Journal of Probability
and Statistical Science
24(1), 1-15 Feb. 2026

Advancing Randomized Response: Logistic Regression with Zero-Truncated Negative Binomial Design

Neelam
Department of Statistics
University of Peshawar, Pakistan

Syed Muhammad Asim
Department of Statistics
University of Peshawar, Pakistan

Qamruz Zaman
Department of Statistics
University of Peshawar, Pakistan

Farooq Shah
Department of Statistics
University of Peshawar, Pakistan

ABSTRACT

Sensitive survey research often suffers from non-response and misreporting due to privacy concerns. Randomized Response Techniques (RRTs) provide a mechanism to elicit truthful answers while preserving confidentiality, but existing designs face challenges when data are heterogeneous or overdispersed. This paper introduces a new extension of the Zero-Truncated Negative Binomial (ZTNB) randomization device to logistic regression with covariates. The proposed framework integrates the ZTNB distribution into the logistic likelihood, thereby accommodating overdispersion while maintaining unbiased estimation of regression parameters. Theoretical results establish the consistency and asymptotic normality of the maximum likelihood estimators. A comprehensive simulation study evaluates the finite-sample properties of the method under varying coefficient settings and sample sizes. Results demonstrate negligible bias, close agreement between empirical and model-based variability, and coverage probabilities near the nominal 95% level. Compared to the Zero-Truncated Poisson device, the ZTNB approach exhibits improved stability and precision, particularly in small to moderate samples. These findings confirm the ZTNB logistic regression model as a flexible and efficient tool for analyzing sensitive survey data, expanding the scope of RRT-based inference.

Keywords: Randomized Response Technique (RRT); Zero-Truncated Negative Binomial (ZTNB); Logistic Regression; Sensitive Survey; Overdispersion, Maximum Likelihood Estimation; Simulation Study

1. Introduction

Collecting accurate information on sensitive topics such as income, tax evasion, criminal behavior, drug use, or health-related issues poses a persistent challenge in survey research. Respondents often feel reluctant to disclose truthful answers when questions involve behaviors that are illegal, stigmatized, or socially undesirable. This reluctance leads to two major problems: non-response bias, where individuals refuse to participate altogether, and response bias*,* where they provide intentionally false answers to protect their privacy. Conventional direct questioning is therefore inadequate for such contexts, as it fails to balance the twin goals of obtaining reliable estimates and safeguarding respondents' confidentiality. Over the decades, statisticians and survey methodologists have recognized this dilemma as one of the central challenges in social and behavioral data collection.

To address this problem, Warner [1] introduced the Randomized Response Technique (RRT), a groundbreaking method that allowed respondents to answer sensitive questions indirectly while maintaining plausible deniability. The key innovation of Warner's design was to introduce a randomization mechanism that obscures individual responses but still enables unbiased

- Received October 12, 2025, in final form February 2026.

- Neelam (corresponding author) is affiliated with Department of Statistics,
University of Peshawar, Pakistan
neelam_arbab@uop.edu.pk

estimation of population proportions. Following Warner's seminal proposal of RRT to mitigate evasive-answer and social desirability bias in sensitive surveys, a broad methodological lineage emerged that refines the privacy–efficiency trade-off. Warner's original device established unbiased estimation with respondent-level masking, cementing the paradigm for indirect questioning in statistics. Building on this, Greenberg et al. [2] introduced the unrelated-question model, giving respondents a known-probability route to answer a neutral item, which improved deniability and offered efficiency comparisons with Warner's scheme. Methodological refinements continued through the 1990s and 2000s; Mangat [3] proposed a simpler, more efficient strategy under clear conditions, sparking a stream of analyses and alternatives; two-stage and controlled variants further tuned efficiency and practicality. Parallel strands extended RRT to multiple sensitive attributes and complex designs (e.g., joint proportions, multi-level attributes), while comparative reviews synthesized decades of variants and their operating characteristics [4]. More recently, comprehensive surveys of the field e.g., **L**e et al. [5] document the spread of RRT across disciplines (behavioral, socio-economic, psychological, epidemiology/biomedical, and public health), summarizing advances and open issues after nearly six decades. Collectively, these works demonstrate the adaptability of RRT and its central role in protecting privacy in sensitive surveys.

Beyond methodology, applications demonstrate RRT's practical value for topics prone to misreporting. Early work showed higher estimated prevalence and fewer refusals for drug use relative to direct questioning, and the approach has been repeatedly revisited in substance-use epidemiology. In sexual health and HIV contexts, controlled and hybrid designs using RRT improved cooperation rates and yielded credible estimates in key populations; newer studies even combine RRT with network scale-up to size hidden groups [6]. Public-health and campus surveys likewise report that RRT uncovers materially higher, arguably more realistic, prevalence of sensitive behaviors (e.g., cognitive-enhancing drug use) than standard questionnaires [7]. At the same time, recent evaluations caution that implementation details (instructions, device understanding, context) can affect disclosure, underscoring the need for designs that balance protection with efficiency and usability. In criminology, it has been applied to estimate the prevalence of offenses such as theft, tax evasion, and bribery. In sociology and psychology, RRT has helped measure stigmatized behaviors, attitudes, and mental health indicators. These diverse applications reinforce the practical value of RRT in generating truthful responses in contexts where direct questioning would likely fail.

Against this backdrop, recent attention has turned to Zero-Truncated distributions as promising randomization devices. The motivation stems from their ability to exclude zero outcomes, thereby enhancing anonymity and stabilizing estimator properties. Building on this line of thought, Neelam [8] proposed the Zero-Truncated Poisson (ZTP) model as a randomization device for sensitive data estimation in randomized response surveys. Unlike earlier designs, the ZTP was developed as a distinct methodological innovation, exploiting the properties of the truncated Poisson distribution to achieve gains in estimation efficiency while maintaining competitive levels of respondent protection. Simulation studies confirmed its robustness across a range of prevalence values, underscoring the potential of zero-truncated approaches in advancing the RRT literature. Compared to the ZTP model, which assumes equidispersion, the ZTNB introduces a dispersion parameter to handle overdispersion in survey data. This flexibility makes it especially suitable for heterogeneous populations or clustered responses.

