Current Applied Science and Technology Vol. 23 No. 2 (March-April 2023)

Research article

Forced Randomized Response Protocol Using Arbitrary Random Variable

Touch Sungkawichai, Peerakarn Thongsata, Tithivich Paka, Atiratch Laoharenoo and Pat Vatiwutipong*

Kamnoetvidya Science Academy, Rayong, Thailand

Received: 10 August 2021, Revised: 3 July 2022, Accepted: 15 September 2022

DOI: 10.55003/cast.2022.02.23.015

Abstract

Anonymous polls nowadays rely solely on centralization, which means that respondents should trust the owner's company not to share responses or look through all the submitted answers. However, there is a concept of randomized response which trades the need for trust with some error. In this paper, the classical forced randomized response protocol is extended by using an arbitrary random variable. We try to optimize the tradeoff between accuracy and privacy of the polling. The Gaussian random variable is chosen to perform simulations of our method. For the best model, the poll maker has to choose the parameters that maximize a utility function, which has to be defined due to the priority between privacy and accuracy. If the poll maker prioritizes voters' privacy, our simulation shows that the best Gaussian random variable model, in this case, will be the model with $\sigma = 0.9$ and $\delta = 0.2$. On the other hand, if the poll maker prioritizes accuracy, the best model for our experiment will be the one with $\sigma = 0.9$ and $\delta = 1$.

Keywords

anonymity;

privacy preserving;

randomized response;

sampling method

1. Introduction

Especially in the age of data, gathering information from people is beneficial to companies, organizations, and institutes. Although people may feel safe providing private information to secure, trusted organizations via their anonymous surveys, in reality, the survey owner has complete control and access to networking details, including the IP addresses. So, it is easy for the owner to decode back and find the corresponding responses of every individual. In other words, anonymous polls nowadays rely entirely on the survey organization trustworthiness and policies.

Warner [1] developed a randomized response procedure as a survey method to address the issue of the need to protect participant answers to sensitive questions. According to Blair *et al*. [2], randomized response protocols can be separated into four kinds: mirrored question design, forced response design, disguised response design, and an unrelated question design. The mirrored question

*Corresponding author: Tel.: (+66) 33013888 Fax: (+66) 33013889
E-mail: pat.v@kvis.ac.th

design was put forward by Warner [1]. The idea was to set a question with its negation. The voter was then given one of the two opposing questions at random. As a result, the poll creator could not tell which question a voter had chosen to respond when they received the responses. For the forced response, there was only a single question. A voter was chosen at random to either be forced to provide an answer or to express their views. This method was proposed by Boruch [3]. The next method was the disguised response, which was devised in Kuk [4] to stop the uncomfortable voter from providing a specific response. Two random items having the same set of outcomes but different probabilities were required for this procedure, such as two different weighted coins. Each coin represented each answer; the voter was asked to choose the coin according to their answer, then toss it and report the outcome. Lastly, in the unrelated question design proposed by Greenberg *et al*. [5], voters needed to pick a question randomly: one was the real question, and the other was an unrelated one. Therefore, the poll creator was unable to tell which question the voter had been responding to.

In this research, we focus on the forced response design. Lensvelt-Mulders *et al*. [6] mentioned that the forced response method was one of the most efficient designs among several classic methods. One of the real examples of research using this method was Blair *et al*. [7] who used a questionnaire to ask 2457 civilians in villages affected by militant violence. The voters were asked to roll a die; if one showed up, the answer was forced to be no. If six showed up, the answer was forced to be yes. However, the honest answer was collected if another number showed up. Basically, this design can be illustrated as two random steps. First was to choose whether the answer was forced or not. Second, if the answer had been forced, which one had it been forced to be. This design is the most wildly used in many fields of studies. There are many examples of its use in research, such as estimating the prevalence of xenophobia and anti-semitism in Germany [8], identifying the indicators of illegal behaviour [9], establishing the prevalence of the use of performance enhancing drugs [10], investigating cannabis use by Spanish university students [11], studying physical and cognitive doping in recreational triathletes [12], estimating the prevalence of drug use [13], modelling criminal behaviour among a prison population [14], and measuring individual benefits of psychiatric treatment in non-cannabis and cannabis users [15].

