{i\cdot} \quad , \quad Var(X) = \sum_{i=1}^{K} (x_i - E[X])^2 p_{i\cdot} \quad , \quad Cov(X,Y) = \sum_{i=1}^{K} (x_i - E[X])(y_i - E[Y]) p_{ij}$$

$$P[X = x_i | Y = y_j] = p_{ij}/p_{\cdot j} \quad , \quad E[X|Y = y_j] = \sum_{i=1}^{K} x_i P[X = x_i | Y = y_j] \quad , \quad E[aX + bY + c] = aE[X] + bE[Y] + c$$

Random Sampling From a Finite Population

Consider a population of size N and a sample of size n. Let y_i be the (fixed) values of some variable of interest in the population (such as a person's age, or whether they would vote for Obama). Let

$$Z_i = \begin{cases} 1, & \text{if } i \text{ is in the sample} \\ 0, & \text{else} \end{cases}$$

be the random sample inclusion indicators, and let Y_i be the random observations in the sample. Then the sample average can be written

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i = \frac{1}{n} \sum_{i=1}^{N} Z_i y_i$$

The Z_i's are Bernoulli random variables with

$$E[Z_i] = \frac{n}{N} \quad , \quad Var(Z_i) = \frac{n}{N}\left(1 - \frac{n}{N}\right) \quad , \quad Cov(Z_i, Z_j) = -\frac{1}{N-1}\frac{n}{N}\left(1 - \frac{n}{N}\right)$$

Confidence Intervals and Sample Size

(a) A CLT-based $100(1-\alpha)\%$ confidence interval for the population mean is $(\overline{Y} - z_{\alpha/2} SE \ , \ \overline{Y} + z_{\alpha/2} SE)$.

(b) For sampling with replacement from an infinite population, $SE = SD/\sqrt{n}$.

(c) For sampling without replacement from a finite population, the SE has to be multiplied by the finite population correction (FPC).

(d) For a given margin of error (ME, half the width of the CI) and confidence level $1-\alpha$, we can find the sample size by solving

$$z_{\alpha/2} SE < ME$$

for n. The same approach works for both SRS with replacement (using the SE in (b)) and SRS without replacement (using the SE in (c)).

Some Useful Formulas From the Statistics of Survey Sampling, II

Stratified Sampling

Consider H strata with population counts $N = \sum_{h=1}^{H} N_h$ and sample counts $n = \sum_{h=1}^{H} n_h$. Let $f_h = n_h/N_h$; $W_h = N_h/N$; and $\overline{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{ih}$ in each stratum, and let $s_h^2 = \frac{1}{n_h-1}\sum_i (y_{ih} - \overline{y}_h)^2$ be the sample variance in each stratum. Then

$$\overline{y}_{st} \;=\; \sum_{h=1}^{H} W_h \overline{y}_h \;\;, \quad \mathrm{Var}\,(\overline{y}_{st}) \;\approx\; \sum_{h=1}^{H} W_h^2 (1-f_h)\frac{s_h^2}{n_h} \;\;, \quad DEFF \;=\; \frac{\mathrm{Var}\,(\overline{y}_{st})}{\mathrm{Var}\,(\overline{y}_{srs})} = \frac{\sum_{h=1}^{H} W_h^2 (1-f_h)\frac{s_h^2}{n_h}}{(1-f)\frac{s^2}{n}}$$

Cluster Sampling

Consider a population of N clusters. We take an SRS $\mathcal{S}$ of n clusters, and all units within each sampled cluster (one-stage clustering). Assume clusters all have same size M. Let $\overline{y}_i = \frac{1}{M}\sum_{j=1}^{M} y_{ij}$ in each cluster. Then

$$\overline{y}_{cl} \;=\; \frac{1}{n}\sum_{i\in\mathcal{S}} \overline{y}_i \;\;, \quad \mathrm{Var}\,(\overline{y}_{cl}) \;\approx\; \left(1-\frac{n}{N}\right)\frac{1}{n} s^2_{\overline{y}_i} \;=\; \left(1-\frac{n}{N}\right)\frac{1}{n}\left[\frac{1}{n-1}\sum_{i\in\mathcal{S}} (\overline{y}_i - \overline{y}_{cl})^2\right]$$