Beyond proportion estimation, several scholars have explored the integration of RRTs into regression frameworks. The idea of combining randomized response with regression analysis dates back to early work on limited dependent and qualitative-variable models. Maddala [9] is often cited as laying the groundwork for handling perturbed binary outcomes within a regression context. Later, Scheers and Dayton [10] provided one of the first formal treatments of covariate-based randomized-response models. Their approach demonstrated how known randomization probabilities can be incorporated into likelihoods for logistic-type models and established that RR outcomes can be treated as misclassified or perturbed binary variables. Following this, Magder and Hughes [11] studied logistic regression under outcome uncertainty, deriving bias-correction and likelihood-based approaches that directly apply to RR-type perturbations of binary responses.

During the 1990s and 2000s, methodological development generalized these ideas to more complex and multivariate settings. Van der Heijden [12] extended RRT to univariate and multivariate logistic regression models, formalizing the connection with misclassification models and presenting likelihood-based inference for single and multiple RR responses. Subsequent contributions by Hsieh et al.[13-14] considered logistic regression for RR data with missing covariates, proposing weighted estimation methods to maintain consistency. The broader misclassification literature (e.g., Neuhaus [15] and others) also clarified how non-differential misclassification biases effect estimates and how likelihood or sensitivity-analysis approaches can be used to correct for it, providing direct insight for RR–logistic frameworks.

More recently, researchers have refined computational strategies and extended applicability. Van den Hout et al. [16] outlined both GEE-style and full-likelihood implementations for correlated RR outcomes, while later works [17-18] developed practical EM and GEE algorithms to make RR–logistic regression feasible in applied research. Reviews and validation studies [19] have synthesized evidence of RRT's practical performance across diverse survey settings, emphasizing that implementation details such as device comprehension and context critically affect disclosure. A related stream interprets RR as a form of known misclassification, allowing researchers to apply the broader econometric and epidemiologic toolbox for misclassified outcomes (e.g., bias-corrected likelihoods, sensitivity analysis, SIMEX methods).

Building on this trajectory, the Zero-Truncated Negative Binomial (ZTNB) distribution offers an even more flexible foundation for randomized response designs. Whereas the Zero-Truncated Poisson (ZTP) model by Neelam [20], assumes equidispersion, real survey data often exhibit overdispersion, where the variance exceeds the mean due to heterogeneity in respondent behavior, hidden subgroups, or clustering effects. The ZTNB distribution directly addresses this challenge through the inclusion of an additional dispersion parameter, making it particularly well-suited for applied settings where variability cannot be captured by a single-parameter Poisson structure.

In methodological terms, Neelam [21] formally introduced the ZTNB as a randomization device and derived the properties of the unbiased estimator of the population proportion. Theoretical results showed that the ZTNB-based estimator achieved lower variance than those obtained under existing devices, such as the models of Kuk [22], Singh and Grewal [23], as well as the Zero-Truncated Binomial (ZTB) by Zapata et al.[24] and ZTP distributions. Importantly, the protection index analysis demonstrated that ZTNB could safeguard respondent privacy without sacrificing efficiency, even in scenarios where the sensitive attribute was rare ($\pi \leq 0.30$). These findings underline the dual advantage of ZTNB: higher estimation efficiency combined with robust respondent protection. Moreover, its ability to capture variation beyond the mean makes it a robust choice when survey responses are heterogeneous, a feature often encountered in applied contexts such as health, criminology, and economic surveys.

Beyond theoretical properties, simulation studies further confirmed the stability and reliability of the ZTNB device. Results indicated that the model performed consistently across varying sample sizes, prevalence rates, and parameter values, thereby reinforcing its potential as a versatile tool in practice. Such robustness makes the ZTNB model an attractive candidate for extension to more complex inferential frameworks, particularly regression models that allow the incorporation of covariates. By accommodating overdispersion and eliminating zero outcomes, ZTNB represents a significant methodological advance in the RRT literature.

Despite this progress, most applications of the ZTNB distribution have so far been confined to the estimation of simple population proportions. While useful, this limited scope does not reflect the reality of modern surveys, where researchers are often interested in how sensitive attributes vary with explanatory factors such as age, income, education, or geographic region. Logistic regression has become the standard framework for analyzing such binary outcomes with covariates, making its integration with RRT both timely and necessary. Neelam [21] recently extended the ZTP device to logistic regression, showing that zero-truncated randomization mechanisms can be embedded within logistic likelihoods to obtain consistent and asymptotically normal estimators. However, no extension of the ZTNB device to logistic regression has yet been developed**.** Given that the ZTNB model incorporates an additional dispersion parameter to address overdispersion and heterogeneity, extending it to logistic regression represents a significant and timely methodological advance.

This gap provides the central motivation for the present study, which seeks to establish a framework for logistic regression inference under ZTNB randomization. In what follows, we first revisit the foundational models of Warner [1] and Maddala [9] alongside the ZTNB device of Neelam [23], before presenting the proposed extension of ZTNB to logistic regression with covariates.

2. Warner RRT

Warner [1] first proposed RRT to obtain truthful answers to sensitive questions while preserving respondent privacy. In this model, each respondent is instructed to answer either the sensitive question or its complement according to a randomization device (such as a coin flip) with known probability.

