36-303: Sampling, Surveys and Society
Exam 1
Thu Feb 21, 2008

- You have 80 minutes for this exam.
- The exam is closed-book, closed notes.
- A calculator is allowed.
- **A formula sheet is provided on the next page 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	18	
3	20	
4	18	
5	24	
Total	100	

Name: ____

Signature: ____

Some Useful Formulas From the Statistics of Survey Sampling

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

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

Then, for example

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

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

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

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

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)).

February 21, 2008 Name: ______________________________

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

   (a) [5 pts] When making a public report on a survey, which of the following is **<u>not</u>** required?

      i. Target population, sampling frame, sampling method, response rates.

      ii. The name of the statistical package used to do the analyses.

      iii. Sample size and precision (SE) of estimates.

      iv. Who sponsored it, who carried it out.

   (b) [5 pts] Recall the sample inclusion indicators $Z_i = 1$ if the i^{th} person in the population is in the sample, and $Z_i = 0$ if not. We showed in class that $Cov(Z_i, Z_j) < 0$. This means that

      i. Seeing a high value (say, income) in the sample makes it more likely that the next value we see will be **<u>low</u>**.

      ii. Seeing a high value (say, income) in the sample makes it more likely that the next value we see will also be **<u>high</u>**.

      iii. If the i^{th} person in the population is in the sample, then it's **<u>more</u>** likely that the j^{th} person will also be in the sample.

      iv. If the i^{th} person in the population is in the sample, then it's **<u>less</u>** likely that the j^{th} person will also be in the sample.

   (c) [5 pts] *Beneficence* is one of the ethical obligations survey researchers have toward survey respondents. The best definition of beneficience, according to our textbook, is

      i. Ensuring that no harm comes to respondents as a result of the survey.

      ii. Ensuring that the survey benefits respondents in some way.

      iii. Acting to minimize possible harms and maximize possible benefits to respondents.

      iv. Making sure that the respondents benefit from the survey by getting a copy of the final report.

   (d) [5 pts] Two important fractions in sample surveys are the *sampling fraction* n/N and the *response rate* r/n (where N is the population size, n is the intended sample size, and r is the number in the sample that actually responded). Which of the following is **<u>not</u>** true, for a simple random sample with replacement from the target population?

      i. To decrease variability in sample estimates, increase the sampling fraction.

      ii. To decrease possible bias in sample estimates, increase the response rate.

      iii. You can force the standard error of sample estimates to be zero by making the sampling fraction large enough.

      iv. You can get a more representative sample by increasing n, regardless of the response rate.

2. [18 pts] *Simple Random Sampling (3 parts).*

   (a) [6 pts] Carefully define *Simple Random Sampling (SRS)* **with** *replacement*, and give an example. Give enough detail that it is obvious that this is a good example.

   (b) [6 pts] Carefully define *Simple Random Sampling (SRS)* **without** *replacement*, and give an example. Give enough detail that it is obvious that this is a good example.

(c) [6 pts] Suppose we take a SRS without replacement of n individuals from a population of N. Let $y_i = 1$ if the respondent feels people should be allowed to talk on their cell phone while driving, and $y_i = 0$ if not. Let

$$\hat{p} \;=\; \frac{1}{n}\sum_{i=1}^{n} Y_i \;=\; \frac{1}{n}\sum_{i=1}^{N} Z_i y_i$$

be the estimated proportion of people who favor cell phone use by drivers from the sample, and let

$$p \;=\; \frac{1}{N}\sum_{i=1}^{N} y_i$$

be the population proportion. Show that $\hat{p}$ is an unbiased estimator for p, i.e., show that

$$E[\hat{p}] \;=\; p\,.$$

3. [20 pts] According to the Carnegie Mellon 2006–2007 FactBook (http://cmu.edu/ira/factbook/facts-2007/facts2007_main.htm), Carnegie Mellon employs 1421 full- and part-time faculty at the Pittsburgh campus (and 38 faculty at the CM-Qatar or CM-West Coast campuses). Your survey team is going to survey Pittsburgh campus faculty about their job satisfaction. One of yout survey questions is

   > *"If you had it to do over again, would you apply for a faculty position at Carnegie Mellon? Yes / No"*

   Answer the following questions (4 parts).

