# Wednesday, Mar 1

## The Challenge of Sensitive Questions

Consider a survey with human respondents answering sensitive questions.

1. **Sampling bias**: Failure to account for the fact that some units are more or less likely to be included in the sample.

2. **Non-response bias**: Failure to observe some observational units that were intended to be observed. *Respondents may be less inclined to participate in a survey with sensitive questions.*

3. **Response bias**: Errors in observation/measurement of the variable of interest. *Respondents may be more inclined to be dishonest in response to sensitive questions.*

When a survey uses a *sensitive question*, respondents are more likely to (a) not respond and (b) lie if they do respond. Thus sensitive questions can produce *non-response* and *response* bias.

## Randomized Response Method (Unrelated Question)

**Instructions**: Please roll the die but **do not** tell me the result or show it to anyone else. If the die came up 1-4 answer *Question A*, but if the die came up 5-6, answer *Question B*. Answer by stating "yes" or "no" but **do not** tell me which question you are answering.

**Question A**:

**Question B**:

We define the following three probabilities.

1. $\theta$ is the probability of getting question A. In the example above $\theta = 2/3$.

2. $p_a$ is the probability of answering "yes" if asked question A. It is unknown.

3. $p_b$ is the probability of answering "yes" if asked question B. In the example above $p_b = 0.5$.

Our goal is to *estimate $p_a$* from $n$ responses to the randomized response procedure. There are four outcomes, and we can write the probability of each as follows.

| Question | Response | Probability |
|:---:|:---:|:---:|
| A | no | $\theta(1 - p_a)$ |
| A | yes | $\theta p_a$ |
| B | no | $(1 - \theta)(1 - p_b)$ |
| B | yes | $(1 - \theta)p_b$ |

The probability of getting a response of "yes" from a respondent is

$$p_y = \theta p_a + (1 - \theta)p_b.$$

Thus the probability of a response of "yes" to Question A is

$$p_a = \frac{p_y - (1 - \theta)p_b}{\theta}.$$

This suggests we estimate $p_a$ as

$$\hat{p}_a = \frac{\hat{p}_y - (1 - \theta)p_b}{\theta},$$

where $\hat{p}_y$ is the proportion of observations (out of $n$) where a respondents responds "yes".

**Example**: Please roll the die but **do not** tell me the result or show it to anyone else. If the die came up 1-4 answer *Question A*, but if the die came up 5-6, answer *Question B*. Answer by stating "yes" or "no" but **do not** tell me which question you are answering.

**Question A**: Have you ever been unfaithful to your partner?

**Question B**: Did the die come up even?

Suppose that out of 1000 respondents, 300 responded "yes".

$$\hat{p}_a = \frac{0.3 - (1 - 2/3)0.5}{2/3} = 0.2.$$

What about a margin of error? The margin of error formula is

$$z\frac{1}{\theta}\sqrt{\frac{\hat{p}_y(1 - \hat{p}_y)}{n}}.$$

So our margin error for the above example (using a confidence level of 95%)

$$1.96\frac{1}{2/3}\sqrt{\frac{0.3(1 - 0.3)}{1000}} \approx 0.04.$$

So we estimate that the proportion of people in the population that have been unfaithful to their partner is $0.2 \pm 0.04$.

## Randomized Response Method (Mirrored Question)

**Instructions**: Please roll the die but **do not** tell me the result or show it to anyone else. If the die came up 1-4 answer *Question A*, but if the die came up 5-6, answer *Question B*. Answer by stating "yes" or "no" but **do not** tell me which question you are answering.

**Question A**:

**Question B**:

We define the following three probabilities.

1. $\theta$ is the probability of getting question A. In the example above $\theta = 2/3$.

2. $p_a$ is the probability of answering "yes" if asked question A. It is unknown.

3. $p_b$ is the probability of answering "yes" if asked question B. It is unknown.

Our goal is to *estimate* $p_a$ from $n$ responses to the randomized response procedure. There are four outcomes, and we can write the probability of each as follows. But note that here $p_b = 1 - p_a$.

| Question | Response | Probability |
|----------|----------|-------------|
| A | no | $\theta(1 - p_a)$ |
| A | yes | $\theta p_a$ |
| B | no | $(1 - \theta)p_a$ |
| B | yes | $(1 - \theta)(1 - p_a)$ |

The probability of getting a response of "yes" from a respondent is

$$p_y = \theta p_a + (1 - \theta)(1 - p_a).$$

Thus the probability of a response of "yes" to Question A is

$$p_a = \frac{p_y + \theta - 1}{2\theta - 1}.$$

This suggest we estimate $p_a$ as

$$\hat{p}_a = \frac{\hat{p}_y + \theta - 1}{2\theta - 1},$$

where $\hat{p}_y$ is the proportion of observations (out of $n$) where a respondents responds "yes".

**Example**: Please roll the die but **do not** tell me the result or show it to anyone else. If the die came up 1-4 answer *Question A*, but if the die came up 5-6, answer *Question B*. Answer by stating "yes" or "no" but **do not** tell me which question you are answering.

**Question A**: I have been unfaithful to my partner.

**Question B**: I have not been unfaithful to my partner.

Suppose that out of 1000 respondents, 400 responded "yes".

$$\hat{p}_a = \frac{0.4 + 2/3 - 1}{2(2/3) - 1} = 0.2.$$

What about a margin of error? The margin of error formula is

$$z \frac{1}{|2\theta - 1|} \sqrt{\frac{\hat{p}_y(1 - \hat{p}_y)}{n}}.$$

So our margin of error for the above example (using a confidence level of 95%) is

$$1.96 \frac{1}{|2(2/3) - 1|} \sqrt{\frac{0.4(1 - 0.4)}{1000}} \approx 0.09.$$

So we estimate that the proportion of people in the population that have been unfaithful to their partner is $0.2 \pm 0.09$.

## Design Considerations

1. The *method* matters. Although the unrelated question and mirrored question methods produced the same point estimate, they did not have the same margin of error.

2. The *value of* $\theta$ matters. What happens if we increase it?

**Example**: Please roll this **20-sided** die but **do not** tell me the result or show it to anyone else. If the die came up 1-18 answer *Question A*, but if the die came up 19-20, answer *Question B*.

**Question A**: "Have you ever been unfaithful to your partner?"

**Question B**: "Did the die come up even?"

Note that here the probability of getting question A is $\theta = 0.9$.