Let Y denote the sensitive attribute, where $Y = 1$ if the respondent possesses the attribute and $Y = 0$ otherwise. In Warner's model, each respondent uses a randomization device to answer either the sensitive question or its complement, according to a known probability mechanism. If p is the probability of being instructed to answer the sensitive question, then the probability of observing a "yes" response is

$$P(Z=1) = p\pi + (1-p)(1-\pi) \tag{1}$$

where $\pi = P(Y=1)$ is the true population proportion possessing the sensitive attribute and p is the known probability of being asked the original (sensitive) question.. An unbiased estimator of π is derived from this probability structure as follows.

$$\hat{\pi} = \frac{\overline{Z} - (1-p)}{2p-1} \tag{2}$$

where $\overline{Z}$ is the sample mean of the randomized responses. The sampling variance of $\hat{\pi}$ i

$$Var(\hat{\pi}_W) = \frac{\pi(1-\pi)}{n} + \frac{p(-p)}{n(2p-1)^2} \tag{3}$$

Warner's framework demonstrated that randomized answers can still yield valid population estimates when the randomization probabilities are incorporated into the estimation procedure.

3. Maddala's Framework for Logistic Regression under Randomization

Maddala [9], in his seminal work on limited dependent variable models, showed how regression analysis can be adapted when binary outcomes are subject to misclassification or perturbation. In the context of randomized response, the observed variable is not the true sensitive outcome Y , but a randomized surrogate Z generated through a known probability mechanism.

Suppose the true outcome Y_i for individual i follows a logistic regression model:

$$P\left(Y_i = 1 \mid x_i\right) = \frac{\exp\left(\beta^T x_i\right)}{1 + \exp\left(\beta^T x_i\right)} \tag{4}$$

where x_i is the covariate vector and β is the parameter vector. Under randomized response, the observed response Z_i relates to Y_i through known misclassification probabilities determined by the device. Maddala's contribution was to show that the likelihood of the observed data can be expressed in terms of these known probabilities and the logistic model for Y_i. Maximization of this likelihood yields consistent estimators of β, despite the randomization. This framework forms the foundation for regression analysis in randomized response models.

4. Zero-Truncated Negative Binomial (ZTNB) Device

Building on distribution-based randomization mechanisms, Neelam [8] introduced the Zero-Truncated Negative Binomial (ZTNB) distribution as a randomization device for sensitive surveys. The ZTNB distribution excludes zero outcomes, which enhances respondent anonymity, and incorporates a dispersion parameter that accommodates overdispersion frequently present in survey data.

Under this device, if the true attribute is $Y = 1$, the randomized response Z follows a ZTNB distribution with parameters $\left(r_1, p_1\right)$, while if $Y = 0$, it follows another ZTNB distribution with parameters $\left(r_2, p_2\right)$. The probability mass function for the ZTNB distribution is

$$P\left(Z = z\right) = \frac{\binom{r+z-1}{z} p^r \left(1-p\right)^z}{1 - p^r}, \qquad z = 1, 2, ... \tag{5}$$

The observed distribution of Z is therefore a mixture:

$$P(Z = z) = \pi f_{ZTNB}\left(z; r_1, p_1\right) + \left(1 - \pi\right) f_{ZTNB}\left(z; r_2, p_2\right) \tag{6}$$

Based on this mixture distribution, the unbiased estimator of the population proportion is obtained as

$$\hat{\pi}_{ZTNB} = \frac{p_1(1-p_1^{r_1})p_2(1-p_2^{r_2})\overline{z} - r_2(1-p_2)p_1(1-p_1^{r_1})}{r_1(1-p_1)p_2(1-p_2^{r_2}) - r_2(1-p_2)p_1(1-p_1^{r_1})} \tag{7}$$

With variance

$$Var(\hat{\pi}_{ZTNB}) = \frac{\pi(1-\pi)}{n} +$$

$$\frac{\pi r_1(1-p_1)p_2^{\,2}(1-p_2^{\,r_2})^2\left[1-(1+r_1(1-p_1))p_1^{\,r_1}\right]+(1-\pi)r_2(1-p_2)p_1^{\,2}(1-p_1^{\,r_1})^2\left[1-(1+r_2(1-p_2))p_2^{\,r_2}\right]}{n[r_1(1-p_1)p_2(1-p_2^{\,r_2})-r_2(1-p_2)p_1(1-p_1^{\,r_1})]^2} \tag{8}$$

5. Proposed Work

Logistic Regression with ZTNB Randomization

The distinctive contribution of this study is to extend the Zero-Truncated Negative Binomial (ZTNB) device from simple proportion estimation to a logistic regression framework with covariates. Unlike the classical regression context where the sensitive trait Y_i is observed directly, here the analyst only observes the randomized response Z_i generated via the ZTNB mechanism.

Accordingly, the likelihood function for estimation is constructed by integrating the ZTNB probability mass functions, corresponding to $Y_i = 1$ and $Y_i = 0$, with the logistic model–based probabilities of the latent trait. This yields a mixture likelihood that embeds both the randomization probabilities and the regression structure.

Maximization of this likelihood provides the estimators of β, which retain the desirable properties of consistency and asymptotic normality under regularity conditions. Furthermore, the presence of the ZTNB dispersion parameter introduces additional flexibility, allowing the model to accommodate overdispersed randomized responses, a limitation in earlier ZTP-based extensions.

We now extend the Zero-Truncated Negative Binomial (ZTNB) randomization device to the case of binary logistic regression with covariates. Let $Y_i \in \{0,1\}$ denote the sensitive attribute for respondent i. Conditional on covariates X_i, the distribution of Y_i follows the usual logistic regression model:

$$\pi_i = P(Y_i = 1 \mid x_i) = \frac{\exp(\beta^T x_i)}{1+\exp(\beta^T x_i)}$$

where β is the vector of regression coefficients to be estimated.

In the randomized response setting, the analyst does not observe Y_i directly. Instead, each respondent's reported outcome Z_i is generated through the ZTNB device:

If $Y_i = 1$, then $Z_i \sim ZTNB(r_1, p_1)$.

If $Y_i = 0$, then $Z_i \sim ZTNB(r_2, p_2)$.