and

$$DEFF = \frac{\mathrm{Var}\,(\overline{y}_{cl})}{\mathrm{Var}\,(\overline{y}_{srs})} = \frac{M s^2_{\overline{y}_i}}{s^2_{y_{ij}}} \approx 1 + (M-1)\rho$$

where $s^2_{\overline{y}_i}$ is the sample varance of the cluster means, $s^2_{y_{ij}}$ is the sample variance of the individual observations, and ρ is the intraclass (intracluster) correlation, or ICC.

Post-Stratification Weights and Means

As part of survey data collection it is a good idea to get general demographic information (e.g. in our surveys: sex, age, class, major, hometown, etc.). After data collection we compare the proportions in each of these categories in our sample with the same proportions in the population. If they agree, great. If not, calculate

$$w_i = (N_h/N)/(n_h/n) \text{ for each } i \text{ in post-stratum } h \;\;, \quad \textit{and} \quad \overline{y}_w = \frac{\sum_i w_i y_i}{\sum_i w_i}$$

Post-Stratification Variance Calculations

Taylor series:

$$\mathrm{Var}_{TS}(\overline{y}_w) \approx \frac{1}{(\sum_i w_i)^2}\left[\mathrm{Var}\,(\sum_i w_i y_i) - 2\overline{y}_w \mathrm{Cov}\,(\sum_i w_i y_i, \sum_i w_i) + (\overline{y}_w)^2 \mathrm{Var}\,(\sum_i w_i)\right]$$

where $\overline{y}_w$ is as above, $\overline{w} = \frac{1}{n}\sum_i w_i$, $\overline{wy} = \frac{1}{n}\sum_i w_i y_i$,

$$\mathrm{Var}\,(\textstyle\sum_{i=1}^n w_i) \approx n \cdot \frac{1}{n-1}\sum_{i=1}^n (w_i - \overline{w})^2, \quad \mathrm{Var}\,(\sum_{i=1}^n y_i w_i) \;\approx\; n \cdot \frac{1}{n-1}\sum_{i=1}^n (w_i y_i - \overline{wy})^2,$$

$$\mathrm{Cov}\,(\textstyle\sum_{i=1}^n y_i w_i, \sum_{i=1}^n w_i) \approx n \cdot \frac{1}{n-1}\sum_{i=1}^n (w_i y_i - \overline{wy})(w_i - \overline{w})$$

Jackknife:

- Replicate n times (by removing one obs. each time and recalculating weights):

$$\overline{y}_w^{(r)} = \frac{\sum_{i=1}^n w_i^{(r)} y_i^{(r)}}{\sum_{i=1}^n w_i^{(r)}}$$

- Calculate

$$\overline{y}_{JK} \;=\; \frac{1}{n}\sum_{r=1}^{n} \overline{y}_w^{(r)} \;\;, \quad Var_{JK}(\overline{y}_w) \;\approx\; \frac{n-1}{n}\sum_{r=1}^{n} (\overline{y}_w^{(r)} - \overline{y}_{jk})^2$$

April 13, 2010 Name: _____

1. [20 pts] REVISE. *Multiple Choice (4 parts). For each part, circle the roman numeral of the one best answer.*

   (a) [5 pts] Let Y_i be the number of monthly neighborhood watch meetings attended by the i^{th} resident in a neighborhood, in the last year. You are going to conduct a survey, using a SRS to estimate $\overline{Y}_{pop}$, the population mean number of meetings attended by neighborhood residents. Among the N_R residents who *would respond* to your survey, the mean number of meetings is $\overline{Y}_R$, and among the N_M number of residents who would *not respond*, the mean number of meetings attended is $\overline{Y}_M$. In class we showed that the *bias* between $\overline{Y}_R$ and $\overline{Y}_{pop}$, due to missing responses, is

   $$\overline{Y}_R - \overline{Y}_{pop} = \frac{N_M}{N}(\overline{Y}_R - \overline{Y}_M)$$

   where $N = N_R + N_M$.

