36-303 Sampling, Surveys & Society
Homework 02 Solutions

February 14, 2010

1 Question 1

Let's call $\sum_i^n p_{ij} = p_{+j}$ and $\sum_j^n p_{ij} = p_{i+}$. Note that $\Pr[X = x_i] = p_{i+}$ and that $\Pr[Y = y_j] = p_{+j}$. Note also that $\sum_{i=1}^n \sum_{j=1}^n p_{ij} = 1$

a)

$$
\begin{aligned}
E\left[aX + bY + c\right] &= \sum_{i=1}^{n}\sum_{j=1}^{n}\left(ax_i + by_i + c\right)p_{ij} && (1)\\
&= a\sum_{i=1}^{n}\sum_{j=1}^{n} x_i p_{ij} + b\sum_{i=1}^{n}\sum_{j=1}^{n} y_j p_{ij} + c\sum_{i=1}^{n}\sum_{j=1}^{n} p_{ij} && (2)\\
&= a\sum_{i=1}^{n} x_i p_{i+} + b\sum_{j=1}^{n} y_j p_{+j} + c && (3)\\
&= aEX + bEY + c && (4)
\end{aligned}
$$

b)

$$
V\left[aX + bY + c\right] = \underbrace{E\left[(aX + bY + c)^2\right]}_{\alpha} - \underbrace{E\left[aX + bY + c\right]^2}_{\beta} \qquad (5)
$$

$$
\begin{aligned}
\alpha &= a^2\sum_{i=1}^{n}\sum_{j=1}^{n} x_i^2 p_{ij} + b^2\sum_{i=1}^{n}\sum_{j=1}^{n} y_j^2 p_{ij} + c^2 + 2ab\sum_{i=1}^{n}\sum_{j=1}^{n} x_i y_j p_{ij} && (6)\\
&\quad +2ac\sum_{i=1}^{n}\sum_{j=1}^{n} x_i y_j p_{ij} + 2bc\sum_{i=1}^{n}\sum_{j=1}^{n} x_i y_j p_{ij} + 2c^2\sum_{i=1}^{n}\sum_{j=1}^{n} p_{ij}\\
&= a^2EX^2 + b^2EY^2 + 2abEXY + 2acEXY + 2c^2 && (7)
\end{aligned}
$$

$$\beta = a^2E[X]^2 + b^2E[Y]^2 + c^2 + 2abEXEY + 2acEX + 2c^2 \quad (8)$$

Finally,

$$\begin{aligned}
Var[aX + bY + c] &= \alpha - \beta && (9)\\
&= a^2\left[EX^2 - E[X]^2\right] + b^2\left[EY^2 - E[Y]^2\right] && (10)\\
&\quad 2ab\left[EXY - EXEY\right] \\
&= a^2V[X] + b^2V[Y] + 2ab\text{Cov}[X,Y] && (11)
\end{aligned}$$

c)

If X and Y are independent $p_{ij} = p_i \cdot p_j = p_{i+} \cdot p_{+j}$,

$$\begin{aligned}
E[X|Y = y_j] &= \sum_{i=1}^{k} x_i P(X = x_i|Y = y_j) && (12)\\
&= \sum_{i=1}^{k} x_i \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} && (13)\\
&= \sum_{i=1}^{k} x_i \frac{p_{i+} \cdot p_{+j}}{p_{+j}} && (14)\\
&= \sum_{i=1}^{k} x_i \cdot p_{i+} && (15)\\
&= E[X] && (16)
\end{aligned}$$

2 Question 2

2.1 a)

$$\begin{aligned}
E[\hat{\pi}] &= E\left[\frac{\hat{\lambda} - 1/2(1-p)}{p}\right] && (17)\\
&= \frac{1}{p}E\hat{\lambda} - \frac{1}{2p}(1-p) && (18)
\end{aligned}$$

$\hat{\lambda}$ is defined as the fraction of "Yes" answers in the survey, therefore, $n\hat{\lambda} \sim Binomial(n, P(Yes))$ and $E[n\hat{\lambda}] = nP(Yes) = n\lambda$. Thus, $E\hat{\lambda} = \lambda$. Replacing,

$$\begin{aligned}
E[\hat{\pi}] &= \frac{1}{p}\lambda - \frac{1}{2p}(1-p) && (19)\\
&= \frac{\lambda - 1/2(1-p)}{p} && (20)\\
&= \pi && (21)
\end{aligned}$$

2.2 b)

$$V[\hat{\pi}] = V\left[\frac{\hat{\lambda} - 1/2(1-p)}{p}\right] \tag{22}$$
$$= V\left[\frac{\hat{\lambda}}{p}\right] \tag{23}$$
$$= \frac{1}{p^2}V[\hat{\lambda}] \tag{24}$$

