36-303: Sampling, Surveys and Society
Exam 2
Tue Apr 15, 2008

- 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 15, 2008 Name: _____

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

   (a) [5 pts] You have conducted a survey in which you have collected data on respondents' ages and incomes; all respondents are between 20 and 60 years old, and earn between $25,000 and $250,000. Some cases (respondents) did not report their incomes: older people were less likely to report their income, and among people who did report incomes, older people tended to earn more than younger people.

   Which imputation method below is **not** recommended (or, circle iv. if all are OK)?

   i. For each missing income, fill in the mean of all of the incomes you do have.
   ii. Build a regression model predicting income from age, using the respondents who have answered both questions, and use the model to predict what the missing incomes should be, based on the respondents' ages.
   iii. For each person who did not report income, find all of the other people in the sample of the same age who did report an income, and fill in the average of those persons' incomes.
   iv. All of the above methods are OK.

   (b) [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**.

   (c) [5 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) [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.

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

One of the joys of summer is driving by a roadside stand and buying bunches of fresh sweet-corn to take home and cook. Of course, roadside stands have to satisfy health inspection requirements like all other food sellers.

One roadside stand sells corn in bunches of 3 ears each; they have 580 such bunches for sale. A food inspector samples 12 of these bunches and counts the number of worm fragments he finds on each ear in all 12 bunches, to determine the average number of worm fragments per ear of corn. The inspector's data are as follows:

	Bunch											
	01	02	03	04	05	06	07	08	09	10	11	12
Ear 1	1	4	0	3	4	0	5	3	7	3	4	0
Ear 2	5	2	1	6	9	7	5	0	3	1	7	0
Ear 3	7	4	2	6	8	3	1	2	5	4	9	0
mean	4.33	3.33	1.00	5.00	7.00	3.33	3.67	1.67	5.00	2.67	6.67	0.00

If we treat this data as arising from a one-stage cluster sampling design, the sample variance between cluster means is

$$s^2_{\bar{y}_i} = \frac{1}{12-1}\sum_{i=1}^{1} 2(\bar{y}_i - \bar{y}_{cl})^2 = 4.53$$

and the sample variance of all 36 individual observations is

$$s^2_{y_{ij}} = \frac{1}{36-1}\sum_{i=1}^{12}\sum_{j=1}^{3}(y_{ij} - \bar{y}_{srs})^2 = 7.38$$

(a) [6 pts] Compute $\bar{y}_{cl}$ and $\bar{y}_{srs}$. 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($\overline{y}_{cl}$) and use this to create an approximate 95% confidence interval for the mean number of worm fragments per ear in the entire population of $580 \times 3 = 1740$ ears of corn.

(c) [8 pts] Estimate

- The DEFF (design effect) for this design; and
- The ICC (intra-cluster correlation, ρ).

3. [18 pts] *Yale Work & Life Survey, Part I (3 parts).*

   On September 20th, 2005, the New York Times ran a front-page article, *Many Women at Elite Colleges Set Career Path to Motherhood*, alleging that undergraduate women at elite colleges such as Yale plan to choose motherhood over their careers. Dr. Victoria Brescoll and the Yale Women's Center set out to test this claim with a comprehensive study. A summary of her survey results are available at http://www.yale.edu/wc/worklifesurvey. According to http://www.yale.edu/oir/facts05.html, there were 2,707 men and 2,609 women enrolled at Yale in 2005–2006, the academic year in which Brescoll did her study.

   In Brescoll's survey, 154 men responded, of which 134 planned to become parents someday, and 315 women responded, of which 247 planned to be parents someday[1].

   (a) [6 pts] Treating the sample of men as an SRS w/o replacement, compute $\hat{p}_{men}$, the proportion of men at Yale who plan on becoming parents, and its standard error, SE($\hat{p}_{men}$).

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

[1] Other Likert-type questions more directly addressed the tradeoff between career and family, but we will just work with this yes/no question.

(c) [6 pts] A two-sample *z*-test for whether there is a significant difference between men and women on this question can be based on the test statistic

$$z = \frac{\hat{p}_{men} - \hat{p}_{women}}{\sqrt{\mathrm{SE}(\hat{p}_{men})^2 + \mathrm{SE}(\hat{p}_{women})^2}},$$

which follows a Normal distribution with mean 0 and variance 1 under the Null Hypothesis that the population difference $p_{men} - p_{women}$ is really zero.

Perform this test, using your answers to parts (a) and (b).

What conclusion do you draw?

4. [24 pts] *Yale Work & Life Survey, Part II (4 parts).*

   (a) [6 pts] Combining men and women, and treating the survey as an SRS w/o replacement, estimate the overall proportion of students at Yale who plan on becoming parents. (You do not have to compute the SE).

   (b) [6 pts] We really do not know if Brescoll's survey was an SRS, what the response rate was, etc., but it seems clear there is an imbalance among the respondents between men & women, vs. the Yale student population. Using sex 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 students at Yale who plan on becoming parents (you do not have to compute the SE).

(d) [6 pts] Use the values of the weights to explain[2] the difference between your answers in parts (a) and (c).

[2] 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] A 36-303 group wants to do a survey of student attitudes toward Carnegie Mellon sports teams and sporting events. In their survey proposal they specify the *target population* to be all currently enrolled undergraduates at Cargnegie Mellon, and they indicate that to collect data they will advertise for volunteers to take the survey on FaceBook.com, with a link to a formal survey instrument at QuestionPro.com.

   (a) [6 pts] Did this group specify a *sampling frame* for their survey? If so, say what it is. If not, specify a sampling frame that the group could work with.

   (b) [6 pts] Write down **two** very likely sources of *coverage error* for this survey as currently proposed.

   (c) [6 pts] Make one big suggestion that this group could implement to improve their sampling plan.