   Which statement below is false (or, circle iv. if all are OK)?

   i. All other things being equal, the more people who would *not respond*, the smaller the bias due to missing responses.
   ii. All other things being equal, the bigger the difference between mean number of meetings attended by nonresponders, vs the mean number attended by responders, the bigger the bias due to missing responses.
   iii. The larger your SRS, the better you can estimate $\overline{Y}_R$.
   iv. All of the above statements are true.

   (b) [5 pts] In one-stage clustered sampling, the ICC ρ measures

   i. The correlation between observations in different clusters.
   ii. The correlation between the cluster means of different clusters.
   iii. The correlation between observations in the same cluster.
   iv. The correlation between the cluster mean and the individual observations in the cluster.

   (c) [5 pts] A 36-303 survey project constructs an SRS of undergraduates and directs them to a website where they can answer the survey questions. However some respondents are never presented with the last page of the survey, due to a bug in the website software (the error can happen at any time to any respondent, but only a few respondents fail to complete the survey for this reason). This is most likely an example of

   i. Unit nonresponse, Missing At Random (MAR)
   ii. Item nonresponse, Missing Completely At Random (MCAR)
   iii. Unit nonresponse, Missing Completely At Random (MCAR)
   iv. Item nonresponse, Missing At Random (MAR)

(d) [5 pts] Suppose we divide a sampling frame into groups, which we may treat as either strata for stratified sampling, or clusters for cluster sampling. If we make the groups so that *observations* **within groups** *are more* **similar** *to each other*, and *observations* **between groups** *are more* **different** *from each other*, then, all other things being equal, we expect

i. The variance of the stratified sample mean $\overline{y}_{st}$ will go **up** and the variance of the cluster sample mean $\overline{y}_{cl}$ will go **down**.

ii. The variance of the stratified sample mean $\overline{y}_{st}$ will go **down** and the variance of the cluster sample mean $\overline{y}_{cl}$ will go **up**.

iii. Both variances will go **up**.

iv. Both variances will go **down**.

[This space intentionally left blank]

2. [20 pts] *Stratified Sampling (3 parts).*

Most local advertisers have a choice of TV, print or radio advertisements. In Jefferson County there are two towns: Alpha is built around a factory, and most of the households contain factory workers with school-age children; Berea is an exclusive suburb of the city of Danville in the next county, and contains mostly older people with few children at home. To see if it is worthwhile to advertise on TV, a stratified sample of households with three strata—Alpha, Berea, and the remaining rural area in the county—is taken, and television viewing time, in hours per week, is recorded for each household in the sample.

The data are as follows:

	Stratum 1 *Alpha*	*Stratum 2* *Berea*	*Stratum 3* *Rural Area*
Number of households	$N_1 = 155$	$N_2 = 62$	$N_3 = 93$
Stratum sample size	$n_1 = 20$	$n_2 = 8$	$n_3 = 12$
TV viewing time, hr/wk	35 28 26 41 43 29 32 37 36 25 29 31 39 38 40 45 28 27 35 34	27 4 49 10 25 41 25 30	8 15 21 7 14 30 20 11 12 32 34 24
Stratum Statistics	$\overline{y}_1 = 33.9000$ $s_1^2 = 35.358$	$\overline{y}_2 = 25.125$ $s_2^2 = 232.411$	$\overline{y}_3 = 19.000$ $s_3^2 = 87.636$

(a) [6 pts] Compute $\overline{y}_{st}$ and $\overline{y}_{srs}$. If they should be the same, explain why. If they should be different, explain why *[use the back of this page to show work, if you need to]*.

(b) [6 pts] Compute SE($\overline{y}_{st}$) and use this to create an approximate 95% confidence interval for the mean household TV viewing time (hrs/wk) in Jefferson County.

(c) [8 pts] If we were to treat this as an SRS rather than a stratified sample, the sample SD would be $s^2 = 83.658$.

i. Estimate the DEFF (design effect) for this design.

ii. Was it a good idea to do the stratified sample rather than an SRS? Why or why not?

