Sensitive Questions and Randomized Response Techniques

Kenneth R. Kundert, University of Wisconsin-Platteville, Platteville, WI

In many situations, researchers attempt to determine characteristics of some population of subjects by surveying a sample of members. When the survey questions are of a sensitive or personal nature, there is a real risk that the results will be inaccurate because of dishonest responses. For example, in studies of the extent of the AIDS epidemic many subjects are reluctant to respond candidly to questions about sexual relations. One surprising approach that has been proposed in these cases involves introducing even more uncertainty into the data! The basic idea is to have subjects consult some random event (such as flipping a coin) before choosing an answer. This has the psychological advantage of masking each subject's true response, and may therefore encourage greater honesty. When the added element of randomness is spread across all subjects, its effects can be estimated. Thus, uncontrolled uncertainty caused by self-conscious subjects is exchanged for controlled uncertainty produced by a simple chance event. This article describes how the controlled uncertainty may be included in the inferential model, and how data may be analyzed to take this uncertainty into account.

Drogin and Orkin [*Vital Statistics*, McGraw-Hill, New York, 1975] consider in detail the problem of estimating sensitive parameters, and propose that an individual's response be based not only on the answer to the question under consideration, but also on the result of some random phenomenon such as the tossing of a coin or the drawing of a card. The following procedure, which guarantees anonymity, requires that a person draw a playing card (which the interviewer does not see) and then respond with either a plus (+) or minus (−) in accordance with the following scheme.

During your married life, have you ever had sex with someone other than your spouse?	Card Drawn	Response to Interviewer
YES	SPADE	+
YES	NON-SPADE	−
NO	SPADE	−
NO	NON-SPADE	+

A (+) response could mean that the individual either did or did not have sex with someone other than his or her spouse. The same is true for the (−) response. Thus, anonymity is preserved. If we let p be the proportion of married people who have had sex with someone other than their spouse, p_+ be the probability of a (+) response, and X_+ be the number of (+) responses in a random sample of size n, then

$$p_+ = P(\text{YES and SPADE}) + P(\text{NO and NON-SPADE})$$

$$= \frac{1}{4}p + \frac{3}{4}(1-p) = \frac{3}{4} - \frac{p}{2}. \tag{1}$$

An estimate of p_+ would be $\hat{p}_+ = X_+/n$, where X_+ can be modeled using a binomial distribution with mean np_+ and variance $np_+(1-p_+)$. Solving equation (1) for p, we obtain $p = 3/2 - 2p_+$. It follows that an estimator of p, $\hat{p}_1$, is given

by

$$\hat{p}_1 = \frac{3}{2} - 2\hat{p}_+ = \frac{3}{2} - 2\frac{X_+}{n}.$$

Since

$$E(\hat{p}_1) = \frac{3}{2} - \frac{2}{n}E(X_+) = \frac{3}{2} - \frac{2}{n}np_+ = \frac{3}{2} - 2\left(\frac{3}{4} - \frac{p}{2}\right) = p,$$

$\hat{p}_1$ is unbiased. Recalling that $\mathrm{Var}(aX + b) = \mathrm{Var}(aX) = a^2\,\mathrm{Var}(X)$, the variance of $\hat{p}_1$ is

$$\mathrm{Var}(\hat{p}_1) = \frac{4}{n^2}\mathrm{Var}(X_+)$$

$$= \frac{4}{n^2}np_+(1-p_+) = \frac{4}{n}\left(\frac{3}{4} - \frac{p}{2}\right)\left(\frac{1}{4} + \frac{p}{2}\right) = \frac{\frac{3}{4} + p(1-p)}{n}. \quad (2)$$

If individuals were asked the question directly, the variance of the unbiased estimator $\hat{p}$, the proportion who said "yes," would be $p(1-p)/n$. The variance (2) is larger since it contains the additional randomness induced by the drawing of a card.