This research aimed to improve the security of the polling method while limiting the increase in error. The expected result was an anonymous surveying model that guaranteed high privacy for the voters, and therefore reduced certain biases from the data that the collectors received. This paper would present an extension of the classic version of the forced randomized response protocol by allowing the random item to any arbitrary random variable.

2. Materials and Methods

2.1 Mathematical background

The likelihood function is the function that measures how well a static model fits sample data. The likelihood function describes a hypersurface, where its maximum represents the combination of model parameter values that maximizes the probability of drawing the sample obtained. The procedure for obtaining these arguments of the maximum likelihood function is known as maximum likelihood estimation.

Let $X_1, X_2, \dots, X_n$ be observations from n independent and identically distributed random variables drawn from a probability distribution f that depend on some parameter θ on parameter space Θ , then the likelihood function is:

$$L = f(X_1, X_2, \dots, X_n|\theta) = f(X_1|\theta) \times f(X_2|\theta) \times \cdots \times f(X_n|\theta)$$

The maximum likelihood estimate is a method of estimating the probability distribution parameters by maximizing the likelihood function so that under the assumed statistical model, the observed data is most probable. The point in the parameter space that maximizes the likelihood function is called the maximum likelihood estimator (MLE). The goal of maximum likelihood estimation is to find the values of the model parameters that maximize the likelihood function over the parameter space Θ, that is:

$$\hat{\theta} = \arg_{\theta\in\Theta} \max L$$

The specific value $\hat{\theta} \in \Theta$ is called the maximum likelihood estimate. If it is measurable, then it is called the maximum likelihood estimator.

Lastly, Bayes' Theorem describes how to update the probabilities of hypotheses when given evidence. It follows simply from the axioms of conditional probability. However, it can be used to reason a wide range of problems involving belief updates powerfully. Given an experiment, the universe U includes n unsimultaneous events $A_1, A_2, \dots, A_n$, and E be an event in the sample space given by A_i for $i = 1, \dots, n$. The conditional probability of A_i given by E, which had already occurred, can be determined by the following equation [16]:

$$P(A_i|E) = \frac{P(E|A_i)\cdot P(A_i)}{\sum_{i=1}^{n} P(E|A_i)\cdot P(A_i)} = \frac{P(E|A_i)\cdot P(A_i)}{P(E)}$$

2.2 Forced randomized response protocol

The classical forced randomized response protocol can be simply visualized, as shown in Figure 1. Firstly, the voter casts their vote, either 1 or 0, denoted as v. Then, they pass their intended answer into the predefined algorithm, which gives back a randomized value, v', based on the intended answer v. Note that $v' \in \{0, 1\}$, the diagram in Figure 1 shows that when $v = 1$, $P(v' = 1) = f + (1-f)q$, and similarly, when $v = 0$, we get $P(v' = 1) = (1-f)q$.

From the above observation, it can be said that v' has a Bernoulli distribution with success probability $f + (1-f)q$. A similar argument applies in the other case. Therefore, the classical forced randomized response method can also be viewed as a random variable, as shown in the following equation:

$$v' = \begin{cases} B(f + (1-f)q) & \text{if} \quad v = 1 \\ B((1-f)q) & \text{if} \quad v = 0 \end{cases}$$

where $B(p)$ is the Bernoulli random variable with success probability p. Note that the expectation $E[B(f + (1-f)q)] = E[B((1-f)q)]$ applied only if $f = 0$, and the model with $f = 0$ is not valid.

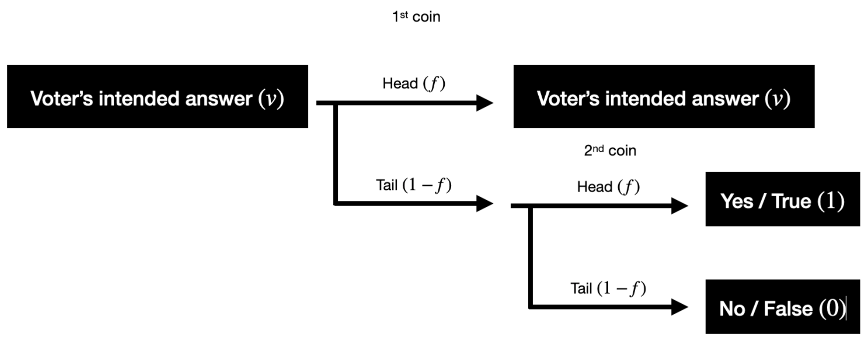

Figure 1. Classical forced randomized response protocol

From the previous observation, the forced randomized response protocol can be extended to any arbitrary random variable as in the following equation:

$$v' = \begin{cases} X & \text{if} \quad v = 1, \\ Y & \text{if} \quad v = 0, \end{cases}$$

where X and Y are arbitrary same type random variables (both continuous or both discrete) such that $E[X] \neq E[Y]$.

An approximation of the numbers of voters who intentionally choose $v = 1$ can be derived from the point estimator. First, denote p as the probability that each independent voter will vote for $v = 1$, this is the preferential bias of a certain population, and n as the expected number of voters who intentionally choose $v = 1$, which means that $n = pN$. The point estimator of the numbers of voters who intentionally vote $v = 1$ is $\hat{n} = \hat{p}N$. From the assumption regarding population bias, p, the distribution of v' is shown in the equation:

$$v'(p, X, Y) = pX + (1-p)Y$$

The point estimator of p, is then derived as:

$$\hat{p} = \frac{\sum_{i=1}^{N} v_i{}' - E[Y] \cdot N}{E[X-Y] \cdot N}$$

which is unbiased, since

$$\begin{aligned} E[\hat{p}] &= E\left[\frac{\sum_{i=1}^{N} v_i{}' - E[Y] \cdot N}{E[X-Y] \cdot N}\right] \\ &= \frac{E\left[\sum_{i=1}^{N} v_i' - E[Y] \cdot N\right]}{E[X-Y] \cdot N} \\ &= \frac{E[v'] - E[Y]}{E[X-Y]} \end{aligned}$$

$$
\begin{aligned}
&= \frac{E[pX + (1-p)Y] - E[Y]}{E[X-Y]} \\
&= \frac{pE[X-Y]}{E[X-Y]} \\
&= p
\end{aligned}
$$

2.3 Performance

The performance of each model can be measured in two aspects, which are privacy and accuracy. Privacy is the measure of the difficulty of a bystander to guess the intended answer knowing only the randomized value. On the other hand, accuracy is the measure of the real-world error in using the process, which is simulated by a computer.

For every value of v', it is possible to know the probability that $v = 1$ and $v = 0$. The idea is that distinguishing $v = 1$ and $v = 0$ can be done with high confidence when the difference of probability of $v = 1$ and $v = 0$ is high. If that is the case, then the voter's intention is not private. Therefore, let $G(z) = |f_X(z) - f_Y(z)|$, where f_X is the probability density function of X and f_Y is the probability density function of Y, then $G(z)$ represents the confidence of determining v given that $v' = z$.

The insecurity function is then defined as the product of the area under the curve of $G(x)$ and the probability of having $v' = x$ over all possible xs. We formally define the insecurity I by the following equation:

$$
\begin{aligned}
I &= \int_{-\infty}^{\infty} |P(v=1|v'=x) - P(v=0|v'=x)| \cdot P(v'=x) \quad dx \\
&= \int_{-\infty}^{\infty} |P(v'=x|v=1)P(v=1) - P(v'=x|v=0)P(v=0)| \quad dx \\
&= \int_{-\infty}^{\infty} |f_X(x)p - f_Y(x)(1-p)| \quad dx
\end{aligned}
$$

Note that integration should be changed to summation over all x when using discrete random variables.

Accuracy is measured using a computer simulation. A simulation takes two random variables, applies the extended forced randomized response protocol, gathers data, and decodes the result using the point estimator. The accuracy is reversely defined using the error, which is the difference between real value and the value yielded by the model in each trial.

3. Results and Discussion

The Gaussian random variable is chosen as an example of using the extended forced randomized response technique. The process is started by defining the random function which, in this case, will be as:

$$
v' = \begin{cases} N(\delta, \sigma) & \text{if} \quad v = 1 \\ N(0, \sigma) & \text{if} \quad v = 0 \end{cases}
$$

for some value of σ and $\delta \neq 0$. Here, $N(\mu, \sigma)$ denotes a Gaussian random variable with mean μ and standard deviation σ.