3. [18 pts] *Approval Rating of Sen. Diane Feinstein, Part I (3 parts).*

   In late February 2009, the Field Research Corporation (http://www.field.com/fieldpollonline) conducted a poll of Democrats and Republicans in California to assess voters' approval of US Sentator Diane Feinstein, and judge whether any potential candidates might be able to unseat her in the midterm elections in November 2010.

   Approximately equal numbers of Democrats (342) and Republicans (298) were contacted by phone; 267 of the Democrats and 98 of the Republicans approved of the job Sen. Feinstein was doing.

   According to the Ballot Access News (http://www.ballot-access.org/2008/120108.html), there were 7,683,495 registered Democrats and 5,428,052 registered Republicans in California at approximately the same time as the Field Poll was being taken.

   (a) [6 pts] Treating the sample of Democrats as an SRS w/o replacement, compute $\hat{p}_{dem}$, the proportion of Democrats who approve of Sen. Feinstein, and its standard error, SE($\hat{p}_{dem}$).

   (b) [6 pts] Do the same for Republicans: compute $\hat{p}_{rep}$ and SE($\hat{p}_{rep}$).

(c) [6 pts] A confidence interval for the difference between the true proportions of Democrats and Republicans who approve of Senator Feinstein can be computed with the help of the pooled SE, $\sqrt{\text{SE}(\hat{p}_{dem})^2 + \text{SE}(\hat{p}_{rep})^2}$. Compute this pooled SE and give an approximate 95% confidence interval for this difference in proportions of voters who approve of Sen. Feinstein.

4. [24 pts] *Approval Rating of Sen. Diane Feinstein, Part II (4 parts).*

   (a) [6 pts] Combining Democrats and Republicans, and treating the survey as an SRS w/o replacement, estimate the overall proportion of voters in California who approve of the job Sen. Feinstein has been doing. (You do not have to compute the SE).

   (b) [6 pts] We really do not know if the Field Poll was an SRS, what the response rate was, etc., but it seems clear that the proportions of Democrats and Republicans in the poll were different from the population proportions in California. Using political party registration to define two post-strata, compute post stratification weights to apply to this survey.

(c) [6 pts] Using the weights you computed in part (b), compute the weighted proportion of voters in California who approve of Sen. Feinstein's job (you do not have to compute the SE).

(d) [6 pts] Use the values of the weights to explain[1] the difference between your answers in parts (a) and (c).

[1] If you were unable to complete part (c), use the weights to make a prediction about whether your answer to part (c) would be higher or lower than your answer to part (a), and explain why.

5. [18 pts] *Data Collection Methods (3 parts.)* In each situation below, choose the best mode of data collection, and give a reason why. For the "why" question, write complete, clear sentences. Your response does not need to fill all of the available space[2]. Strong reasoning and clear writing count much more than length.

   Think about tradeoffs between feasibility, cost, response rate & nonresponse bias, availability of information on which respondents can base their answers, etc. You may assume there is a good sampling frame, and that a strong sampling method like SRS, stratified or clustered sampling, will be used.

   (a) [6 pts] Your market research firm has been asked to conduct a survey in Pittsburgh to see who would purchase a new, sweet, fruity carbonated soft drink.

   - Which mode of data collection (circle the best answer)?
     i. Telephone survey.
     ii. Face to Face interviews.
     iii. Internet (web-based) survey.
   - Why?

[2] On the other hand, if you need more space, use the back of the page.

(b) [6 pts] The Governor of your state has asked you to design a survey to assess the views of residents of the state toward public transportation and to evaluate how likely they would be to use alternative modes of transportation.

- Which mode of data collection (circle the best answer)?
  i. Telephone survey.
  ii. Face to Face interviews.
  iii. Internet (web-based) survey.
- Why?

(c) [6 pts] The American Statsitical Association wishes to do a survey of its members across the United States on their views toward certification of statisticians (statisticians would need pass a test in order to get a license to practice statistics, like doctors, some engineering professions, etc.).

- Which mode of data collection (circle the best answer)?
  i. Telephone survey.
  ii. Face to Face interviews.
  iii. Internet (web-based) survey.
- Why?