   (a) [5 pts] Let *p* be the proportion of Pittsburgh-based CM faculty that would say yes. Assuming you are doing SRS without replacement, how large must your sample size be, to estimate *p* within a margin of error of ±0.10, with a 95% confidence interval?

(b) [5 pts] It turns out you only have enough resources (time and money to process the responses, mainly) to do an email survey of 75 faculty, chosen at random from a complete email address list for the Pittsburgh-based CM faculty. Amazingly, all 75 faculty reply, and of those, 60 respond "yes" to the question above. Compute a 95% confidence interval for p.

(c) [5 pts] Now suppose that you sent out emails to a random sample of 200 Pittsburgh-based CM faculty, but only 75 replied. Of those 75 replies, 60 respond “yes”. Compute the 95% confidence interval for p.

(d) [5 pts] Which of the confidence intervals, (b) or (c), is harder to believe as evidence about the whole population of Pittsburgh-based CM faculty? Why?

4. [18 pts] There are many situations in which respondents may not report their true behaviors to survey researchers. This typically happens when the behaviors themselves are socially sensitive.

   One researcher wanted to find out how accurately women report having had an abortion. From medical records, she obtains the names and addresses of women wh have had an abortion at a particular clinic, and then sends each of them a letter inviting them to participate. Among other things, the letter states:

   > "The aim of the investigation is the collection of data on women's health and the factors influencing it, as well as how satisfied women are with the organization of medical care and their opinion about ways it might be reorganized. The substantive part of the research is a survey in which data are collected directly from people. Respondents, such as yourself, were selected from the address register by a method of random selection. . . "

   Answer the following questions (3 parts).

   (a) [6 pts] What, if any, ethical principles do you think this study violates? (circle the roman numeral of the one best answer)

      i. Beneficence
      ii. Justice
      iii. Respect for Persons; Confidentiality
      iv. Informed Consent
      v. None of the above; it's just fine.
      vi. Not listed above; I think ______________________________

      ____.

   (b) [6 pts] Suggest a different survey design and/or a different letter to respondents that could answer the initial research question and not raise any ethical flags. (continue onto the next page if needed)

(b) (More space to answer (b), if needed.)

(c) [6 pts] Suppose this was a study of how accurately people report their voting behavior, and the names came from voter registration lists (which are publicly available), but again the respondents were not told the real way their names ended up in the sample.

What, if any, ethical principles do you think this study violates?

i. Beneficence
ii. Justice
iii. Respect for Persons; Confidentiality
iv. Informed Consent
v. None of the above; it's just fine.
vi. Not listed above; I think ______________________________

____ .

5. [24 pts] Below are several survey questions. For each question: (i) indicate a potential problem with the question *using terms we discussed in class on question writing*; (ii) suggest a way to rewrite it (as one or more questions, by providing more information, by improving grammar, etc.) that gets at the same thing while avoiding the problem you raised.

   (a) *"What brand of computer do you own?*

   __ *Dell*

   __ *Apple"*

   i. [3 pts] A Potential Problem:

   ii. [3 pts] Suggestion(s) For Rewrite:

   (b) *"The United States should withdraw its armed forces from Iraq and Afghanistan within one year after the new US President takes office. (Agree or disagree)."*

   i. [3 pts] A Potential Problem:

   ii. [3 pts] Suggestion(s) For Rewrite:

(c) *"In the past week, what fraction of your time online have you spent on instant messaging? Answer here: ______________________"*

i. [3 pts] A Potential Problem:

ii. [3 pts] Suggestion(s) For Rewrite:

(d) *"Given the choice, would you prefer increasing employer contributions to 401(k) plans or increasing employee salaries and raising the investing limits on IRA's?"*

i. [3 pts] A Potential Problem:

ii. [3 pts] Suggestion(s) For Rewrite: