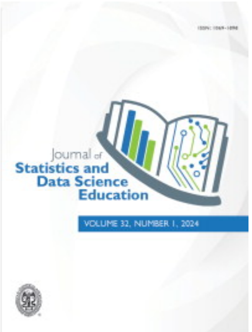

Journal of Statistics and Data Science Education

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/ujse21

Investigating Sensitive Issues in Class Through Randomized Response Polling

Christian Genest, James A. Hanley & Sahir R. Bhatnagar

To cite this article: Christian Genest, James A. Hanley & Sahir R. Bhatnagar (08 Feb 2024): Investigating Sensitive Issues in Class Through Randomized Response Polling, Journal of Statistics and Data Science Education, DOI: 10.1080/26939169.2024.2302179

To link to this article: https://doi.org/10.1080/26939169.2024.2302179

© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC

View supplementary material

Published online: 08 Feb 2024.

Submit your article to this journal

Article views: 172

View related articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=ujse21

JOURNAL OF STATISTICS AND DATA SCIENCE EDUCATION
2024, VOL. 00, NO. 0, 1–7
https://doi.org/10.1080/26939169.2024.2302179

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Investigating Sensitive Issues in Class Through Randomized Response Polling

Christian Genest[a], James A. Hanley[b], and Sahir R. Bhatnagar[c]

[a]Department of Mathematics and Statistics, McGill University, Montréal, Québec, Canada; [b]Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montréal, Québec, Canada; [c]Five Prime Sciences, Westmount, Québec, Canada

ABSTRACT
This article provides an introduction to randomized response polling, a technique which was designed to allow for questioning on sensitive issues while protecting the respondent's privacy and avoiding social desirability bias. It is described in terms that are suitable for presentation and use in any classroom environment. Instructions for plain users are included, along with the results of a small in-class implementation. The underpinnings of the method, which are laid out for the statistically savvy, illustrate the tradeoff between data acquisition and privacy protection. A few suitable references are also included for those who wish to dig further into the subject. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received October 2021
Accepted November 2023

KEYWORDS
Poll; Privacy; Randomization; Sensitive question; Social desirability bias

1. Introduction

As documented by the American social psychologist Leon Festinger (1919–1989) in his social comparison theory, human beings have an innate drive to evaluate themselves and to compare themselves with others. This is especially true of teenagers, for whom measuring themselves up to others is an important part of their identity formation. At the same time, however, they are often reluctant to share personal information, as they dread a negative evaluation, which could compromise their position in a group and affect their well-being.

Schoolmates are a natural population to which students tend to compare themselves, not only academically, but in matters of preference, beliefs, and life habits. Polling in the classroom is a tool that educators can use to open the dialogue and instructors also resort to it for other pedagogical reasons, for example, to ask participants about their personal opinions or experiences. In general, however, respondents only feel comfortable participating in such activities and providing honest answers if anonymity is preserved. This is especially so for sensitive topics such as sex, drugs, or mental health.

In principle, anonymous online polls are one way to elicit information in a class about sensitive topics, but given a growing sense of distrust of technology in society, such an initiative is likely to be met with skepticism and lukewarm participation. Starting from the principle that there is no better way to instill confidence in respondents than to give them the means to protect their own privacy, a randomized response design seems very well suited for the task. This method, due to Warner (1965), consists of asking each participant to answer a question selected individually using a personal randomization device.

In a randomized response design, the pollster does not know which question has been answered. Therefore, the respondent's privacy is protected and an honest answer can be freely consented to a sensitive question such as "Did you cheat on the exam?" While this polling technique makes it impossible to tell whether any given individual cheated or not, it can be used to derive an estimate of the number or proportion of respondents who did.

For questions in which students themselves have a personal interest, for example, the number or proportion of them who had sex, took drugs or used ChatGPT to write an essay, this approach may encourage participation in an in-class poll, especially if it is carried out live. In addition to satisfying their curiosity, such an experience could spark their interest in statistics and show its practical use.

The purpose of this article is to describe the randomized response design in terms that are suitable for presentation in just about any class. The focus is on concepts and the simplest possible analysis is presented, so that users without a statistical background can implement the approach and make sense of the results. Various teaching points for statistics instructors are also highlighted along the way. A distinctive feature of the setup is that the class is the *population* of interest, not a *sample*; the poll is then a *census* with an element of randomness stemming from the uncertainty as to which question each respondent answered.

The randomized response design is introduced in Section 2 and a simple way of estimating the number of respondents with the trait/behavior is described in Section 3. By necessity, this design cannot provide an exact answer even if everyone

CONTACT Christian Genest christian.genest@mcgill.ca Department of Mathematics and Statistics, McGill University, 805, rue Sherbrooke ouest, Montréal, Québec, H3A 0B9, Canada.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/ujse.
© 2024 The Author(s). Published with license by Taylor and Francis Group, LLC.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

co-operates and truthfully answers the question put to them. Therefore, a way to compute a confidence interval reflecting the estimation error is given in Section 4. It is further shown in Section 5 how a number of independent repeats of the survey can narrow the margin of error and improve precision. Section 6 reports a small class experience with the randomized response design, and clues for further reading can be found in Section 7. Additional material is provided in an Online Supplement, including instructions for running an online poll, and R code for the main figures.