After defining the random variable, we will elaborate on the protocol. First, the poll maker distributes the pre-determined value of δ and σ to every voter. Then, each voter can choose their intended answer, keep it a secret, and use the δ and σ to blind their response. After they come up with the blinded result, they can submit it to the system. The privacy level of them doing so will be discussed later. Next, the system gathers the responses from all the voters and then proceeds with the calculation. In the calculation, the machine sums up all the responses and uses the unbiased point estimator to map back the value.

The unbiased point estimator of this model can be derived as:

$$\hat{p} = \frac{\sum_{i=1}^{N} v_i{}'}{\delta N}$$

The definition of insecurity in the equation can be adapted to match with the case of Gaussian variables. The derivation of the privacy formula is included in the following equation:

$$\begin{aligned} I &= \int_{-\infty}^{\infty} |f_X(x)p - f_Y(x)(1-p)| \quad dx \\ &= \int_{-\infty}^{\infty} \left| \frac{p}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\delta)^2}{2\sigma^2}} - \frac{1-p}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \right| \quad dx \\ &= (p-1)\,\mathrm{erf}\left(\frac{x-\delta}{\sigma\sqrt{2}}\right) - p\,\mathrm{erf}\left(\frac{-x}{\sigma\sqrt{2}}\right) \end{aligned}$$

where

$$x = \frac{\delta^2 - 2\sigma^2\left(\ln\left(\frac{1}{p}-1\right)\right)}{2\delta}.$$

The Gaussian random variable model is simulated using the method for various quadruples of (N, p, δ, σ) . The classical model simulates the same way for various quadruples of (N, p, f, q) as a comparison set. In each simulation, the model is tested for error 100 times. The root mean square of the errors, defined in Jeffreys [16], was used in all testing. The results from the simulation are shown in Tables 1-4.

Table 1 shows the insecurity of both the Gaussian random variable model and the classical model. From the Table, it is observed that for the cases where the population is extremely biased toward one end of preference, when the population bias p is 0.1 or 0.9, the insecurity of all models is higher than in other cases. This means that it is harder to protect voters' intentions from being discovered in those kinds of situations. However, for a common population with a population bias p around 0.5, it is possible to choose a model that yields arbitrary insecurity using both the classical and Gaussian random variable models.

From Table 2, it is observed that a high value of f decreases the root mean square error of the model significantly. The classical model with higher value of f has a lower probability of convincing the voters' answers with a random one. With lower mutation rate, the decoding process can be done more efficiently. These phenomena occur regardless of p, N, and q.

Similarly, Table 3 shows the relation between δ and the accuracy of the model. Higher values of δ make the model more accurate for every value of p, N, and σ. The value of δ directly causes differences in expectation of two models, which is the noise layer in this model. Therefore, the larger δ is, the harder it is to reverse the process and determine voters' intention.

