36-303 Sampling, Surveys & Society
Homework 02 Solutions

February 8, 2011

Question 1

(a) Given that the subject matter of this survey may cover more topics that some individuals may find it awkward to discuss, CATI telephone may provide more degree of privacy. Another advantage of CATI telephone is potentially higher response rates since repeated call attempts can be made.

CATI telephone survey is most appropriate for gathering relatively simple informatoin requiring a short close response. Another disadvantage of CATI telephone survey is that some of the Low-income families in rural and inner city areas may not have access to telephones.

Given that the subject matter of this survey may cover more topics that some individuals may find it awkward to discuss, CAPI face-to-face interviews may provide less degree of privacy. Another disadvantage of CAPI face-to-face interviews is that face-to-face interviewing is the most expensive method of gathering data.

CAPI face-to-face interview is a direct and visual contact between the interviewer and the resondent. Another advantage of it is that one can give an incentive for the collaboration.

(b) telephone survey should be the quickest since it saves travel time from face-to-face interview and delivery and response time that of by mail

I think either mail or telephone would be the less costly option. Generally, face-to-face would be more costly because the travel and human hours expense.

I think face-to-face may provide higher response rate among the three. Because mail and random telephone calls can be ignored easily.

face-to-face interview should be the choice for covering populations speaking a language different from the majority since personal attentions can be provided and suitable interviewers can be selected for every group of such people.

mail should be the one with the lowest sampling error because it is easy to mail out to everybody in the household sample. Telephone and face-to-face on the other hand take more human hours to complete.

(c) (i). I think Fence should have the greatest coverage error because the targeted population is undergraduates on CMU campus Pittsburgh and during noon hour anybody from graduate students to visitors can be seen by the fence. Facebook should be next because I don't think there is a way to systematically sample all the undergraduates from CMU. The least one should be Email because all CMU undergraduates have CMU email accounts and they can be sampled from.

(ii) I think the greatest nonresponse error could be Email because if no incentive is provided many undergraduate students may not want to take the time to do the survey. The next should be Facebook because facebook is for connecting with your friends not very suitable for serious work such as filling out a survey. The least one should be Fence because human contact it is harder to refuse.

(iii) Fence should be the one with the greatest measurement error because of the different course schedules of undergraduates. You are more likelily to talk to a certain group of students. Facebook should be next because undergraduates may be under pressure from their friends to response to survey on facebook and answers may be systematically biased. The least one should be Email because every undergraduates have an equal chance to fill the survey.

Question 3

(a)

$$\begin{aligned} E[X_i] &= E[X_1] \quad \text{(because the } X_i\text{'s are iid)} \\ &= 1 \cdot p + 0 \cdot (1-p) \\ &= p \end{aligned}$$

$$\begin{aligned} V[X_i] &= E[X_i^2] - E[X_i]^2 \\ &= E[X_1^2] - E[X_1]^2 \\ &= 1^2 \cdot p + 0^2 \cdot p - p^2 \\ &= p - p^2 = p(1-p) \end{aligned}$$

(b)

$$\begin{aligned} E\left[\sum_{i=1}^{n} X_i\right] &= \sum_{i=1}^{n} E[X_i] \\ &= \sum_{i=1}^{n} E[X_1] \\ &= np \end{aligned}$$

$$\begin{aligned} V\left[\sum_{i=1}^{n} X_i\right] &= \sum_{i=1}^{n} V[X_i] \\ &= \sum_{i=1}^{n} V[X_1] \\ &= np(1-p) \end{aligned}$$

Note that the identity $V\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} V[X_i]$ is true because we know that the random variables $X_1, X_2, ..., X_n$ are independent. This is **not** true in general.

(c)

Using the results from the previous sections,

$$\begin{aligned} E[\hat{p}] &= E\left[\frac{Y}{n}\right] \\ &= \frac{1}{n} E[Y] \\ &= \frac{1}{n} \cdot np \\ &= p \end{aligned}$$

The above property is called "unbiasedness". In this case we say that $\hat{p}$ is an *unbiased* estimator of p.

Again using the results from the previous sections,

$$\begin{aligned} V[\hat{p}] &= V\left[\frac{Y}{n}\right] \\ &= \frac{1}{n^2}V[Y] \\ &= \frac{1}{n^2} \cdot np(1-p) \\ &= \frac{p(1-p)}{n} \end{aligned}$$

Question 4

(a)

$$\begin{aligned} P[X=1] = P[X=1, Y=4] + P[X=1, Y=3] &= \frac{1}{2} \\ P[X=2] = 1 - P[X=1] &= \frac{1}{2} \\ P[Y=3] = P[X=1, Y=3] + P[X=2, Y=3] &= \frac{1}{2} \\ P[Y=4] = 1 - P[Y=3] &= \frac{1}{2} \end{aligned}$$

Now, to prove that X and Y are not independent, we just need to show a counter example:

$$(P[X=1, Y=4] = \frac{1}{8}) \neq (P[X=1] * P[Y=4] = \frac{1}{4})$$

(b)

assuming X and Y are independent, then

$$P[X=x|Y=y] = \frac{P[X=x, Y=y]}{P[Y=y]} = \frac{P[X=x] * P[Y=y]}{P[Y=y]} = P[X=x]$$