2. The Notion of Randomized Response

The notion of a randomized response design was first proposed by Stanley Warner (1965).[1] His work was motivated by the desire to allow respondents to provide an honest "Yes" or "No" answer to a question about a sensitive issue while protecting their privacy. In Warner's original work, chance decided, unknown to the interviewer, whether the interviewee would answer the direct question or the same question in the negative. This way, the meaning of a respondent's "Yes" could not be decoded.

Many variants and extensions of Warner's approach were developed over the years; see Blair, Imai, and Zhou (2015) for a review. For simplicity, attention is restricted here to Warner's original design. In contrast to most applications of this design, however, the respondents are regarded as the *entire* population, rather than as a sample from a larger population of interest. For simplicity also, the focus is limited at first to cheating on an exam.

Thus, given a class attended by N students, the quantity of interest to all present is the number n of them who cheated. The question and its negative form are thus:

Question A: Did you cheat on the exam?
Question B: Were you honest on the exam?

Each respondent then determines privately whether they will answer Question A or Question B. The determination is made individually and at random in such a way that

Question A is selected with probability p;
Question B is selected with probability $1-p$.

The value of $p \in [0,1]$ is agreed in advance and the same for all participants in the poll. If $p = 3/4$, say, the respondent could report "Yes" or "No" in an anonymous in-class online poll after executing the following R commands:

```
p = 0.75
question <- c("Did you cheat on the exam?",
     "Were you honest on the exam?")
sample(question, 1, prob = c(p, 1 - p))
```

This procedure ensures complete anonymity of the responses because, unless p is 0 or 1, nobody except the respondent knows which question they answered. A "Yes" could mean that the respondent answered Question A and cheated on the exam, or that this respondent answered Question B and was honest on the exam. A similar ambiguity exists in interpreting a "No."

As will be seen, this procedure makes it possible to (i) estimate very easily the proportion of participants who cheated on the exam and (ii) control the precision of this estimate, so long as the participants genuinely choose at random between Question A and Question B with probabilities p and $1-p$, respectively. For statistics instructors, this setting also offers an opportunity to discuss and illustrate many statistical concepts, as highlighted below.

3. Estimation

The data from a single online poll consist of a value X, namely the number of "Yes" responses. The value of X (the statistic) is to be distinguished from the unknown number n of participants who cheated on the exam (the parameter). However, n can be estimated easily from X using the method of moments (MoM), as detailed next. Heuristics for students unfamiliar with this method are provided in the Online Supplement.

Whatever the value of $n \in \{0,\ldots,N\}$, one can write $X = X_1 + \cdots + X_N$ as a sum of mutually independent Bernoulli random variables, that is, variables taking the value 1 if a respondent's answer is "Yes" and 0 otherwise. However, the variable X is not binomial unless $p = 1-p = 1/2$, because $\Pr(X_i = 1)$ varies with $i \in \{1,\ldots,N\}$. Indeed, $\Pr(X_i = 1) = p$ for a cheater while $\Pr(X_i = 1) = 1-p$ for an honest student. For this reason, the variables $X_1,\ldots,X_N$ are called Poisson trials; see p. 130 of Feller (1968).

One can write X as the sum of two independent binomial random variables, viz.

$$X = V_n + W_{N-n},$$
$$V_n \sim \mathcal{BIN}(n,p),$$
$$W_{N-n} \sim \mathcal{BIN}(N-n, 1-p).$$

Thus, $\mathrm{E}(X) = \mathrm{E}(V_n) + \mathrm{E}(W_{N-n}) = np + (N-n)(1-p)$ or, equivalently,

$$\mathrm{E}(X) = n(2p-1) + N(1-p), \tag{1}$$

from which an MoM estimate of n is obtained by replacing $\mathrm{E}(X)$ by its observed value, X, and solving for n in (1). This yields

$$\hat{n} = \frac{X - N(1-p)}{2p-1}, \tag{2}$$

provided that $p \neq 1/2$. The value $1/2$ must be avoided because X is then simply binomial with parameters N and $p = 1/2$, and thus provides no information about n.

Summary for the plain user: *Explain the procedure involving two questions, one of which is selected at random by each respondent. Agree with the group on a value of p other than 0, $1/2$ or 1. Run the survey and count the number, X, of "Yes" responses. Compute $\hat{n}$ using (2). This is your estimate of the number of cheaters in the group.*

[1] For more information about the life and works of Stanley L. Warner (1928–92), refer to vol. 21.1 (1995) of the bilingual journal *Survey Methodology/Techniques d'enquête* published by Statistics Canada.

Discussion points that could be brought up at this stage include curtailing the MoM estimate to ensure that $\hat{n}$ is an integer from 0 to N. However, the gain in interpretability involves more complex variance expressions.

In an advanced statistics class, students could be asked to determine the maximum likelihood estimator of n. As the parameter space is discrete, they often find this task difficult, and there is no algebraic expression for the solution. In contrast, a numerical search is relatively easy and R code to this end is provided in the Online Supplement. It could be used to explore how much more efficient this estimator would be than (2). For simplicity, however, the rest of this article concentrates on the MoM estimator, for which explicit calculations illustrate the determinants of the margin of error.

4. Margin of Error

While the poll yields an estimate of the number of cheaters in the group, it is crucial to understand that this estimate is not the value of n itself. Indeed, if the poll were rerun, one would likely get another estimate, simply because every participant would proceed to choose Question A or Question B through randomization and may not necessarily end up answering the same question as in the first poll.

This variability can be quantified by looking at the variance of X, the number of "Yes" responses in the poll. Given that the variables V_n and W_{N-n} are binomial and independent,

$$\begin{aligned}\operatorname{var}(X) &= \operatorname{var}(V_n) + \operatorname{var}(W_{N-n}) \\ &= np(1-p) + (N-n)p(1-p) \\ &= Np(1-p),\end{aligned}$$

which is the same as the variance of a single binomial random variable with N trials and success probability p. Therefore,

$$\operatorname{var}(\hat{n}) = \frac{1}{(2p-1)^2}\operatorname{var}(X) = \frac{Np(1-p)}{(2p-1)^2}. \tag{3}$$

It is noteworthy that this expression does not depend on the parameter of interest, n, but that it is only a function of known quantities, namely the class size, N, and the constant p. The expression is also symmetric about $p = 1/2$.

In a calculus-based statistics course, questions for discussion might include:

a) For what value(s) of p is the variance the smallest?
b) Is this value or are these values feasible? If not, what might be a good compromise?

The answer to the first question is $p = 0$ or 1 but in both cases the randomization would be removed; values near 0 or 1 mean that there is a high probability that nearly everyone taking the survey will be asked the same question. The graph of the map $p \mapsto p(1-p)/(2p-1)^2$ sketched in Figure 1 suggests that while $\operatorname{var}(\hat{n})$ is decreasing in $p \in (1/2, 1)$, what would be gained past $p = 3/4$ might be offset by increasing reluctance to participate, or the temptation to be dishonest in responding to the question.

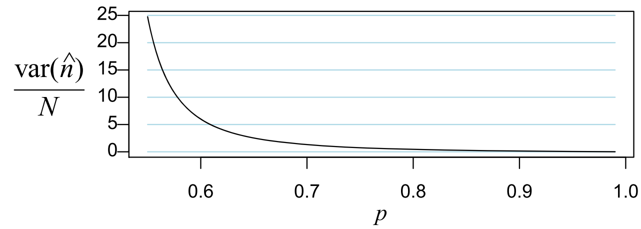

Figure 1. Graph of $\operatorname{var}(\hat{n})/N$, given in (3), as a function of $p \in (1/2, 1)$.

Approximate 95% confidence intervals for the number, n, of cheaters and their proportion, n/N, in the group can be found. Their endpoints are respectively given by the point estimate plus or minus the corresponding margin of error, viz.

$$\hat{n} \pm 2 \times \sqrt{\frac{Np(1-p)}{(2p-1)^2}}, \quad \frac{\hat{n}}{N} \pm \frac{2}{N} \times \sqrt{\frac{Np(1-p)}{(2p-1)^2}}, \tag{4}$$

where the multiplicative factor 2 in front of the square root is a rounded-up value of the familiar 97.5th quantile of the $\mathcal{N}(0,1)$ Gaussian distribution, that is, $z = 1.95996\ldots$

One can check that if $p = 0.75$ is used, the margin of error for the estimate of n simplifies to $1.73 \times \sqrt{N}$. It would be reduced to $1.33 \times \sqrt{N}$ if $p = 0.8$, but such a high value would likely be considered to provide insufficient privacy protection.

Summary for the plain user: *For the value of $p \in (0,1)$ used in the poll, which is such that $p \neq 1/2$, compute ME = $2\sqrt{Np(1-p)/(2p-1)^2}$. Just as in election polls, one can then assert that the margin of error is ± ME individuals at the 95% confidence level.*

As described in the Online Supplement, one might explain to statistics students that the Normal approximation is based on the Central Limit Theorem. In more advanced classes, this could also be checked. When the class size, N, is small, an argument of the type $N \to \infty$ is not compelling. One may then wonder how good the approximation is, and this is yet another issue that could be investigated numerically in class.

As an illustration, Figure 2 shows how the distribution of X varies when $p = 3/4$ and either $n/N = 0.75$ (left) or $n/N = 0.95$ (right) when $N \in \{20, 40, 80, 160\}$. In spite of a slight asymmetry, the Normal approximation is quite serviceable. The figure makes it clear that the Central Limit Theorem kicks in very quickly. For alternative approximations and algorithms, including R code, to improve on the Normal approximation, see, for example, Butler and Stephens (2017) and Liu and Quertermous (2018).

While it may be suboptimal, the approximate margin of error appearing in (4) has the merit of simplicity. It is also instructive on at least two counts:

a) The interval has the same width whatever the value of n, although one could choose, to enhance interpretability, to curtail it when the estimate is close to 0 or N to avoid it extending beyond the interval $[0, N]$.
b) Given that the length of the interval is a function of constants that are known in advance, the compromise that must be struck between privacy protection and precision of the estimate is easy to see.

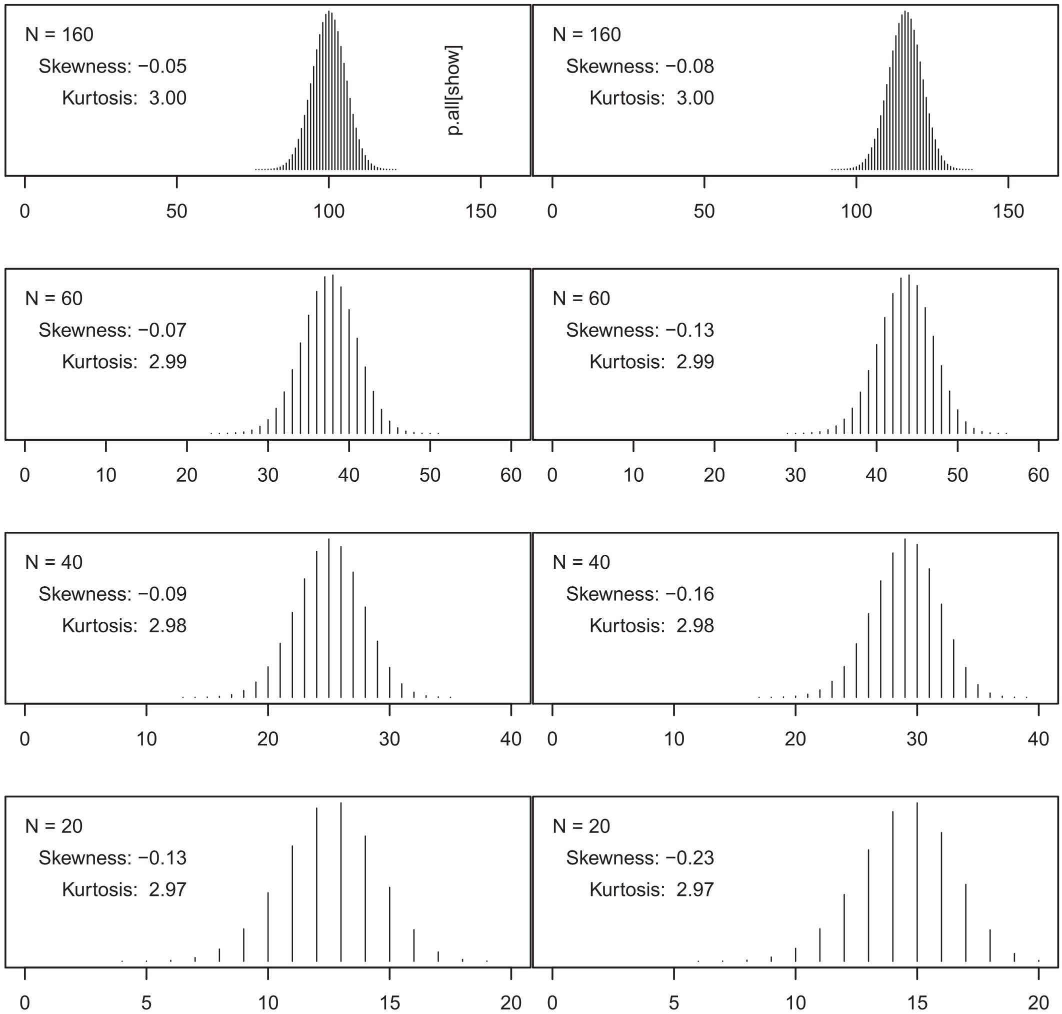

Figure 2. Frequency distribution of X for $N \in \{20, 40, 80, 160\}$ when $p = 3/4$ and either $n/N = 0.75$ (left) or $n/N = 0.95$ (right).

Point a) is clearly illustrated by Figure 3, called a nomogram, in which the black curve shows how the estimated proportion, $\hat{n}/N$, of cheaters varies as a function of the observed fraction, X/N, of respondents who answered "Yes." As seen from (2), this relationship depends only on p, and the slope of the non-flat part of the black curve is $1/(2p-1)$, which equals 2 when $p = 3/4$, as in the figure.

While the black curve in Figure 3 does not depend on the population size, N, it is clear from (4) that the approximate 95% margin of error associated with $\hat{n}/N$ is proportional to $1/\sqrt{N}$. The lower and upper limits of the corresponding confidence interval are shown in red and blue, respectively. For example, if $N = 160$, and $100\,X/N = 65\%$ of participants answered "Yes," the estimated proportion of cheaters is 0.8 or 80%. With this sample size, the lower and upper limits of the confidence interval are approximately 66% and 94%, respectively, so the margin of error is ± 14 percentage points.

To illustrate point b), suppose that it is desired to have a margin of error, k, for $\hat{n}$, that is, one would like to state (with 95% assurance) that the number of cheaters is $\hat{n} \pm k$. Because the same value of k has different implications for different values of N, one might recast the planning exercise as one where the desired margin of error for $\hat{n}$ is some specified fraction, f, of N. Denote by z the multiple (1.96 or 2) of the standard error that is typically used. Given that N is fixed, one must thus choose p in advance so that

$$z\sqrt{N}\sqrt{\frac{p(1-p)}{(2p-1)^2}} = f \times N.$$

High-school students could be asked to check that the only solution of this quadratic equation in p in the interval $(1/2, 1)$ is

$$p_{N,f} = \frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{1+4Nf^2/z^2}}. \tag{5}$$

Summary for the plain user: *For a class of size N, the constant $p_{N,f}$ gives the value of p that produces a margin of error that is a specified fraction, f, of N. By computing p for different values of f, one can get a sense of how a specific demand on the precision of the estimated count of cheaters in the class affects the level of privacy protection provided to the individual respondents. The smaller f is, the closer p is to 1, the less privacy there is, because a larger and larger proportion of individuals will answer Question A.*

Even with a proper introduction to the notion of randomized response, most participants in the poll would likely not perceive their response as confidential if p is too close to 1. For various class sizes, Figure 4 shows the tradeoffs between feasibility and precision.

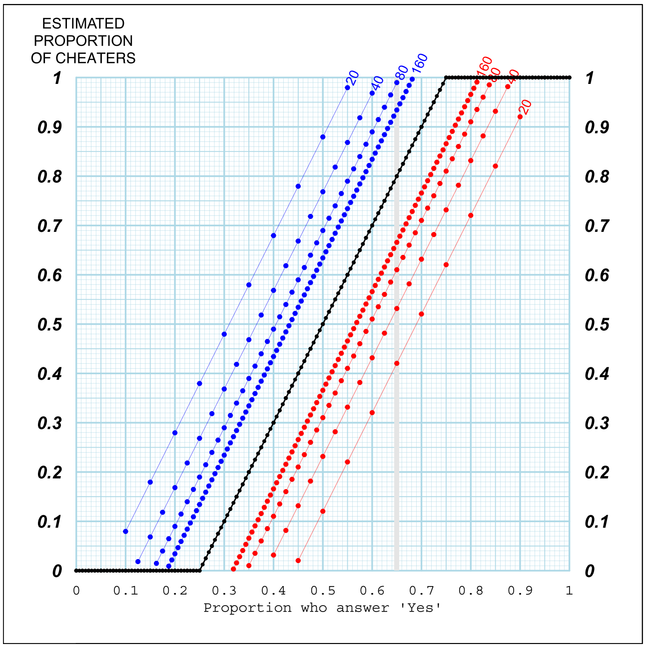

Figure 3. Nomogram showing, for various class sizes $N \in \{20, 40, 80, 160\}$, how the estimated proportion, $\hat{n}/N$, of cheaters (vertical axis) varies as a function of the observed proportion, X/N, of "Yes" (horizontal axis) responses to an anonymous online poll. It is based on Warner's version of the randomized response design, in which the probability of being asked "Did you cheat on the exam?" is $p = 0.75$. The lower and upper limits of the large-sample 95% confidence intervals are shown in red and blue, respectively.

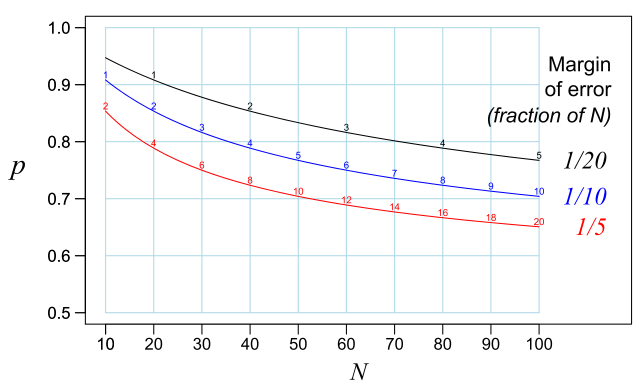

Figure 4. The probability p defined in (5) as a function of N, when the margin of error for $\hat{n}$ is a specified fraction, f, of N. The absolute magnitudes of selected margins of error are displayed as the integers 1, 2, 3, and so on.

5. Repeating the Poll

The wide margin of error is the price paid for introducing an additional level of privacy protection through the randomized response mechanism. However, if respondents are willing to repeat the poll, the effect of the "noise" stemming from the randomization process can be reduced without compromising confidentiality. With online polling facilities available, most groups would likely be willing to go through the exercise a few times.

Suppose that R repeats are considered. In the present context, repeated polls also provide a unique opportunity to witness the effect of random sampling because the values $X_1, \ldots, X_R$ are mutually independent by virtue of the randomized response mechanism, and this even though they involve the very same respondents. Let $\bar{n} = (\hat{n}_1 + \cdots + \hat{n}_R)/R$ be the mean estimate of the number of cheaters over the R trials. It is important here to use the raw estimates, not corrected estimates modified to be between 0 and N. Invoking the mutual independence between the estimates $\hat{n}_1, \ldots, \hat{n}_R$, one finds that

$$\text{var}(\bar{n}) = \frac{N}{R}\,\frac{p(1-p)}{(2p-1)^2}\,. \qquad (6)$$

If, for example, the participants were willing to take part in $R = 4$ independent polls, one could halve the margin of error because $1/\sqrt{R} = 1/2$ in that case.

Given the interest in knowing the answer with a reasonable margin of error, $\pm k$, while ensuring a high degree of privacy, one could illustrate the compromise that must be struck between the two objectives in yet another way using the number, R, of repeats.

Specifically, given a population size, N, one could ask the group to fix two parameters among p, k, and R, and then discuss the effect on the third, knowing that for an approximate 95% confidence interval, one has

$$2\sqrt{\frac{N}{R}}\sqrt{\frac{p(1-p)}{(2p-1)^2}} = k.$$

A more ambitious statistics instructor could also bring in the effect of the degree of confidence $1-\alpha$, here taken to be 95% throughout, by replacing the factor 2 by $z_{\alpha/2}$, where $z_\alpha = \Phi^{-1}(1-\alpha)$ is the αth upper quantile of the $\mathcal{N}(0,1)$ distribution.

6. A Classroom Experience

Warner's polling mechanism was tested over two sessions with a class of $N = 12$ first-year graduate students in biostatistics at McGill University, Montréal (Québec), Canada. In the first session, a motivation was presented for Warner's introduction of the randomized response device for face-to-face interviews, and why it can add an extra layer of protection to today's anonymous live online polls.

Class discussion then moved to how large a value of p students would be comfortable with, and the tradeoffs between precision and privacy. It was shown, using several Bernoulli draws in R, what $p = 0.75$ means in practice. After discussing some possibly embarrassing attributes (related, e.g., to illegal or morally reprehensible behavior), an innocuous question was settled on: ownership of an `Apple` phone.

The fact that all students had a smart phone and that there is only one major competitor meant that the reverse question could simply be about ownership of an Android phone. All but one had their laptops with them, and the 12th was able to run `RStudio cloud` on a tablet, so each student used the following sample function in R:

```
sample(c( "Do you have an iPhone?",
   "Do you have an Android phone?"),
  size = 1, prob = c(0.75, 0.25))
```

Students were asked to report their "Yes/No" answers via the `Vevox` platform, which allows those directed to its website to use an instructor-prepared template to respond in real time. Respondents do not log in or identify themselves, and the instructor and all classmates only see the tally of the responses, never individual ones. Refer to the Online Supplement for details. So that $\sqrt{R}$ was a readily computed quantity, the survey was run $R = 9$ times using $p = 0.75$ and the successive values of X were 9, 9, 8, 8, 8, 10, 7, 8, and 6, yielding $\bar{X} = 73/9 = 8.1$.

The remainder of the hour was devoted to deriving the MoM estimator $\hat{n}$. It was then applied to all values of X, yielding the following values of $\hat{n}$: 12, 12, 10, 10, 10, 14, 8, 10, and 6. Note that one of the estimates, namely 14, exceeded the class size; it was used "as is" in computing $\bar{n} = 92/9 = 10.2$.

Next, it was found using (6) that $\mathrm{var}(\bar{n}) = 1$ and hence that an approximate 95% confidence interval for n has endpoints $\bar{n} \pm 2$, that is, 8.2 and 12.2. The actual number of Apple phones turned out to be 9, which is well within these limits.

In the second session, the attribute of interest was having been fully vaccinated against COVID-19 or not. To avoid double negatives, the two alternatives were stated as follows:

```
I AM fully vaccinated against COVID-19;
I AM NOT fully vaccinated against COVID-19
```

The students were then asked to respond "True" or "False" to whichever version their randomization device selected. As each round took less than a minute, the poll was again run $R = 9$ times using $p = 0.75$. The observed values of X were 10, 8, 8, 8, 9, 9, 7, 8, and 11, so that $\bar{X} = 78/9 = 8.67$ and $\hat{n} = 2 \times (8.67 - 3) = 11.3$. As the margin of error is $1.73 \times \sqrt{12}/\sqrt{9} \approx 2$, the estimate 11.3 is compatible with 9 or more of the 12 students having been fully vaccinated. Considering the sensitive nature of vaccination status, however, a non-randomized census was not run to check the true value of n.

The opportunity was taken to explore how many rounds it would have taken to narrow the margin of error to 1. That it would take 36 rounds (rather than 18) surprised some students even though, unlike Newton (Stigler 2016, chap. 2), they were all aware of the "$\sqrt{n}$ law." Another question that could generate discussion is: With only $N = 1$ person in the class, after how many rounds would one be confident to know the vaccinated status of that individual?

7. Concluding Remarks

By describing how a simple randomized response scheme can be presented in class to investigate an interesting but sensitive issue, we hope to have shown the relevance of this technique and, at the same time, rekindled statisticians' interest in using it to illustrate several basic statistical concepts involved in its implementation. Table 1 lists the concepts covered and proposes questions in increasing level of difficulty to spark class discussion.

In this article, we focused for simplicity on the first version proposed by Warner (1965), where a dichotomous issue

Table 1. Concepts and discussion.

Mathematical concepts	Prompts and discussion
Section 3: Estimation	
Random variable (RV)	X is the number of individuals who responded "Yes."
Sums of independent RVs	$X = V_n + W_{N-n}$, where $V_n \sim \mathcal{BIN}(n, p)$ and $W_{N-n} \sim \mathcal{BIN}(N-n, 1-p)$ are independent.
Model parameters	Which of the quantities n, N, x, p is fixed in advance, observable, or to be estimated?
Expected value	What are the expected values of V_n, W_{N-n}, and X?
Method of moments (MoM)	Derive the MoM estimate $\hat{n}$ of n.
	Is $p = 1/2$ a valid selection probability? Provide a non-algebraic reasoning.
Intrinsic estimator	Is $\hat{n}$ intrinsic, that is, is it always an integer between 0 and N?
	What are the advantages and disadvantages of making $\hat{n}$ intrinsic?
Conditioning	For any $x \in \{0, \ldots, N\}$, $\Pr(X = x) = \sum_{y=\max(0,x+n-N)}^{\min(x,n)} \Pr(V_n = y) \Pr(W_{N-n} = x - y)$.
Likelihood	Write down the likelihood expression as a function of n, N, x, p.
Maximum likelihood	Assume $p = 0.7$ and that a count of $X = 120$ positive responses was observed in a class of $N = 200$. Compute the maximum likelihood estimate $\hat{n}_{ML}$.
Section 4: Margin of error	
Central limit theorem	The variance of the MoM estimator $\hat{n}$ is given by $\mathrm{var}(\hat{n}) = \frac{1}{(2p-1)^2}\,\mathrm{var}(X) = \frac{Np(1-p)}{(2p-1)^2}$.
Proof by simulation	For what values of p is the variance the smallest?
Quadratic equations	Is this value or are these values feasible? If not, what might be a good compromise?
	Derive a 95% confidence interval for the proportion n/N. State your assumptions.
Variance of the sum of independent RVs	Prove that when a series of mutually independent Bernoulli trials $X_1, X_2, \ldots$ with possibly different success probabilities $p_1, p_2, \ldots$ is performed, the distribution of the proportion $\bar{X}_N$ of successes, properly standardized, is asymptotically Gaussian whenever $\lim_{N\to\infty} \sum_{i=1}^{N} p_i(1-p_i) = \infty$.
	For $p = 0.75$, $n/N = 0.75$, and $N \in \{20, 40, 80, 160\}$ use simulations to show when the Normal approximation is appropriate.
	Derive an expression for p, as a function of N, such that the desired margin of error is k, that is, one would like to state (with 95% assurance) that the number of vaccinated respondents is $\hat{n} \pm k$.
	Derive an expression for p, as a function of a given fraction f of N, such that the desired margin of error is k.
Section 5: Repeating the poll	
The curse of $\sqrt{n}$	Assume the poll is repeated R times to obtain $X_1, \ldots, X_R$. Let $\bar{n} = (\hat{n}_1 + \cdots + \hat{n}_R)/R$ be the average estimate of the number of vaccinated individuals over the R trials.
Variance calculation	Derive the variance of $\bar{n}$.
	What happens to the margin of error if the number of repeats is raised from $R = 4$ to $R = 16$?
	Can you describe the pattern more generally?
	Would one get the same conclusion if the one-sample analysis were applied to $X^* = X_1 + \cdots + X_R$ and $n^* = nR$ and n^* were divided by R at the end?