The probability mass function of the ZTNB distribution is as given in (5):

$$f_{ZTNB}\left(z;r_y,p_y\right)=\frac{\binom{r_y+z-1}{z}p_y^{r_y}\left(1-p_y\right)^z}{1-p_y^{r_y}},\quad z=1,2,...$$

where $\left(r_y,p_y\right)$ are the parameters associated with the randomization device for outcome. $r>0$ is the number of successes, $p\in\left(0,1\right)$ is the success probability, and the denominator $1-p^r$ ensures truncation at zero.

Hence, the unconditional distribution of Z_i, given covariates X_i, is a mixture:

$$D_i\left(\beta\right)=\pi_i\left(\beta\right)f_{ZTNB}\left(z\,|\,r_1,p_1\right)+\left(1-\pi_i\left(\beta\right)\right)f_{ZTNB}\left(z\,|\,r_2,p_2\right)$$

Thus the sample likelihood is

$$L\left(\beta\right)=\prod_{i=1}^{n}\left\{\pi_i\left(\beta\right)f_{ZTNB}\left(Z_i;r_1,p_1\right)+\left(1-\pi_i\left(\beta\right)\right)f_{ZTNB}\left(Z_i;r_2,p_2\right)\right\}. \tag{9}$$

The log-likelihood for a sample of n respondents is then:

$$l\left(\beta\right)=\sum_{i=1}^{n}\log\left[\pi_i\left(\beta\right)f_{ZTNB}\left(Z_i\,|\,r_1,p_1\right)+\left(1-\pi_i\left(\beta\right)\right)f_{ZTNB}\left(Z_i\,|\,r_2,p_2\right)\right] \tag{10}$$

where $l\left(\beta\right)$ represents the log-likelihood function.
By differentiating $l\left(\beta\right)$ with respect to β, we arrive at the score vector given by

$$U_n\left(\beta\right)=\frac{\partial}{\partial\beta}l\left(\beta\right)=\frac{\partial}{\partial\beta}\left[\sum_{i=1}^{n}\log\left\{\pi_i\left(\beta\right)f_{ZTNB}\left(Z_i\,|\,r_1,p_1\right)+\left(1-\pi_i\left(\beta\right)\right)f_{ZTNB}\left(Z_i\,|\,r_2,p_2\right)\right\}\right]. \tag{11}$$

$$=\sum_{i=1}^{n}\frac{\frac{\partial\pi_i\left(\beta\right)}{\partial\beta}f_{1i}-\frac{\partial\pi_i\left(\beta\right)}{\partial\beta}f_{2i}}{\pi_i\left(\beta\right)f_{1i}+\left(1-\pi_i\left(\beta\right)\right)f_{2i}} \tag{12}$$

$$=\sum_{i=1}^{n}\psi_i\left(\beta\right)$$

where

$$\psi_i\left(\beta\right)=\left(w_i-\pi_i\left(\beta\right)\right)x_i,f_{1i}=f_{ZTNB}\left(Z_i\,|\,r_1,p_1\right),f_{2i}=f_{ZTNB}\left(Z_i\,|\,r_2,p_2\right)$$

and

$$w_i=P\left(Y_i=1\,|\,Z_i\right)=\frac{\pi_i\left(\beta\right)f_{1i}}{\pi_i\left(\beta\right)f_{1i}+\left(1-\pi_i\left(\beta\right)\right)f_{2i}}$$

Where w_i is the posterior probability.

To facilitate the derivation of asymptotic properties, the score function can be re-expressed in a variance-weighted form that highlights the role of the ZTNB expectation and its variability, as presented in Lemma 1.

Lemma 1

For the ZTNB-based logistic regression model, the individual score contribution can be equivalently written as

$$\psi_i(\beta)=\left(\frac{\partial g_i(\beta)}{\partial \beta}\right)^T V_i^{-1}(\beta)\left(Z_i-g_i(\beta)\right), \qquad i=1,2,\ldots,n$$

where

- Z_i is the observed randomized response generated by the ZTNB device,
- $g_i(\beta)=E\left[Z_i \mid x_i,\beta\right]$ is the ZTNB-based expectation, and
- $V_i(\beta)=Var\left(Z_i \mid x_i,\beta\right)$ is the corresponding variance.

A full proof of Lemma 1 is provided in the Appendix. The subsequent section is devoted to establishing the asymptotic properties of the maximum likelihood estimators.

Large Sample Properties

Let the overall score function be defined as

$$U_n(\beta)=\sum_{i=1}^{n}\psi_i(\beta)$$

Where $\psi_i(\beta)$ is the individual contribution derived in Lemma 1. To study the asymptotic behavior of the maximum likelihood estimator $\hat{\beta}$, the following assumptions are imposed.

Regularity conditions

(C1) The data $(Z_i,X_i), i=1,2,\ldots,n,$ are independent and identically distributed. The covariates satisfy $E\|X_i\|^2<\infty$. The device parameters (r_1,p_1,r_2,p_2) of the ZTNB distribution are fixed and known by design.

(C2) The true parameter β_0 lies in the interior of a compact parameter space $\Theta \subset \mathbb{R}^p$.

(C3) The mixture pmf

$$D_i(\beta)=\pi_i(\beta)f_{1i}+\left(1-\pi_i(\beta)\right)f_{2i}$$

satifies $D_i(\beta)=D_i(\beta_0)$ with positive probability only if $\beta=\beta_0$.

(C4) The log-likelihood contribution $l_i(\beta)=\log D_i(\beta)$ is three-times continuously differentiable in a neighborhood of β_0. All derivatives up to order two are dominated by an integrable envelope function, permitting differentiation under the expectation operator.

(C5) The Fisher information matrix

$$I\left(\beta_0\right)=E\left[\psi_i\left(\beta_0\right)\psi_i\left(\beta_0\right)^T\right]$$

Is positive definite.

(C6) The conditional mean $g_i\left(\beta\right)=E\left[Z_i \mid X_i\right]$ and variance $V_i\left(\beta\right)=Var\left(Z_i \mid X_i\right)$ under the ZTNB device are finite in a neighborhood of β_0 and $V_i\left(\beta\right)$ is non-degenerate.