As an example, suppose that a sample of $n = 1000$ individuals produced $X_+ = 650$ (+) responses. An estimate of the proportion of married people who have had sex with someone other than their spouse, $\hat{p}_1$, would be

$$\hat{p}_1 = \frac{3}{2} - 2\frac{X_+}{n} = \frac{3}{2} - 2\left(\frac{650}{1000}\right) = .2$$

and an estimate of the associated variance, $\hat{v}(\hat{p}_1)$, would be

$$\hat{v}(\hat{p}_1) = \frac{3/4 + \hat{p}_1(1-\hat{p}_1)}{n} = \frac{3/4 + (.2)(.8)}{1000} = .00091.$$

The April 24, 1987 issue of *Science* [p. 382 and June 19, 1987, p. 1503] cites a similar method for estimating p. The procedure requires that the person being interviewed toss a coin and then respond with a NO only when the coin comes up tails *and* the individual has not had sex with someone other than his or her spouse. This time, the scheme appears as follows.

During your married life, have you ever had sex with someone other than your spouse?	Coin Face	Response to Interviewer
YES	HEAD	YES
YES	TAIL	YES
NO	HEAD	YES
NO	TAIL	NO

A YES response will not indicate to the interviewer whether the individual did or did not have sex with someone other than the spouse. However, a NO response does indicate a true NO, so this procedure may not be as good as the first one. As with the first method, anonymity is retained. If, as before, p is the proportion of married people who have had sex with someone other than their spouse, p_N is the

probability of the response NO, and X_N is the number of NO responses in a random sample of size n, then

$$p_N = P(\text{NO and TAIL}) = \frac{1}{2}(1-p). \tag{3}$$

Solving equation (3) for p, we have $p = 1 - 2p_N$. An estimator of p, $\hat{p}_2$, is given by

$$\hat{p}_2 = 1 - 2\hat{p}_N = 1 - 2\left(\frac{X_N}{n}\right)$$

where X_N has a binomial distribution with mean np_N and variance $np_N(1-p_N)$. As before, the estimator $\hat{p}_2$ is unbiased. Furthermore,

$$\begin{aligned}\text{Var}(\hat{p}_2) &= \frac{4}{n^2}\text{Var}(X_N) = \frac{4}{n^2}np_N(1-p_N) \\ &= \frac{4}{n}\frac{(1-p)}{2}\frac{(1+p)}{2} = \frac{(1-p)+p(1-p)}{n}.\end{aligned} \tag{4}$$

The variance is again larger than would be the case if the question were asked directly.

Comparing $\text{Var}(\hat{p}_1)$ and $\text{Var}(\hat{p}_2)$ given in equations (2) and (4), we see that for $p < 1/4$, $\text{Var}(\hat{p}_1) < \text{Var}(\hat{p}_2)$. In this case, $\hat{p}_1$ is the better estimator of p. For $p > 1/4$, $\hat{p}_2$ is the better estimator. For p = 1/4, the variances of the two estimators are the same.

The normal approximation to the binomial enables us to use the estimators $\hat{p}_1$ and $\hat{p}_2$ to obtain approximate $(1-\alpha)100\%$ confidence intervals for p (see Mendenhall, *Introduction to Probability and Statistics*, 7th ed., Duxbury, Boston, 1987 for a discussion of interval estimates of p). The appropriate formulas for $\hat{p}_1$ and $\hat{p}_2$, respectively, are

$$\hat{p}_1 \pm Z_{\alpha/2}\sqrt{\frac{\frac{3}{4} + \hat{p}_1(1-\hat{p}_1)}{n}}$$

and

$$\hat{p}_2 \pm Z_{\alpha/2}\sqrt{\frac{(1-\hat{p}_2) + \hat{p}_2(1-\hat{p}_2)}{n}}$$

where $Z_{\alpha/2}$ is the value of the standard normal variable with $(\alpha/2)100\%$ of the area under the curve to its right.

———o———

Power Series and Exponential Generating Functions

G. Ervynck and P. Igodt, Katholieke Universiteit Leuven, Kortrijk, Belgium

The following problem originated during a working session with first year undergraduates:

> *Let $P_k(n)$ denote a kth degree polynomial, with real coefficients, in the variable n. Find*
>
> $$\sum_{n=0}^{\infty} \frac{P_k(n)}{n!}x^n. \tag{1}$$