Table 1. Insecurity of models

p				**Classical Model**						**Gaussian Variable Model**		
	q	***f*=0.1**	***f*=0.3**	***f*=0.5**	***f*=0.7**	***f*=0.9**	**σ**	**δ=0.2**	**δ=0.4**	**δ=0.6**	**δ=0.8**	**δ=1.0**
0.1	0.1	0.800	0.800	0.800	0.800	0.836	0.1	0.860	0.976	0.998	1.000	1.000
	0.3	0.800	0.800	0.800	0.800	0.868	0.3	0.800	0.814	0.860	0.911	0.951
	0.5	0.800	0.800	0.800	0.800	0.900	0.5	0.800	0.800	0.808	0.830	0.860
	0.7	0.800	0.800	0.800	0.800	0.932	0.7	0.800	0.800	0.801	0.806	0.819
	0.9	0.800	0.800	0.820	0.892	0.964	0.9	0.800	0.800	0.800	0.801	0.805
0.3	0.1	0.400	0.400	0.400	0.604	0.868	0.1	0.723	0.959	0.998	1.000	1.000
	0.3	0.400	0.400	0.420	0.652	0.884	0.3	0.428	0.572	0.723	0.838	0.914
	0.5	0.400	0.400	0.500	0.700	0.900	0.5	0.402	0.452	0.541	0.635	0.723
	0.7	0.400	0.412	0.580	0.748	0.916	0.7	0.400	0.415	0.463	0.527	0.595
	0.9	0.400	0.524	0.600	0.796	0.932	0.9	0.400	0.404	0.428	0.470	0.520
0.5	0.1	0.100	0.300	0.500	0.700	0.900	0.1	0.683	0.954	0.997	1.000	1.000
	0.3	0.100	0.300	0.500	0.700	0.900	0.3	0.261	0.495	0.683	0.818	0.904
	0.5	0.100	0.300	0.500	0.700	0.900	0.5	0.159	0.311	0.451	0.576	0.683
	0.7	0.100	0.300	0.500	0.700	0.900	0.7	0.114	0.225	0.332	0.432	0.525
	0.9	0.100	0.300	0.500	0.700	0.900	0.9	0.088	0.176	0.261	0.343	0.421
0.7	0.1	0.400	0.524	0.660	0.796	0.932	0.1	0.723	0.959	0.998	1.000	1.000
	0.3	0.400	0.412	0.580	0.748	0.916	0.3	0.428	0.572	0.723	0.838	0.914
	0.5	0.400	0.400	0.500	0.700	0.900	0.5	0.402	0.452	0.541	0.635	0.723
	0.7	0.400	0.400	0.420	0.652	0.884	0.7	0.400	0.415	0.463	0.527	0.595
	0.9	0.400	0.400	0.400	0.604	0.868	0.9	0.400	0.404	0.428	0.470	0.520
0.9	0.1	0.800	0.800	0.820	0.892	0.964	0.1	0.860	0.976	0.998	1.000	1.000
	0.3	0.800	0.800	0.800	0.800	0.932	0.3	0.800	0.814	0.860	0.911	0.951
	0.5	0.800	0.800	0.800	0.800	0.900	0.5	0.800	0.800	0.808	0.830	0.860
	0.7	0.800	0.800	0.800	0.800	0.868	0.7	0.800	0.800	0.801	0.806	0.819
	0.9	0.800	0.800	0.800	0.800	0.836	0.9	0.800	0.800	0.800	0.801	0.805

Table 2. Root mean square error of the classical model

q	***p***			***N* = 500**					***N* = 3000**		
		***f*=0.1**	***f*=0.3**	***f*=0.5**	***f*=0.7**	***f*=0.9**	***f*=0.1**	***f*=0.3**	***f*=0.5**	***f*=0.7**	***f*=0.9**
0.1	0.1	67.985	24.452	11.761	6.737	3.113	170.247	51.158	27.899	15.485	6.921
	0.3	76.961	25.349	15.467	9.501	4.309	162.130	56.552	32.682	18.927	11.085
	0.5	80.685	25.421	18.859	10.293	5.020	182.592	62.097	40.64	26.206	13.965
	0.7	75.153	33.338	18.514	12.766	5.834	208.370	67.94	48.982	28.475	13.553
	0.9	76.177	35.977	21.868	13.064	6.952	228.473	78.713	53.636	30.599	20.776
0.3	0.1	98.204	31.505	16.840	10.179	4.539	260.002	80.421	39.387	21.831	11.123
	0.3	108.908	34.413	16.420	10.852	4.754	267.013	89.145	43.434	23.967	12.916
	0.5	102.479	37.265	19.144	11.056	5.045	246.585	79.975	47.097	29.248	11.757
	0.7	101.124	35.539	21.324	12.506	5.852	283.251	89.086	50.148	28.666	14.849
	0.9	102.538	32.452	20.152	12.967	5.558	260.709	82.655	52.202	29.258	17.442
0.5	0.1	133.195	35.727	19.532	12.015	4.418	296.231	84.275	46.330	27.542	12.941
	0.3	110.603	31.633	20.398	11.898	4.810	238.967	88.072	47.530	28.301	13.653
	0.5	100.215	32.395	20.539	10.535	5.952	274.124	90.005	48.067	25.432	13.973
	0.7	116.452	34.898	19.127	11.577	5.782	294.403	79.056	47.777	29.349	13.657
	0.9	108.798	31.565	19.549	11.303	5.160	302.772	88.756	46.943	25.523	14.339
0.7	0.1	113.142	36.411	19.921	12.469	6.153	251.803	87.314	52.831	30.468	13.206
	0.3	104.862	36.712	19.093	11.014	5.206	272.149	79.441	45.474	27.376	13.774
	0.5	113.618	34.686	18.487	11.099	6.587	237.051	90.624	45.093	28.433	12.785
	0.7	89.247	34.158	15.505	9.919	4.978	246.66	72.670	45.837	27.444	12.135
	0.9	93.301	32.590	17.374	11.111	4.469	244.121	81.700	43.517	20.524	11.224
0.9	0.1	82.219	33.568	22.320	11.602	5.582	193.505	90.346	53.962	35.248	17.828
	0.3	86.475	36.278	21.516	12.192	6.050	185.610	78.723	48.704	31.296	15.259
	0.5	86.470	28.445	17.249	11.433	5.556	179.483	80.955	41.721	25.081	14.191
	0.7	70.788	23.345	15.590	9.203	4.238	146.161	59.124	37.441	23.412	9.835
	0.9	67.424	20.854	11.370	6.690	2.773	148.432	48.470	28.978	16.424	8.687