The asymptotic properties of the maximum likelihood estimator $\hat{\beta}$ are summarized in the following theorem.

Theorem 1

Under the regularity conditions (C1)-(C6) for likelihood inference, the estimator $\hat{\beta}$ possesses the following properties:

1. As the sample size $n \to \infty$, the estimator converges in probability to the true parameter value, i.e,

$$\hat{\beta}\xrightarrow{p}\beta$$

2. The scaled deviation of the estimator from the true parameter vector is asymptotically normal,

$$\sqrt{n}\left(\hat{\beta}-\beta\right)\xrightarrow{d}N\left(0,I^{-1}\left(\beta\right)\right)$$

where $I\left(\beta\right)$ is the Fisher Information matrix per observation, defined as

$$I\left(\beta\right)=E\left[-\frac{\partial^2 l_i\left(\beta\right)}{\partial\beta\partial\beta^T}\right]$$

With $l_i\left(\beta\right)$ denoting the individual log-likelihood contribution.

Complete proof of theorem is given in the appendix.

6. Simulation Study

The simulation study aims to evaluate the finite-sample performance of the proposed Zero-Truncated Negative Binomial (ZTNB) randomized response model under logistic regression. Specifically, the bias, standard deviation (SD), asymptotic standard error (ASE), and coverage probability of the estimated regression parameters are assessed. The results are compared across a realistic parameter setting.

Simulation Design

In order to evaluate the finite-sample performance of the proposed estimator, data were generated under a controlled setting. A single covariate X_i was drawn from the standard normal distribution. The true binary response Y_i was then generated according to the logistic regression model

$$\pi_i=P\left(Y_i=1 \mid x_i\right)=\frac{\exp\left(\beta^T x_i\right)}{1+\exp\left(\beta^T x_i\right)}$$

where β_0 and β_1 denote the regression coefficients. The observed randomized response Z_i was subsequently drawn from a mixture of two zero-truncated negative binomial distributions, conditional on the latent value of Y_i. Specifically, if $Y_i = 1$, then $Z_i \square ZTNB(r_1, p_1)$ whereas if $Y_i = 0$, then $Z_i \square ZTNB(r_2, p_2)$. For the baseline configuration, the parameters were fixed at $(r_1.p_1) = (2, 0.3)$ for the sensitive group and $(r_2, p_2) = (1, 0.5)$ for the non-sensitive group. This specification ensures sufficient distinction between the two response distributions while preserving realistic variability.

The probability mass function of a zero-truncated negative binomial distribution with parameters (r, p) is given by

$$f_{ZTNB}(z \mid r, p) = \frac{\binom{r+z-1}{z} p^r (1-p)^z}{1-p^r}, \qquad z = 1, 2, ...$$

where truncation at zero is imposed to maintain consistency with the randomized response framework. Each simulated dataset consisted of $n = 1000, 2000, 3000$ independent observations, and the entire simulation was repeated $r = 2000$ times in order to assess the empirical properties of the estimators. The true regression coefficients were set to $\beta = (\beta_0, \beta_1) = (0, -0.1), (0, 0.1).(0.1, 0.1), (-0.1, -0.1), (0, 0.05), (0, -0.05), (-0.2, 0.1), (0.2, -0.1),$

$(-0.1, 0.2), (0.1, -0.2)$.

Estimation

Estimation was based solely on the observed pairs (Z_i, X_i), with the latent responses Y_i integrated out through the mixture representation. The likelihood contribution for observation *i* can be written as

$$L_i(\beta) = \pi_i(\beta) f_{ZTNB}(Z_i \mid r_1, p_1) + (1 - \pi_i(\beta)) f_{ZTNB}(Z_i \mid r_2, p_2)$$

where $\pi_i(\beta)$ is defined as in the logistic model. The overall log-likelihood function is therefore

$$l(\beta) = \sum_{i=1}^{n} \log L_i(\beta)$$

Maximum likelihood estimation was carried out by numerically maximizing $l(\beta)$ with respect to the regression coefficients $\beta = (\beta_0, \beta_1)^T$. Standard errors were obtained from the inverse of the observed information matrix, that is, the negative Hessian of the log-likelihood evaluated at the maximum likelihood estimates. Wald-type confidence intervals at the 95% nominal level were constructed using these estimates.

For each parameter, empirical performance was assessed across simulation replications using the average estimate (Mean), bias relative to the true parameter, average model-based standard error (ASE), empirical standard deviation (SD) and coverage probability of the confidence intervals. These measures provide a comprehensive evaluation of estimator accuracy, variability, and reliability.

Table 1: Simulation results of the ZTNB logistic regression model for $n = 1000$

	True	Mean	Bias	ASE	SD	CP
β_0	0	0.0.00565	0.00565	0.11967	0.12019	0.949
β_1	-0.1	-0.10709	-0.00709	0.12096	0.12271	0.945
β_0	0	0.00208	0.00208	0.11960	0.11771	0.961
β_1	0.1	0.10227	0.00227	0.12091	0.12013	0.958
β_0	0.1	0.10302	0.00302	0.12065	0.11931	0.953
β_1	0.1	0.10051	0.00051	0.12194	0.12158	0.9615
β_0	-0.1	-0.09553	0.00447	0.11906	0.12026	0.9475
β_1	-0.1	-0.10530	-0.00530	0.12032	0.12073	0.9535
β_0	0	0.00243	0.00243	0.11942	0.11744	0.9585
β_1	0.05	0.05016	0.00016	0.12037	0.11945	0.953
β_0	0	0.00520	0.00520	0.11947	0.12033	0.945
β_1	-0.05	-0.05499	-0.00499	0.12035	0.12069	0.948
β_0	-0.2	-0.20058	-0.00058	0.11889	0.11918	0.9505
β_1	0.1	0.09955	-0.00045	0.12016	0.11904	0.953
β_0	0.2	0.20832	0.00832	0.12233	0.12375	0.9475
β_1	-0.1	-0.10477	-0.00477	0.12346	0.12145	0.9565
β_0	-0.1	-0.10004	-0.00003	0.11971	0.11896	0.9565
β_1	0.2	0.20349	0.00349	0.12234	0.12258	0.9525
β_0	0.1	0.10793	0.00793	0.12153	0.12189	0.955
β_1	-0.2	-0.20857	-0.00857	0.12417	0.12566	0.946

