36-303: Sampling, Surveys and Society
Exam 1
Thu Feb 17, 2011

- 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	18	
4	20	
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 17, 2011 Name: ____________________

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

   (a) [5 pts] Following are some of the most important decisions that can be made as part of designing a survey. Which one **doesn't belong**?

      i. How will the potential sample members be identified and selected?
      ii. What approach will be used to contact those sampled, and how much effort will be devoted to trying to collect data from those who are hard to reach or reluctant to respond?
      iii. How much effort will be devoted to evaluating and testing the questions that are asked?
      iv. What mode will be used to pose questions and collect answers from respondents?
      v. None of the above (I.e., all are important decisions in survey design).

   (b) [5 pts] It is possible to construct a representative sample without random sampling, but it is more difficult to argue that it is really representative.

      i. True
      ii. False

   (c) [5 pts] *Beneficence* is a basic ethical principle for research with human subjects. Which of the following is **not** an aspect of beneficence?

      i. Maximize possible benefits and minimize possible harms, to subjects.
      ii. Decide when to do a study because the benefits outweigh the risks to subjects; and when to forego a study because the risks outweigh the benefits.
      iii. Make sure the subjects get the benefit of some compensation for their participation.
      iv. None of the above (I.e., all are aspects of beneficence).

   (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 **not** true, for a simple random sample without replacement from the target population?

      i. You can force the standard error of sample estimates to be zero by making the sampling fraction large enough, as long as the response rate is equal to 1.
      ii. To decrease variability in sample estimates, increase the sampling fraction.
      iii. To decrease possible bias in sample estimates, increase the response rate.
      iv. You can get a more representative sample by increasing n, regardless of the response rate.

2. [18 pts] According to a 2005 survey of American workers reported in the San Francisco Chronicle (http://www.sfgate.com/cgi-bin/article.cgi?f=/g/a/2005/07/11/wastingtime.TMP), the average worker admits to frittering away 2.09 hours per 8-hour workday, not including lunch and scheduled break-time. 44.7% of the respondents polled cited web surfing as their #1 distraction at work. Socializing with co-workers came in second at 23.4%.

   Answer the following questions (3 parts):

   (a) [6 pts] What is the **Target Population** for this survey?

   (b) [6 pts] The "Survey Methodology" section of the article states,

   > America Online and Salary.com conducted in-depth research relating to time wasted at work among 10,044 respondents during May and June, 2005. Populations surveyed included AOL users, Salary.com Salary Wizard users and corporate human resource professionals. Respondents were asked to indicate how much time they wasted in an average workday, assuming a workweek consisting of five 8-hour days. Demographic information, including respondent geography, job category and gender was also collected in the study. Data was analyzed by Salary.com's team of Certified Compensation Professionals.

   Do you think the **coverage error** for this sampling method, for the target population in part (a), is high or low? Give at least one reason for your answer.

[Continued on next page]

(c) [6 pts] Give two other potential problems with the sampling method chosen for this survey. Use words from class to name the problem, and then use specific ideas from class, and/or evidence from the survey methodology paragraph, to explain what you mean.

Problem 1:

Problem 2:

3. [18 pts] *[continuation]* Suppose you wish to replicate this survey, for the Target Population consisting of employees of Carnegie Mellon on the Pittsburgh campus. According to the 2009–2010 CMU FactBook, there are 4732 employees (faculty and staff) on the Pittsburgh campus.

   (a) [6 pts] Suppose you are able to get a simple random sample of 100 CMU Pittsburgh employees, and (remarkably) all of them respond to your survey. You find that 40% of the respondents list web surfing as their #1 distraction. Construct a confidence interval for p, the true proportion of Pittsburgh Campus CMU employees for whom web surfing is the #1 distraction. *(Hint: The SD is $\sqrt{p(1-p)}$.)*

   (You may not need all the space provided.)

[continued on next page]

(b) [6 pts] Suppose you wanted to get a confidence interval in part (a) with a margin of error of 0.05 (5%). What would the sample size need to be, if you were doing simple random sampling *with replacement*?

(c) [6 pts] Suppose you wanted to get a confidence interval in part (a) with a margin of error of 0.05 (5%). What would the sample size need to be, if you were doing simple random sampling *without* *replacement*?

4. [20 pts] Each situation below raises a potential issue about **informed consent**. If you were a member of the IRB, would you approve the research activity, or would you require a change? Justify your "approve" or "disapprove" answer in each case.

   (a) [5 pts] As a part of a face-to-face survey, the interviewer is asked to observe how many books are visible in the respondent's living room.

   __ Approve.

   __ Disapprove. *Please Change the Following:*

   Reason:

   (b) [5 pts] As part of a telephone survey, the first few moments of the interaction between the interviewer and the potential respondent are recorded, for use as examples in teaching other interviewers.

   __ Approve.

   __ Disapprove. *Please Change the Following:*

   Reason:

(c) [5 pts] Before approaching a parent-child pair for a face-to-face interview to study parent-child interations, the interviewer makes structured observations of how the parent and child interact in a public park. The observations are used to select parent-child pairs for the sample, and become part of the data collected for each parent-child pair in the sample.

__ Approve.

__ Disapprove. *Please Change the Following:*

Reason:

(d) [5 pts] As part of a study of whether voter intentions are affected by the sponsor of the survey, some respondents are told that the sponsor is an organization different from the one actually conducting the survey.

__ Approve.

__ Disapprove. *Please Change the Following:*

Reason:

5. [xx pts] Below are several survey questions. For each question: (i) indicate a potential problem with the question *using specific ideas from the lecture notes 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 was your income in the past year?*
   Answer here: ____________________*"*

      i. [3 pts] A Potential Problem:

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

   (b) *"Do you favor or oppose universal health care in the United States?"*

      i. [3 pts] A Potential Problem:

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

(c) *"I am happier than usual these days. (Circle the best answer below)*
1 2 3 4 5
Strongly agree Strongly disagree."

i. [3 pts] A Potential Problem:

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

(d) *"What kind of music do you listen to when you study?*

__ *Rap.* __ *Classic Rock.*
__ *Classical.* __ *Industrial Music.*

i. [3 pts] A Potential Problem:

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