36-303: Sampling, Surveys and Society
Exam 2
Tue Apr 12, 2011

- 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.
- If you need more room, continue onto the back of the same page as the question you are answering *(and let us know that is what you are doing!)*.

Question	Points Possible	Points Earned
1	24	
2	26	
3	24	
4	26	
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

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

Then, for example

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

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

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

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

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

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

and

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

where $s_{\overline{y}_i}^2$ is the sample varance of the cluster means, $s_{y_{ij}}^2$ 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

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

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 12, 2011 Name: ______________________

1. [24 pts] *Multiple Choice (4 parts). For each part, circle the roman numeral of the one best answer.*

   (a) [6 pts] Which of the following is *not* a usual part of post-survey processing?

      i. Data entry
      ii. Sample size calculation
      iii. Imputation
      iv. Checking post-strata and building weights if needed
      v. All of the above *are* usually part of post-survey processing!

   (b) [6 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**.

   (c) [6 pts] Which of the following is *not* one of the recommended things to work on, to reduce the tendency of survey subjects to not respond?

      i. Followup.
      ii. Choice of stratified or cluster sampling.
      iii. Amount of effort it takes respondents to undeerstand/respond to questions.
      iv. Assurance of confidentiality, especially for sensitive questions.

   (d) [6 pts] Weights can be calculated and applied to individual observations for a variety of reasons. Circle the reason below that is *not* appropriate.

      i. Weights may be calculated in designing a stratified sample designs.
      ii. Weights may be calculated in designing certain kinds of surveys in which not every respondent has an equal chance of being selected.
      iii. Weights may be calculated after the survey to compensate for some kinds of informative (non-ignorable) missingness.
      iv. Weights may be calculated after the survey to adjsut sample proportions in various post-strata to equal the population proportions.
      v. All of the above *are* appropriate reasons to compute weights!

2. [26 pts] *Cluster sampling*. A survey is conducted to find out the proportion of cell phone users in a certain city. From a population of 2500 residential blocks, 10 are sampled at random without replacement, and each person residing on that block is asked whether they use a cell phone. We will assume there are exactly 40 people living in each block[1]. This yields the following table of data:

Block h	Total # of People M	# of Cell Phone Users c_h	Proportion of Cell Phone Users p_h
1	40	10	0.25
2	40	8	0.20
3	40	16	0.40
4	40	15	0.38
5	40	24	0.60
6	40	17	0.42
7	40	12	0.30
8	40	13	0.32
9	40	16	0.40
10	40	13	0.32
Total	400	144	

(a) [6 pts] This is an example of one-stage clustered sampling. Circle one word in each pair of choices in the following sentences:

- The primary sampling units (psu's) are the **(blocks, residents)**.
- The secondary sampling units (ssu's) are the **(blocks, residents)**.

[continued on next page]

[1] This is a reasonable approximation as long as all the blocks have close to 40 residents.

(b) [4 pts] Ignoring the clustering and treating this as an SRS of 400 residents, estimate the proportion of cell phone users and its standard error.

[continued on next page]

(c) [6 pts] Now re-estimate the proportion of cell phone users and its standard error, using appropriate cluster sampling methods. *Hint: to reduce calculation, use the fact that in the table above,* $s^2_{p_h} = 0.1214333$.

[continued on next page]

(d) [6 pts] Calculate the design effect DEFF for this design.

(e) [4 pts] Calculate ρ, the correlation between responses from residents on the same block.

3. [24 pts] *Response rates and missing data.* You are completing a telephone survey of an SRS of 1000 members of a much larger professional organization, regarding their level of involvement in support of the organization. Currently the response rate is 80%, with 52.5% of those responding saying they attend every monthly meeting of the local chapter of the organization, and 47.5% saying they do not.

   (a) [8 pts] Does it seem likely that the 200 (20% of 1000) who did not respond to your survey are missing completely at random (MCAR, ignorable missingness) or missing not at random (MNAR, non-ignorable missingness)? Choose MCAR or MNAR and *briefly* explain your reasoning.

   (b) [8 pts] Which of the following is more likely to be correct (circle one):

   - An unbiased estimate of the population proportion that attends every meeting is 52.5%
   - An unbiased estimate of the population proportion that attends every meeting would probably be less than 52.5%.
   - An unbiased estimate of the population proportion that attends every meeting woudl probably be more than 52.5%

[continued on next page]

(c) [8 pts] To get the sample size nearer to the target of $n = 1000$, you could either

- Ask the organization to pay for you to call a new SRS of size 250, hoping that $(0.80)(250) = 200$ people choose to respond *(cost: $200 because it just involves routine single calls to each new phone number)*; or
- Ask the organization to pay for you to followup with the 200 in your original sample who didn't respond yet, to try to get their responses *(cost: $800 because it involves repeated call-backs until each of the 200 non-respondents either responds or refuses)*.

Which do you choose, and why *(briefly)*?

[The space below this line intentionally left blank]

4. [26 pts] *Imputation methods.*

   (a) One method of imputation for missing responses to individual survey items is *mean imputation*.

      i. [5 pts] Explain briefly how mean imputation works.

      ii. [4 pts] Under what assumption (MCAR, MAR, MNAR) is mean imputation OK? (Choose one, no explanation needed.)

      iii. [4 pts] Identify a possible problem with mean imputation.

[continued on next page]

(b) Another method of imputation for missing responses is *hot-deck imputation*.

i. [5 pts] Explain briefly how hot-deck imputation works.

ii. [4 pts] Under what assumption (MCAR, MAR, MNAR) is hot-deck imputation OK? (Choose one, no explanation needed.)

iii. [4 pts] Identify a possible problem with hot-deck imputation.