Table 3. Root mean square error of Gaussian random variable model

σ	p	N = 500					N = 3000				
		δ=0.2	δ=0.4	δ=0.6	δ=0.8	δ=1.0	δ=0.2	δ=0.4	δ=0.6	δ=0.8	δ=1.0
0.1	0.1	67.985	31.710	24.452	16.985	11.761	170.247	75.593	51.158	42.435	27.899
	0.3	76.961	35.366	25.349	18.424	15.467	162.130	92.189	56.552	46.715	32.682
	0.5	80.685	43.506	25.421	20.561	18.859	182.592	98.555	62.097	49.731	40.640
	0.7	75.153	49.835	33.338	24.031	18.514	208.370	118.919	67.940	61.294	48.982
	0.9	76.177	50.540	35.977	25.291	21.868	228.473	116.636	78.713	60.930	53.636
0.3	0.1	98.204	46.979	31.505	20.792	16.840	260.002	113.402	80.421	51.673	39.387
	0.3	108.908	57.617	34.413	22.501	16.420	267.013	113.487	89.145	62.048	43.434
	0.5	102.479	48.492	37.265	28.109	19.144	246.585	123.712	79.975	57.109	47.097
	0.7	101.124	51.186	35.539	28.084	21.324	283.251	125.614	89.086	65.588	50.148
	0.9	102.538	54.376	32.452	25.752	20.152	260.709	122.108	82.655	63.380	52.202
0.5	0.1	133.195	56.493	35.727	26.239	19.532	296.231	124.602	84.275	68.299	46.330
	0.3	110.603	60.673	31.633	23.971	20.398	238.967	132.058	88.072	54.546	47.530
	0.5	100.215	51.819	32.395	27.590	20.539	274.124	137.142	90.005	59.800	48.067
	0.7	116.452	46.634	34.898	23.808	19.127	294.403	131.896	79.056	58.106	47.777
	0.9	108.798	56.840	31.565	24.033	19.549	302.772	143.224	88.756	69.214	46.943
0.7	0.1	113.142	49.221	36.411	25.558	19.921	251.803	139.285	87.314	65.766	52.831
	0.3	104.862	60.073	36.712	25.690	19.093	272.149	113.063	79.441	68.456	45.474
	0.5	113.618	54.282	34.686	25.522	18.487	237.051	126.803	90.624	58.097	45.093
	0.7	89.247	45.318	34.158	24.091	15.505	246.660	117.806	72.670	55.959	45.837
	0.9	93.301	51.105	32.590	22.109	17.374	244.121	118.591	81.700	53.886	43.517
0.9	0.1	82.219	44.905	33.568	27.879	22.320	193.505	116.713	90.346	64.796	53.962
	0.3	86.475	43.972	36.278	24.867	21.516	185.610	110.580	78.723	53.422	48.704
	0.5	86.470	39.265	28.445	18.611	17.249	179.483	107.509	80.955	47.550	41.721
	0.7	70.788	38.425	23.345	18.891	15.590	146.161	85.016	59.124	45.018	37.441
	0.9	67.424	30.859	20.854	15.554	11.370	148.432	76.327	48.470	36.099	28.978

Table 4. Average of root mean square error of all the models in percentage of *N*