Table 2: Simulation results of the ZTNB logistic regression model for $n = 2000$

	True	Mean	Bias	ASE	SD	CP
β_0	0	0.000153	0.000153	0.08427	0.08314	0.9555
β_1	-0.1	-0.10277	-0.00277	0.08477	0.08294	0.9605
β_0	0	0.00261	0.00261	0.08427	0.08279	0.9515
β_1	0.1	0.09993	-0.00007	0.08481	0.08297	0.959
β_0	0.1	0.10440	0.00440	0.08503	0.08574	0.953
β_1	0.1	0.09917	-0.00083	0.08556	0.08445	0.9605
β_0	-0.1	-0.10157	-0.00157	0.08386	0.08298	0.9535
β_1	-0.1	-0.10193	-0.00193	0.08437	0.08320	0.953

β_0	0	0.00284	0.00284	0.08415	0.08312	0.949
β_1	0.05	0.04969	-0.00031	0.08445	0.08246	0.961
β_0	0	-0.00014	-0.00014	0.08414	0.08307	0.9565
β_1	-0.05	-0.05261	-0.00261	0.08442	0.08211	0.9625
β_0	-0.2	-0.20166	-0.00166	0.08379	0.08258	0.9565
β_1	0.1	0.09894	-0.00106	0.08429	0.08199	0.96
β_0	0.2	0.20436	0.00436	0.08610	0.08485	0.953
β_1	-0.1	-0.10200	-0.00200	0.08662	0.08694	0.9505
β_0	-0.1	-0.09828	0.00172	0.08434	0.08338	0.956
β_1	0.2	0.20155	0.00155	0.08583	0.08435	0.954
β_0	0.1	0.10108	0.00108	0.08553	0.08419	0.9535
β_1	-0.2	-0.20432	-0.00432	0.08694	0.08512	0.954

Table 3: Simulation results of the ZTNB logistic regression model for $n = 3000$

	True	Mean	Bias	ASE	SD	CP
β_0	0	0.00172	0.00172	0.06874	0.06868	0.954
β_1	-0.1	-0.10118	-0.00118	0.06913	0.06868	0.954
β_0	0	-0.000006	-0.000006	0.06873	0.06919	0.952
β_1	0.1	0.10021	0.00021	0.06918	0.07037	0.945
β_0	0.1	0.10038	0.00038	0.06933	0.07022	0.951
β_1	0.1	0.09973	-0.00027	0.06975	0.07054	0.946
β_0	-0.1	-0.09868	0.00132	0.06840	0.06862	0.9485
β_1	-0.1	-0.10059	-0.00059	0.06880	0.06855	0.946
β_0	0	-0.00040	-0.00040	0.06863	0.06924	0.946
β_1	0.05	0.04966	-0.00033	0.06888	0.07001	0.9465
β_0	0	0.00209	0.00209	0.06864	0.06874	0.9545
β_1	-0.05	-0.05075	-0.00075	0.06885	0.06926	0.95
β_0	-0.2	-0.19898	0.00102	0.06833	0.06817	0.9485
β_1	0.1	0.10053	0.00053	0.06876	0.06930	0.951
β_0	0.2	0.20081	0.00081	0.07020	0.07084	0.947
β_1	-0.1	-0.10162	-0.00162	0.07059	0.07166	0.9465
β_0	-0.1	-0.09963	0.00037	0.06878	0.07013	0.9405
β_1	0.2	0.20069	0.00069	0.06999	0.07176	0.943

β_0	0.1	0.10148	0.00148	0.06976	0.07112	0.9505
β_1	-0.2	-0.20167	-0.00167	0.07087	0.07062	0.9505

7. Results and Interpretation

The results of the simulation study provide a comprehensive evaluation of the finite-sample behavior of the proposed ZTNB logistic regression estimator. The mean of the simulated estimates is consistently close to the corresponding true parameter values, indicating that the estimator successfully recovers the underlying regression coefficients across repeated samples. The empirical bias remains very small in all cases, typically well below 0.01 in absolute magnitude, which demonstrates that the estimator is essentially unbiased even in moderate sample sizes. The observed variability of the estimates, measured by the empirical standard deviation, aligns closely with the model-based variability represented by the average standard error (ASE), derived from the asymptotic variance formula based on the observed information matrix. The close agreement between ASE and SD across all settings confirms that the theoretical variance approximation is valid in practice. Furthermore, the coverage probabilities of the nominal 95% confidence intervals remain stable and close to the intended level, generally fluctuating around 0.95, with no indication of systematic under- or over-coverage.

When considering different sample sizes, the results in Tables 1 to 3 reveal the expected efficiency gains with larger n. At $n=1000$, the estimator already performs well, with negligible bias and reliable confidence interval coverage, though the variability of the estimates is somewhat larger. As the sample size increases to 2000 and then 3000, both ASE and SD decrease markedly, illustrating the precision improvements predicted by asymptotic theory. By $n=3000$, the empirical bias is virtually eliminated, often appearing at the order of 10^{-4}, and ASE and SD are nearly indistinguishable. Coverage probabilities also stabilize tightly around the nominal level, showing that inference based on the proposed model remains valid across all sample sizes.

Across the range of coefficient settings considered, including both positive and negative values of β_0 and β_1, the qualitative performance of the estimator remains unchanged. The patterns of negligible bias, close agreement between ASE and SD, and nominal coverage are consistently observed, demonstrating that the estimator is robust to different underlying logistic structures. The simulation study therefore confirms that the proposed ZTNB logistic regression model provides accurate and efficient parameter estimation, with reliable inference and strong finite-sample performance. Importantly, when contrasted with the previously studied ZTP-based model, the ZTNB estimator shows slightly improved stability and reduced variability, particularly at smaller sample sizes, highlighting its potential as a more flexible and efficient randomization device for sensitive survey data.

