36-303 Sampling, Surveys & Society
Homework 03 Solutions

February 26, 2012

Question 1

(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,

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

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,

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

Question 2

(a)

We can calculate two marginal probabilities by summing over the possible joint probabilities with the other variable.

$$P[X=1] = P[X=1, Y=4] + P[X=1, Y=3] = \frac{1}{8} + \frac{3}{8} = \frac{1}{2}$$
$$P[Y=3] = P[X=1, Y=3] + P[X=2, Y=3] = \frac{3}{8} + \frac{1}{8} = \frac{1}{2}$$

Now, to prove that X and Y are not independent, it suffices to demonstrate that in one instance the joint probability is not equal to the product of marginal probabilities:

$$(P[X=1, Y=3] = \frac{3}{8}) \neq (P[X=1] * P[Y=3] = \frac{1}{4})$$

(b)

assuming X and Y are independent, then

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

Question 3

(a)

$$E[aX + bY + c] = \sum_{x_i=1}^{K} \sum_{y_j=1}^{K} [ax_i + by_j + c]p_{ij}$$

$$= \sum_{x_i=1}^{K} \sum_{y_j=1}^{K} ax_i p_{ij} + \sum_{x_i=1}^{K} \sum_{y_j=1}^{K} by_j p_{ij} + \sum_{x_i=1}^{K} \sum_{y_j=1}^{K} cp_{ij}$$

$$= a \sum_{x_i=1}^{K} \sum_{y_j=1}^{K} x_i p_{ij} + b \sum_{x_i=1}^{K} \sum_{y_j=1}^{K} y_j p_{ij} + c \sum_{x_i=1}^{K} \sum_{y_j=1}^{K} p_{ij}$$

$$= a \sum_{x_i=1}^{K} \sum_{y_j=1}^{K} x_i p_{ij} + b \sum_{y_j=1}^{K} \sum_{x_i=1}^{K} y_j p_{ij} + c \sum_{x_i=1}^{K} \sum_{y_j=1}^{K} p_{ij}$$

where, in the last line, we exchanged the order of summation in the second sum. We can now pull out terms from summations that are constant with respect to the index of summation.

$$= a \sum_{x_i=1}^{K} x_i \sum_{y_j=1}^{K} p_{ij} + b \sum_{y_j=1}^{K} y_j \sum_{x_i=1}^{K} p_{ij} + c \sum_{x_i=1}^{K} \sum_{y_j=1}^{K} p_{ij}$$

Taking note of the following facts about summing joint probabilities:

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

and plugging them in where applicable yields:

$$= a \sum_{x_i=1}^{K} x_i P[X = x_i] + b \sum_{y_j=1}^{K} y_j P[Y = y_j] + c = aE[X] + bE[Y] + c$$

(b)

First we establish two properties of variance and covariance that will help us later; we will show how they follow from the definitions. For any random variables X and Y,

$$Var(X) = E[(X - E[X])^2] = E[X^2 - 2XE[X] + (E[X]^2)]$$
$$= E[X^2] - 2E[X]E[X] + (E[X])^2 = E[X^2] - 2(E[X])^2 + (E[X])^2$$
$$= E[X^2] - (E[X])^2$$

Also,

$$Cov(X,Y) = E[(X - E[X])(Y - E[Y])] = E[XY - XE[Y] - YE[X] + E[X]E[Y]]$$
$$= E[XY] - E[Y]E[X] - E[X]E[Y] + E[X]E[Y] = E[XY] - 2E[X]E[Y] + E[X]E[Y]$$
$$= E[XY] - E[X]E[Y]$$

Keeping the first of these in mind,

$$Var(aX + bY + c) = E[(aX + bY + c)^2] - (E[aX + bY + c])^2$$
$$= E[(aX + bY + c)^2] - (aE[X] + bE[Y] + c)^2$$

where the rightmost expression is simplified using part 3(a). Now we multiply out:

$$= E[a^2X^2 + b^2Y^2 + c^2 + 2abXY + 2acX + 2bcY] -$$
$$(a^2(E[X])^2 + b^2(E[Y])^2 + c^2 + 2abE[X]E[Y] + 2acE[X] + 2bcE[Y])$$
$$= a^2E[X^2] + b^2E[Y^2] + c^2 + 2abE[XY] + 2acE[X] + 2bcE[Y] -$$
$$(a^2(E[X])^2 + b^2(E[Y])^2 + c^2 + 2abE[X]E[Y] + 2acE[X] + 2bcE[Y])$$

Fortunately for us, all the terms of degree one or lower cancel out...

$$= a^2E[X^2] + b^2E[Y^2] + 2abE[XY] - a^2(E[X])^2 - b^2(E[Y])^2 - 2abE[X]E[Y]$$
$$= a^2(E[X^2] - (E[X])^2) + b^2(E[Y^2] - (E[Y])^2) + 2ab(E[XY] - E[X]E[Y])$$
$$= a^2Var(X) + b^2Var(Y) + 2abCov(X,Y)$$

where in the last line we made use of the properties we established initially.

(c)

If X and Y are independent,

$$E[X|Y=y] = \sum_{i=1}^{n} x_i P(X=x_i|Y=y) = \sum_{i=1}^{n} x_i P(X=x_i) = E[X]$$

where the second equality follows from the result in question 2(b) for independent random variables.

Question 4

In mail surveys and some web-based surveys, subjects will have the opportunity to view all questions before they answer any, while in telephone and person-to-person surveys, subjects may answer later questions based on conclusions drawn from earlier questions. Thus, telephone and face-to-face surveys are more likely to produce question-order effects than mail and (some) web-based surveys.

Note that this is a question about **question**-order effects and not **choice**-order effects.

Question 5

- a.(e) The target population is the work force, defined as individuals 16 years and older who are not institutionalized nor in the armed forces. For possible errors, there are many acceptable answers. For example, because this is a sample (60,000 households) and not a census, sampling error is unavoidable. Coverage error is possible but reduced by recruiting via postal mail, so that the sample frame is not simply restricted to people with reachable phones. There could be nonresponse error if selected households are not willing to participate in the voluntary survey.

- a.(f) The mode of data collection is a combination of CATI and CAPI. Households are in sample for eight months; the first and fifth months' interviews are conducted (if at all possible) via personal interview,

with the remaining months' interviews preferably conducted by telephone. See http://www.census.gov/cps/methodology/collecting.html for a more complete exposition.

- b. Unemployment rate is defined as the percentage of the work force (unemployed and employed) people who are deemed unemployed. The definitions of both employed and unemployed exclude those under 16 and in the work force. People are considered unemployed if they do not have a job and have looked for work within the past four weeks. See http://www.bls.gov/cps/cps_htgm.htm#unemployed for a more complete exposition. Many may find this definition inadequate, as people who gave up on finding employment more than a month ago are not deemed to be unemployed or part of the 'work force' at all.