Random Response Models
or
R-rated Questions of PG Audiences

General Introduction

Example: **(DON'T RESPOND)** How many of you are HIV positive?

Consider two methods of obtaining answers to this question:

- Heads up — no privacy
- Heads on desks — limited privacy

General areas where sensitive questions arise:

- Sexual behavior
- Drug or alcohol use
- Criminal actions
- Political or religous preferences

Probability

Population characteristics:

- Proportion with characteristic of interest is p. (*i.e.* $100p$% HIV positive)

- Proportion without characteristic is $1 - p$.

Sample characteristics:

- Sample size is N.

- Response for the i^{th} person may be represented by

Feature	Symbol	Response "Yes"	Response "No"
Random Variable	X_i	1	0
Probability	Pr	p	$1-p$

Simple Experiments

- Each X_i is called a *Bernoulli trial.*

- If we sum all of the X_i's we get the total number of "Yes" responses. Denoted by $\sum_{i=1}^{n} X_i$.

- The *average response,* $\bar{X} = \sum_{i=1}^{n} X_i/N$, is the proportion of our sample which answered "Yes." It is also our best guess at the underlying proportion, p, of the population with the characteristic of interest.

Example: If $p = 0.75$ and $N = 100$ then

- Observing $\bar{x} = 0.72$ and $\bar{x} = 0.79$ wouldn't be too surprising.

- Seeing $\bar{x} = 0.2$ would be a bit out of the ordinary.

Example: If $p = 0.75$ and $N = 2$ then we can observe 0, 0.5 and 1 and nothing else.

- Repeated sampling gives us multiple estimates of p. The distributions of the sums from these samples are *Binomial.*

- The closeness of the estimates to the true proportion is measured by the *variance* of $\bar{X}$,

$$\text{Var}(\bar{X}) = \frac{p(1-p)}{N}$$

Example: For $p = 0.75$ and $N = 100$ we have variance

$$\begin{aligned}\text{Var}(\bar{X}) &= \frac{0.75 \times 0.25}{100} \\ &= 0.001875\end{aligned}$$

Example: For $p = 0.75$ and $N = 2$ our variance is

$$\begin{aligned}\text{Var}(\bar{X}) &= \frac{0.75 \times 0.25}{2} \\ &= 0.09375\end{aligned}$$

Direct Question Model

To find the proportion of HIV positive individuals we could ask each individual in our sample.

- $Q_S =$ sensitive question
(*e.g.* Are you HIV positive?)

- $p =$ proportion of population with characteristic of interest
  - Estimated by $\bar{X}$
  - Variance is $\text{Var}(\bar{X}) = p(1-p)/N$

We write

$$\begin{aligned}\widehat{p}_{DQ} &= \text{guess for } p \\ &\equiv \bar{X}\end{aligned}$$

and

$$\begin{aligned}\text{Var}\left(\widehat{p}_{DQ}\right) &= \frac{p(1-p)}{N} \\ &= \text{Var}(\bar{X})\end{aligned}$$

Can we trust the responses to Q_S?

Randomized Response Model (Innocuous Question)

Notation:

- Greek letters for known probabilities: α, π
- Latin letters for *unknown* probabilities: r, p, a
- $Q_I =$ innocuous question
  (*e.g.* Is your favorite color blue?)
- $\alpha =$ proportion of population answering "Yes" to Q_I

Randomly assign a person to Q_S or Q_I. The interviewer is *blind* to the question being answered.

Example: Respondent flips a biased coin which the interviewer can't see. If a "Head" then respondent answers Q_S. If a "Tail" then Q_I is answered.

More notation:

- $\pi =$ probability respondent answers Q_S
- $1 - \pi =$ probability respondent answers Q_I
- $r_I =$ probability of a "Yes" response to the complete procedure
- $\bar{X}_I =$ observed proportion of "Yes" responses to the randomized procedure

We know π and α and want to know p.