8. Conclusion

This article advances the literature on randomized response techniques (RRT) by formally extending the Zero-Truncated Negative Binomial (ZTNB) distribution to logistic regression with covariates. Unlike existing approaches such as Warner's model, Kuk's design, Singh and Grewal's geometric device, and more recent zero-truncated proposals, the ZTNB framework incorporates an additional dispersion parameter that accommodates overdispersion and heterogeneity in sensitive survey data. Theoretical derivations establish the consistency and asymptotic normality of the maximum likelihood estimators, while simulation studies confirm their excellent finite-sample performance in terms of negligible bias, close agreement between empirical and model-based standard errors, and valid coverage probabilities. Importantly, the ZTNB model improves efficiency relative to its ZTP counterpart, particularly under heterogeneous response settings, without compromising respondent protection. These results highlight the ZTNB logistic regression model as a robust and flexible inferential tool for analyzing sensitive survey data with covariates, offering both methodological innovation and practical utility for applied research in health, criminology, sociology, and economics.

References

[1]. Christofides, T. C. (2005). Randomized response technique for two sensitive characteristics at the same time. *Metrika*, 62(1), 53-63.

[2]. Dietz, P., Striegel, H., Franke, A. G., Lieb, K., Simon, P., and Ulrich, R. (2013). Randomized response estimates for the 12-month prevalence of cognitive-enhancing drug use in university students. *Pharmaco: The Jour of Hum Pharma and Drug Ther*, 33(1), 44-50.

[3]. Greenberg, B. G.; Abul-Ela, A. L. A.; Simmons, W. R., and Horvitz, D. G. (1969). The unrelated question randomized response model: Theoretical framework. *Jour of the Amer Stat Assoc*, 64(326), 520-539.

[4]. Hsieh, S. H., Lee, S. M., and Shen, P. S. (2009). Semiparametric analysis of randomized response data with missing covariates in logistic regression. *Comp Stat & Data Anal*, 53(7), 2673-2692.

[5]. Hsieh, S. H., Lee, S. M., and Shen, P. S. (2010). Logistic regression analysis of randomized response data with missing covariates. *Jour of Stat Plan and Inf*, 140(4), 927-940.

[6]. Hsieh, S. H., Lee, S. M., Li, C. S., and Tu, S. H. (2016). An alternative to unrelated randomized response techniques with logistic regression analysis. *Stat meth & appl*, 25(4), 601-621.

[7]. Hsieh, S. H., & Perri, P. F. (2022). A logistic regression extension for the randomized response simple and crossed models: Theoretical results and empirical evidence. *Sociological Methods & Research*, 51(3), 1244-1281.Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive
answer bias. *Jour Amer Stat Assoc*, 60(309), 63-69.

[8]. Kuk, A. Y. (1990). Asking sensitive questions indirectly. *Biometrika*, 77, 436-438

[9]. Lensvelt-Mulders, G. J., Hox, J. J., Van der Heijden, P. G., & Maas, C. J. (2005). Meta-analysis of randomized response research: Thirty-five years of validation. *Sociol Meth & Res,* 33(3), 319-348.

[10]. Le, T. N., Lee, S. M., Tran, P. L., and Li, C. S. (2023). Randomized response techniques: a systematic review from the pioneering work of warner (1965) to the present. *Mathe*, 11(7), 1718.

[11]. Maddala GS. Limited-dependent and qualitative variables in econometrics. *Camb Uni Press, Camb* **1983**

[12]. Mangat, N. S. (1994). An improved randomized response strategy. *Jour of the Rol Stat Soc: Seri B (Method)*, 56(1), 93-95.

[13]. Magder, L. S., and Hughes, J. P. (1997). Logistic regression when the outcome is measured with uncertainty. *Ameri jour of epidem*, 146(2), 195-203.

[14]. Neuhaus, J. M. (1999). Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika*, 86(4), 843-855.

[15]. Neelam., SM Asim., Murad, S., and Shah, S.F. (2025). Extending the Zero-Truncated Poisson Randomized Response Device to Logistic Regression with Covariates. *Jour for Curr Sign*, 3(3), 1088–1109.

[16]. Neelam. (2025) A Zero-Truncated Poisson Model for Sensitive Data Estimation in Randomized Response Surveys [Unpublished manuscript]. Department of Statistics, University of Peshawar, Pakistan.

[17]. Neelam. (2025) Design and Evaluation of a ZTNB-Based Randomization Device for Sensitive Attributes [Unpublished manuscript]. Department of Statistics, University of Peshawar, Pakistan.

[18]. Scheers, N. J., and Dayton, C. M. (1988). Covariate randomized response models. *Jour of the Amer Stat Assoc*, 83(404), 969-974.

[19]. Singh, S., and Grewal, I. S. (2013). Geometric distribution as a randomization device: implemented to the Kuk's model. *Inter Jour of Cont Math Sci***,** 8(5), 243-248.

[20]. Van den Hout, A. D. L., Van der Heijden, P. G., and Gilchrist, B. (2007). A multivariate logistic regression model for randomized response data. In *Proceedings of the 21th Workshop on Stat Model,* 234-241.

[21]. Van den Hout, A., van der Heijden, P. G., and Gilchrist, R. (2007). The logistic regression model with response variables subject to randomized response. *Comp Stat & Data Anal*, 51(12), 6060-6069.

[22]. Williams, B. L., Suen, H. K., and Baffi, C. R. (1993). A controlled randomized response technique. *Eval & the Health Profess*, 16(2), 225-238.

[23]. Zapata, Z., Sedory, S. A., & Singh, S. (2022). Zero-truncated binomial distribution as a randomization device. *Sociol Meth & Res*, 51(2), 800-815.