¿From the last result,

$$\lim_{p\to 1} V[\hat{\pi}] = \lim_{p\to 1}\frac{1}{p^2}V[\hat{\lambda}] = V[\hat{\lambda}] \tag{25}$$

2.3 c)

Using a normal approximation, a 95% confidence interval for π can be constructed as $\hat{\pi} \pm 2 \cdot se(\hat{\pi})$. Therefore, the width of the interval is $w = 4 \cdot se(\hat{\pi})$. If we want the confidence interval be only 0.02 wide then, $se(\hat{\pi}) = 0.02/4 = 0.005$.

$$se(\hat{\pi})^2 = V[\hat{\pi}] \tag{26}$$
$$= \frac{1}{p^2}V[\hat{\lambda}] \tag{27}$$

Since $n\hat{\lambda} \sim Binomial(n, P(Yes))$,

$$se(\hat{\pi})^2 = \frac{1}{p^2}V[\frac{1}{n}n\hat{\lambda}] \tag{28}$$
$$= \frac{1}{p^2n^2}V[n\hat{\lambda}] \tag{29}$$
$$= \frac{1}{p^2n^2}n(1-\lambda)\lambda \tag{30}$$
$$= \frac{\lambda}{np^2}(1-\lambda) \tag{31}$$

Since we know $p = 1/2$ and we assume $\pi = 0.1$,

$$\lambda = \pi p + \frac{1}{2}(1-p) = 0.3 \tag{32}$$

and

$$se(\hat{\pi})^2 = \frac{4}{n} \times 0.3(1-0.3)$$

Solving for n we get

$$n = \frac{4}{0.005^2} \times 0.3(1-0.3) = 36000 \tag{33}$$

3 Question 3 (Groves Ch2 Q1)

a) The target population is US adults

b)
1. coverage error: The target population is US adults. However, many teenagers also have access to E-mails from their desktop PCs, laptops and handheld devices. There may be ineligible units in the sample.
2. nonresponse error: 47% of people surveyed did not respond to the E-mail solicitation. For example, if the people who did not respond to the survey mostly bought PCs or laptops, the survey could overestimate the expected purchases of handheld digital devices. One of the reasons that the nonresponse rate is high could be due to the automatic spam filtering of E-mail solicitations.
3. measurement error: There may be response bias. For example, people with PDAs, Palm Pilots and other handheld digital devices are more likely to check and respond to E-mails promptly, and thus the sampling frame probably results in an overestimation of the expected purchases of handheld digital devices.

c)
1. coverage error: We could add in the E-mails that only adults are asked to complete the survey.
2. nonresponse error: I think there is a better chance people will respond if there is some kind of incentives for doing the survey. Also, make the E-mails normal so that they won't be automatically filtered by spam filters.
3. measurement error: I think include incentives for doing the survey would help reduce the response bias.

d)
1. sampling error: there is no sampling error because the sampling frame was cut in half and all the samples in the reduced sampling frame are selected.
2. coverage error: all the coverage errors in the original survey design still remains. Also, there is going to be larger undercoverage.

4 Question 4

Groves Ch7, #4

This is the "closed questions with categorical response options" format specified in the book. One of the common problems with this format is that respondents frequently only select the first reasonable answer they see. For example, a respondent may select oil changes but not realizing fluid replacement is usually also a part of oil change nowadays. Also, the question ask respondent to specify one or two which is very vague. I would change it to ask respondents to specify all that applies from the list.

Groves Ch7, #5

Social desirability might induce people not with the truth but with what they think it is socially desirable, introducing measurement error. One way to reduce the effects of social desirability on this question is to deliberate load the wording so that it suggests that many people do not attend religious services for "legitimate" reasons. Another way could be to ask about specific religious services (and perhaps mixing "important" with "not-so-important" ones).

Groves Ch7, #6

a) The "... but not including any activities carried out as part of your job." part appears to be designed to exclude paid domestic service, but it is confusing. Also, it is conceivable that some homemakers assume that their job is to do those chores and therefore answer negatively. It would easier if the question say specifically or give some examples of what type of cases should be excluded.

   Another way to fix the question would be to split it in two:

   - "Have you done any housework, including ..." (and leave out the "but not including" part), and
   - "For which of the previous activities were you paid (hourly or salary)?"

   Then the survey researchers can do the inferential task of figuring out which housework activities they are intersected, rather than the respondent.

b) This is a sensitive question where the respondents might feel compelled to under report. The question might be improved using some of the techniques in section 7.3.7 like forgiving wording or randomized response.

c) It would be easier for the respondents to estimate what is the net amount of money they need each month. The deductions can be calculated or estimated during the processing of the survey.

d) 12 months might be too long a period for a person to remember accurately.