NOTE: For each topic, questions are ordered in increasing level of difficulty.

is investigated through a mirrored question design in which, unbeknownst to the interviewer, the respondent answers honestly the sensitive question or its inverse. One limitation that was identified early is that respondents sometimes find this confusing and may fail to comply because of lack of confidence in the actual level of privacy guaranteed by this approach, even when they understand the instructions.

A simple variant of Warner's method was proposed by Horvitz, Shah, and Simmons (1967) and Greenberg et al. (1969) in which randomization determines whether the respondent should answer a sensitive question with probability p or an unrelated dichotomous question with probability $1 - p$. As shown by Moors (1971), this design can lead to increased efficiency if the success probability of the unrelated question is known. Such a question might be "Is your birthdate in the first half of the year?" The proportion of students in the class who were born in the first half of the year could be determined in a standard anonymous poll (i.e., without randomization), given that this is not a sensitive issue.

Other variants such as the forced response design and the disguised response design are described, for example, in the review paper by Blair, Imai, and Zhou (2015), where estimation procedures are detailed and their performance is compared in theoretical and numerical terms. Readers can also refer to Blair, Imai, and Zhou (2015) for strategies that can be applied when the randomization distribution is unknown or when, in spite of all guarantees of anonymity provided by the randomized response design, respondents still exhibit a noncompliant behavior.

By relying on the method of moments and a standard Normal approximation for the computation of the margin of error, it is possible to make the material accessible to a broad audience. Refinements involving the method of maximum likelihood or exact confidence intervals as discussed, for example, in Frey and Pérez (2012), could be envisaged in more advanced courses for statistical trainees. Moreover, one could consider Bayesian approaches to this problem. One particularly elegant solution which is sufficiently simple to be discussed in an elementary class is described by O'Hagan (1987) using Bayes linear estimators.

Supplementary Materials

The Supplementary Materials consist of five parts, all cast in the context of the vaccination issue described in Section 6 of the main text. Section 1 gives an arithmetic-only description of the method of moments. Section 2 describes ways in which the estimator $\hat{n}$ of the number of vaccinated individuals can be constrained to produce an integer result. Section 3 provides a justification for the Gaussian approximation used in the construction of confidence intervals for n and the corresponding proportion, n/N, of vaccinated individuals in the population class. Section 4 contains a brief user's guide for conducting a live online poll that uses the randomized response technique to estimate anonymously, and with an added layer of privacy, how many people in a group have been vaccinated. Finally, Section 5 provides the R code for maximum likelihood estimation and the two key figures in the main text.

Disclosure Statement

The author declare that they have no relevant or material financial interests that relate to the research described in this article.

Funding

Funding in partial support of this work was provided to C. Genest by the Canada Research Chairs Program (950-231937) and the Natural Sciences and Engineering Research Council of Canada (RGPIN-2016-04720).

References

Blair, G., Imai, K., and Zhou, Y.-Y. (2015), "Design and Analysis of the Randomized Response Technique," *Journal of the American Statistical Association*, 110, 1304–1319.

Butler, K., and Stephens, M. A. (2017), "The Distribution of a Sum of Independent Binomial Random Variables," *Methodology and Computing in Applied Probability*, 19, 557–571.

Feller, W. (1968), *An Introduction to Probability Theory and Its Applications* (Vol. 1, 3rd ed.), New York: Wiley.

Frey, J., and Pérez, A. (2012), "Exact Binomial Confidence Intervals for Randomized Response," *The American Statistician*, 66, 8–15.

Greenberg, B. G., Abul-Ela, A.-L. A., Simmons, W. R., and Horvitz, D. G. (1969), "The Unrelated Question Randomized Response Model: Theoretical Framework," *Journal of the American Statistical Association*, 64, 520–539.

Horvitz, D. G., Shah, B. V., and Simmons, W. R. (1967), "The Unrelated Question Randomized Response Model," *Social Statistics Section Proceedings of the American Statistical Association*, 1967, 65–72.

Liu, B., and Quertermous, T. (2018), "Approximating the Sum of Independent Non-identical Binomial Random variables," *The R Journal*, 10, 472–483.

Moors, J. J. A. (1971), "Optimization of the Unrelated Question Randomized Response Model," *Journal of the American Statistical Association*, 66, 627–629.

O'Hagan, A. (1987), "Bayes Linear Estimators for Randomized Response Models," *Journal of the American Statistical Association*, 82, 580–585.

Stigler, S. M. (2016), *The Seven Pillars of Statistical Wisdom*, Cambridge: Harvard University Press.

Warner, S. L. (1965), "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias," *Journal of the American Statistical Association*, 60, 63–69.