Appendix

Lemma 1

Proof

$\psi_i(\beta)$ in (12) can be written as

$$\psi_i(\beta)=\frac{\pi_i(\beta)\left(1-\pi_i(\beta)\right)x_i\left(f_{1i}-f_{2i}\right)}{\pi_i(\beta)f_{1i}+\left(1-\pi_i(\beta)\right)f_{2i}} \tag{13}$$

As $\dfrac{\partial\pi_i(\beta)}{\partial\beta}=\pi_i(\beta)\left(1-\pi_i(\beta)\right)x_i$ (14)

Since the conditional mean is $g_i(\beta)=\pi_i(\beta)\mu_1+\left(1-\pi_i(\beta)\right)\mu_2$, where $\mu_k=E\left[Z\mid r_k,p_k\right]$

Differentiating $g_i(\beta)$, we get

$$\frac{\partial g_i(\beta)}{\partial\beta}=\left(\mu_1-\mu_2\right)\frac{\partial\pi_i(\beta)}{\partial\beta} \tag{15}$$

From exponential family theory, the log likelihood derivative wrt $\pi_i(\beta)$ can be written in terms of the centered outcome $\left(Z_i-g_i\right)$ scaled by its variance, i.e,

$$\frac{f_{1i}-f_{2i}}{\pi_i(\beta)f_{1i}+\left(1-\pi_i(\beta)\right)f_{2i}}=\frac{Z_i-g_i(\beta)}{V_i(\beta)}\left(\mu_1-\mu_2\right) \tag{16}$$

Using (16) into (13) We have

$$\psi_i(\beta)=\pi_i(\beta)\left(1-\pi_i(\beta)\right)x_i\left(\mu_1-\mu_2\right)\frac{Z_i-g_i(\beta)}{V_i(\beta)} \tag{17}$$

Using (14) in (17)

$$\psi_i(\beta)=\frac{\partial\pi_i(\beta)}{\partial\beta}\left(\mu_1-\mu_2\right)\frac{Z_i-g_i(\beta)}{V_i(\beta)} \tag{18}$$

Using (15) in (18) we get

$$\psi_i(\beta)=\left(\frac{\partial g_i(\beta)}{\partial\beta}\right)^T V_i^{-1}(\beta)\left(Z_i-g_i(\beta)\right)$$

Which is exactly the lemma form.

Proof of Theorem 1

Let $l_n(\beta)=\sum_{i=1}^{n} l_i(\beta)$ denote the log-likelihood function, where $l_i(\beta)=\log D_i(\beta)$ is the individual contribution under the ZTNB randomized response device. The score is $U_n(\beta)=\frac{\partial l_n(\beta)}{\partial\beta}$ and the observed Hessian is $H_n(\beta)=\frac{\partial^2 l_n(\beta)}{\partial\beta\partial\beta^T}$.

(a) Consistency

Defining the normalized log-likelihood function as:

$$Q_n(\beta)=\frac{1}{n}l_n(\beta)=\frac{1}{n}\sum_{i=1}^{n} l_i(\beta) \tag{19}$$

By (C1)–(C4), the observations are i.i.d. and $l_i(\beta)$ is measurable and dominated by an integrable envelope. Thus, by the Law of Large Numbers,

$$Q_n(\beta)\xrightarrow{p}Q(\beta)=E\left[l_i(\beta)\right] \text{ uniformly in } \beta.$$

By (C3), the mixture pmf is identifiable, so $Q(\beta)$ is uniquely maximized at the true parameter β_0 . By (C2), the parameter space is compact, ensuring that $\hat{\beta}=\arg\max Q_n(\beta)$ exists.

Hence,

$$\hat{\beta}\xrightarrow{p}\beta_0$$

which proves consistency**.**

(b) Asymptotic Normality

The score function satisfies

$$U_n(\beta_0)=\sum_{i=1}^{n} U_i(\beta_0), \quad \text{with } E\left[U_i(\beta_0)\right]=0$$

By (C1)–(C4), $\left\{U_i\left(\beta_0\right)\right\}$ are i.i.d. with finite variance. Hence, by the Central Limit Theorem,

$$\frac{1}{\sqrt{n}}U_n\left(\beta_0\right)\xrightarrow{d}N\left(0;I\left(\beta_0\right)\right)$$

where

$$I\left(\beta_0\right)=E\left[\psi_i\left(\beta_0\right)\psi_i\left(\beta_0\right)^T\right]$$

Is the Fisher Information per observation.

By Taylor expansion of the overall score about β_0 there exists $\tilde{\beta}$ on the line segment between $\hat{\beta}$ and β_0 such that

$$0=U_n\left(\hat{\beta}\right)=U_n\left(\beta_0\right)+H_n\left(\tilde{\beta}\right)\left(\hat{\beta}-\beta_0\right) \tag{20}$$

Rearranging and scaling

$$\sqrt{n}\left(\hat{\beta}-\beta_0\right)=-\left\{-\frac{1}{n}H_n\left(\tilde{\beta}\right)\right\}^{-1}\frac{1}{\sqrt{n}}U_n\left(\beta_0\right)$$

By (C4)–(C5) and the Law of Large Numbers,

$$-\frac{1}{n}H_n\left(\tilde{\beta}\right)\xrightarrow{p}I\left(\beta_0\right)$$

and $I\left(\beta_0\right)$ is positive definite by (C5). Combining the CLT limit for $n^{-1/2}U_n\left(\beta_0\right)$ with Slutsky's theorem yields

$$\sqrt{n}\left(\hat{\beta}-\beta_0\right)\xrightarrow{d}I\left(\beta_0\right)^{-1}N\left(0,I\left(\beta_0\right)\right)=N\left(0,I\left(\beta_0\right)^{-1}\right)$$

Hence $\hat{\beta}$ is asymptotically normal with covariance matrix $I\left(\beta_0\right)^{-1}$.