Our estimate of r_I is $\widehat{r_I} = \bar{X}_I$.

Noting that

$$r_I = \pi p + (1 - \pi)\alpha$$

we have by substitution

$$\widehat{r_I} = \pi \widehat{p_I} + (1 - \pi)\alpha.$$

A little algebra brings us to

- our estimate, $\widehat{p_I} = [\widehat{r_I} - (1-\pi)\alpha]/\pi$, of the proportion of the population with the characteristic of interest

- and our variance

$$\begin{aligned}\text{Var}(\widehat{p_I}) &= \frac{p(1-p)}{N} \\ &+ \left(\frac{1-\pi}{\pi}\right)^2 \left[\frac{\alpha(1-\alpha)}{N}\right] \\ &+ \left(\frac{1-\pi}{\pi}\right) \left[\frac{\pi(1-\alpha)+\alpha(1-\pi)}{N}\right]\end{aligned}$$

Example: Suppose we wish to check for the proportion of virgins within a population. We have $p = 0.75$ (unknown), $\alpha = 0.5$ (known), $\pi = 0.6$ (known) and $N = 100$ (known).

If the sample represents the population exactly, then

- 60 answer Q_S (45 are virgins)
- 40 answer Q_I (20 like blue)

Hence, $\bar{X}_I = \#$ "Yes" $= \frac{20+45}{100} = 0.65 = \widehat{r_I}$

Thus, $\widehat{p_I} = \frac{0.65-0.4\times0.5}{0.6} = 0.45/0.6 = 3/4$

Audience Participation Example

Our questions:

- $Q_S =$ "Have you ever shoplifted?"
- $Q_I =$ "Was the card you cut a heart or a spade?"

Method: Respond to question

- Q_S if cut card is A–3 or 8–10
- Q_I if cut card is 4–7

Because the face cards were removed we have $\pi = 0.6$.

Since there are four suits in a deck we have $\alpha = 0.5$.

We compute

- $\bar{X}_I = \sum X_i / N =$

- $\widehat{p_I} = \frac{\bar{X}_I - (1-0.6)0.5}{0.6}$
  Which leads to $\frac{\bar{X}_I - 0.2}{0.6} =$

Although we can't find the actual variance of this estimate we can get an estimate of the variance.

$$\begin{aligned}\text{Var}(\widehat{p_I}) &= \frac{p(1-p)}{N} \\ &\quad + \left(\frac{1-0.6}{0.6}\right)^2 \left[\frac{0.5(1-0.5)}{N}\right] \\ &\quad + \left(\frac{1-0.6}{0.6}\right) \left[\frac{0.6(1-0.5)+0.5(1-0.6)}{N}\right]\end{aligned}$$

If we cheat and substitute $\widehat{p_I}$ for p we get

$$\begin{aligned}\text{Var}(\widehat{p_I}) &\approx \frac{\widehat{p}(1-\widehat{p})}{N} \\ &\quad + \frac{4}{9}\left[\frac{0.25}{N}\right] \\ &\quad + \frac{2}{3}\left[\frac{0.5}{N}\right] \\ &=\end{aligned}$$

Associated Works

- Zellner, A. "An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias," JASA June 1962, 348–368.

- Warner, S. "Randomized responses: a survey technique for eliminating evasive answer bias," JASA March 1965, 63–69.

- Abul-Ela, A., Greenberg, B., Horvitz, D., "A multiproportions randomized response model," JASA Sept. 1967, 990–1008.

- Gould, A., Shah, B., Abernathy, J., "Unrelated question randomized response techniques with two trials per respondent," Proceedings of the Social Statistics Section of the ASA 1969.

- Warner, S., "The linear randomized response model," JASA December 1971, 884–888.

- Cambell, C. and Joiner, B., "How to get the answer without being sure you've asked the question," The American Statistician, Dec. 1973, 229–231.

- Dowling, T. and Shactman, R., "On the relative efficiency of randomized response models," JASA March 1975, 84–87.