Model	p	N							
		100	500	1000	1500	2000	2500	3000	3500
Classical	0.1	12.860	5.762	4.064	3.310	2.872	2.560	2.354	2.138
	0.3	12.829	5.904	4.105	3.311	2.845	2.540	2.337	2.196
	0.5	13.150	5.804	4.061	3.431	2.857	2.564	2.322	2.219
	0.7	12.806	5.858	4.056	3.324	2.878	2.570	2.376	2.210
	0.9	12.914	5.777	4.064	3.306	2.840	2.538	2.354	2.168
Gaussian	0.1	10.905	4.957	3.592	2.992	2.561	2.309	2.042	1.908
	0.3	11.372	5.016	3.747	2.905	2.517	2.250	2.093	1.942
	0.5	11.201	5.085	3.599	2.913	2.498	2.285	2.076	1.910
	0.7	11.420	5.047	3.606	2.981	2.578	2.306	2.139	1.941
	0.9	11.376	5.214	3.549	2.873	2.546	2.304	2.034	1.931

In Table 4, each model that is evaluated is a combination of $f = 0.1, 0.2, \ldots, 0.9$, $q = 0.1, 0.2, \ldots, 1.0$ for classical model, and $\delta = 0.2, 0.4, \ldots, 1$, $\sigma = 0.1, 0.2, \ldots, 0.9$ for the Gaussian random variable model. Note that since the average is taken from just some instances of the model, the value of the Gaussian model and classical model cannot be directly compared. It is not rational to say that one model is better than the other if it has a lower value. This Table illustrates the relation of the accuracy of models with respect to the size of the population and population bias. It shows that this simulation agrees with the law of large numbers in that when the number of samples goes up, the accuracy of the model also increases.

Some instances of the model perform better with certain populational biases. The better models and the worse models cancel out and the average performance does not vary with the population. These numbers suggest that for making a survey in an unknown populational bias, using a random model from the mentioned list will yield the expected error shown in Table 4.

Defining a global requirement for the survey or prioritizing voters' privacy and accuracy is not rational, as it cannot be useful in general cases. Surveys have different requirements, so there will not be a model that fits with all scenarios. For instance, surveying politics in a dictatorship country may require prioritizing voters' privacy over accuracy (and compensate the accuracy measure by going through more samples), whereas a poll regarding service satisfaction of a small population might need higher accuracy by compensating voters' privacy. The poll makers should apply this model to fit specific situations.

To select the best model, the poll maker has to, consciously or unconsciously, define a utility function that takes into account insecurity measure and expected error, and also other environmental parameters if accessible. After that, the poll maker needs to find the model that gives the maximum utility cost from all of the available models. To illustrate this method, let I denote the insecurity measure of a model and E denotes expected error in a certain situation (population of N with bias p). If the poll maker prioritizes voters' privacy and defines the utility function to be $U_1(I, E) = -10000I - E$, in a population of 1000 people that has a population bias $p = 0.5$, the best Gaussian random variable model in this case will be the model with $\sigma = 0.9$ and $\delta = 0.2$ which yields the highest utility, -1025.740. Note that this value is not the global maximum, but it is the best instance from those that are used in this experiment. On the other hand, if the poll maker prioritizes accuracy, they might define $U_2(I, E) = -1000I - 10E$. The insecurity measure is, on average, smaller than the error term by approximately a factor of N. The best model of our experiment is the one with $\sigma = 0.9$ and $\delta = 1$, which yields -663.027 utility points. Moreover, the poll maker can also use more advanced utility functions such as $U(I, E) = \sum_p -1000I - E$ for all p in $\{0.1, 0.2, \ldots, 0.9\}$ and a population of 1000 people. Using this function, the model with $\sigma = 0.9$ and $\delta = 0.4$ gives the maximum utility points of about -4911.043.

4. Conclusions

The extended forced randomized response protocol can benefit every party involved in the polling process by providing a security preserving sampling technique that does not trade off too much accuracy. This model of extended forced randomized response protocol makes it possible to apply any random variable to be used in the sampling. However, some randomized variables, such as Bernoulli or binomial, can be easily conducted on the client-side, by using coin flips, while some might require other instances.

From the extended model we develop, it is possible to use an arbitrary random variable in the forced randomized response protocol. Some random variables might be better than others in terms of accuracy, privacy, or practical ease of use, but this lies beyond the scope of this study. The extended forced randomized response protocol acts as an interface so that the effectiveness of the model depends on the choice of random variables.

This study has shown that using a Gaussian random variable in the extended forced randomized response protocol is enough to replace the classical forced randomized response protocol. However, since it is not rational to prioritize insecurity and accuracy in general cases, we have also concluded that it is not feasible to find an absolute measure that combines insecurity and accuracy. However, the poll maker can define a survey utility function to help determine the best model to use in certain scenarios. In the future, we plan to further extend this concept and apply it to the randomized response protocol for other type of questions such as multiple choice with multiple answers, Likert scale, rating scale and rank order poll questions.

References

[1] Warner, S.L., 1965. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309), 63-69.

[2] 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.

[3] Boruch, R.F., 1971. Assuring confidentiality of responses in social research: A note on strategies. *American Sociologist*, 6, 308-311.

[4] Kuk, A.Y., 1990. Asking sensitive questions Indirectly. *Biometrika*, 77(2), 436-438.

[5] Greenberg, B.G., Kuebler, R.R., Abernathy, J.R. and Horvitz, D.G., 1971. Application of the randomized response technique in obtaining quantitative data. *Journal of the American Statistical Association*, 66, 243-250.

[6] Lensvelt-Mulders, G.J.L.M., Hox, J.J. and Heijden, P.G.M., 2005. How to improve the efficiency of randomised response designs. *Quality and Quantity*, 39, 253-265.

[7] Blair, G., Imai, K. and Zhou, Y.-Y., 2015. Design and analysis of the randomized response technique. *Journal of the American Statistical Association*, 110(511), DOI: 10.1080/0162 1459.2015.1050028.

[8] Krumpal, I., 2012. Estimating the prevalence of xenophobia and anti-semitism in Germany: A comparison of randomized response and direct questioning. *Social Science Research*, 41, 1387-1403.

[9] St John, F.A.V., Keane, A.M., Edwards-Jones, G., Jones, L., Yarnell, R.W., and Jones, J.P.G., 2012. Identifying indicators of illegal behaviour: carnivore killing in human-managed landscapes. *Proceedings of the Royal Society B: Biological Sciences*, 279, 804-812.

[10] Stubbe, J.H., Chorus, A.M., Frank, L.E., Hon, O. and Heijden, P.G., 2014. Prevalence of use of performance enhancing drugs by fitness centre members. *Drug Testing and Analysis*, 6, 434-438.

[11] Cobo, B., Rueda, M.M. and López-Torrecillas, F., 2016. Application of randomized response techniques for investigating cannabis use by Spanish university students. *International Journal of Methods in Psychiatric Research*, 26(4), DOI: 10.1002/mpr.1517.

[12] Schröter, H., Studzinski, B., Dietz, P., Ulrich, R., Striegel, H. and Simon, P., 2016. A comparison of the cheater detection and the unrelated question models: A randomized response survey on physical and cognitive doping in recreational triathletes. *PLoS One,* 11(5), DOI: 10.1371/journal.pone.0155765.

[13] Kirtadze, I., Otiashvili, D., Tabatadze, M., Vardanashvili, I., Sturua, L., Zabransky, T. and Anthony, J.C., 2018. Republic of Georgia estimates for prevalence of drug use: Randomized response techniques suggest under-estimation. *Drug and Alcohol Dependence*, 187, 300-304.

[14] Cobo, B., Castillo, E., López-Torrecillas, F. and Rueda, M.D.M., 2021. Indirect questioning methods for sensitive survey questions: Modelling criminal behaviours among a prison population. *PLoS One*, 16(1), DOI: 10.1371/journal.pone.0245550.

[15] Zhang, X., Crespo-Facorro, B., Leon, J. and Diaz, F.J., 2020. Measuring individual benefits of psychiatric treatment using longitudinal binary outcomes: Application to antipsychotic benefits in non-cannabis and cannabis users. *Journal of Biopharmaceutical Statistics*, 30(5), 916-940.

[16] Jeffreys, H., 1973. *Scientific Inference*. Cambridge: Cambridge University Press.