RESEARCH REPORT SERIES
(Statistics #2023-03)

Analysis of One-Way ANOVA Model Using Synthetic Data

Biswajit Basak[1],
Bimal Sinha[2, 3]

[1]Department of Statistics, Sister Nivedita University, Kolkata, India
[2]University of Maryland, Baltimore County, Maryland, USA
[3]Center for Statistical Research and Methodology, U.S. Census Bureau, USA

Center for Statistical Research & Methodology
Research and Methodology Directorate
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: April 28, 2023 (Revised: October 17, 2023)

Disclaimer: This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not those of the U.S. Census Bureau.

Analysis of One-Way ANOVA Model using Synthetic Data

Biswajit Basak*[1] and Bimal Sinha[2,3]

[1]Department of Statistics, Sister Nivedita University, Kolkata, India
[2]University of Maryland, Baltimore County, Maryland, USA
[3]Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, DC 20233, USA

Abstract

In this paper we consider the age-old ANOVA problem of testing the equality of means of several univariate normal populations with a common unknown variance, except that the data used for analysis arise from a synthetic version of the original observations. We address two versions of the synthetic data: one obtained under Plug-In sampling(PIS) method and the other under Posterior Predictive Sampling(PPS) method. We study its distributional properties (null and non-null) and provide enough computational details. A comparison of power is also provided. As expected, the power under the PIS method is more than that under the PPS method. A measure of privacy protection is also evaluated and it turns out that the PIS method provides less protection than the PPS method, thus confirming the standard belief that accuracy of inference and privacy protection work in opposite directions. Robustness of the proposed tests under deviations from normality is also studied.

Keywords: ANOVA problem, non-central F distribution, plug-in sampling, posterior predictive sampling, privacy protection

1 Introduction

Statistical agencies dealing with collection and publication of relevant data often face the problem of releasing microdata for public use in view of compromising with the privacy of survey respondents. Most often therefore data are summarized and presented in tabular forms. However, some data users and policy stakeholders may also want to use the microdata to carry out other forms of data analysis, different from what the agencies release. This calls for release of microdata under some perturbation mechanism to ensure privacy protection. Statistical literature is quite rich in terms of data perturbation methods and subsequent data analysis techniques based on perturbed data. Some commonly used data perturbation methods include: noise addition/multiplication ([1], [2], [3], [4]), model-based multiply imputed synthetic data methods ([5], [7], [8], [9]). While the inferential methods developed by Reiter ([8], [9], [10], [11], [12], [13]) are essentially asymptotic in nature, Klein and Sinha ([14], [15], [16]) developed exact data analysis methods for singly imputed synthetic data based on some simple parametric models under two popular types of synthetic data generation.

*Corresponding author :
Biswajit Basak, Sister Nivedita University, Kolkata, India. Email: biswajitbasak.stat@gmail.com

In this paper we consider the age-old ANOVA problem of testing the equality of means of several univariate normal populations with a common unknown variance, except that the data used for analysis arise from a synthetic version of the original observations. Consider k random samples $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k$ where the i^{th} sample $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in_i})$, $i = 1, 2, \ldots, k$, is coming from $N(\mu_i, \sigma^2)$ distribution. Define $N = \sum_{i=1}^{k} n_i$, $\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$, $\bar{x}_w = \frac{1}{N}\sum_{i=1}^{k} n_i\bar{x}_i$, and $S_{\boldsymbol{x}}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$. Our main objective is to develop a testing strategy to test the equality of these k means, which is given by

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k \qquad \text{against} \qquad H_1 : \text{At least one inequality in } H_0. \tag{1}$$

It is well known that, based on the original data $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k)$, the likelihood ratio test (LRT) is provided by the standard F-statistic defined as $F_{\boldsymbol{x}} = \left(\frac{N-k}{k-1}\right)\frac{\text{BSS}}{\text{WSS}}$, where $\text{BSS} = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x}_w)^2$ and $\text{WSS} = S_{\boldsymbol{x}}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$.

Our goal is now to suitably perturb the original data in view of privacy protection requirement and provide appropriate analysis of the resultant perturbed data. As mentioned earlier, there are a variety of methods in the statistics literature to accomplish this task. Here we consider two methods of generating synthetic data and provide appropriate valid inference for testing H_0 based on both types of synthetic data. In Section 2 we discuss Plug-In Sampling method while in Section 3 the Posterior Predictive Sampling method. Our inference is essentially based on the usual F-statistic based on the synthetic data and we study its distributional properties (null and non-null) in both the cases. Computational details and a comparison of power are provided in Section 4. Section 5 is devoted to a discussion of privacy protection offered by the above data perturbation methods. As expected, PIS offers better inference and less privacy protection compared to PPS. Robustness of the proposed F tests under deviations from normality is studied in Section 6 and some concluding remarks are mentioned in Section 7.

2 Plug-In Sampling(PIS) Method

Here we consider the unbiased estimates of μ_i's $(i = 1, 2, \ldots, k)$ and σ^2 based on the original data $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k)$, as $\hat{\mu}_i = \bar{x}_i$ and $\hat{\sigma}^2 = \frac{S_{\boldsymbol{x}}^2}{N-k}$, and hence draw k independent samples $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_k$ where the i^{th} sample $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})$ is a random sample from $N(\bar{x}_i, \hat{\sigma}^2)$ distribution. Define $\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$, $\bar{y}_w = \frac{1}{N}\sum_{i=1}^{k} n_i\bar{y}_i$, Within Sum of Squares(WSS($\boldsymbol{y}$)) $= S_{\boldsymbol{y}}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$, and Between Sum of Squares(BSS($\boldsymbol{y}$)) $= \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y}_w)^2$. Note that $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k, S_{\boldsymbol{y}}^2$ are jointly sufficient for $(\mu_1, \mu_2, \ldots, \mu_k, \sigma^2)$ based on the synthetic data $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_k)$ obtained by the above method (see [14] for a very general method). Next we consider the joint pdf of the sufficient statistic $(\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k, S_{\boldsymbol{y}}^2)$ in the following theorem.

Theorem 2.1. *The joint pdf of* $(\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k, S_{\boldsymbol{y}}^2)$ *is given by*

$$f(\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k, S_{\boldsymbol{y}}^2) \propto \frac{1}{(\sigma^2)^{\frac{N-1}{2}}}\int_0^\infty \frac{(S_{\boldsymbol{y}}^2)^{\frac{N-k-2}{2}}(S_x^2)^{-(\frac{K+1}{2})}}{\left(\sigma^2 + \frac{S_x^2}{N-k}\right)^{\frac{1}{2}}} \times exp\left[-\frac{1}{2}\left\{\frac{S_x^2}{\sigma^2} + \frac{(N-k)S_{\boldsymbol{y}}^2}{S_x^2} + \frac{\sum_{i=1}^{k} n_i(\bar{y}_i - \mu_i)^2}{\left(\sigma^2 + \frac{S_x^2}{N-k}\right)}\right\}\right] dS_x^2$$

Proof. Starting from the fact that $\bar{x}_i \sim N(\mu_i, \frac{\sigma^2}{n_i})$ independently for each $i = 1, 2, \ldots, k$, and $\frac{S_x^2}{\sigma^2} \sim \chi_{N-k}^2$ independently of each $\bar{x}_i$, we can write the joint pdf of $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k, S_x^2)$ as given by,

$$f(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k, S_x^2) \propto \frac{(S_x^2)^{\frac{N-k}{2}-1} e^{-\frac{1}{2\sigma^2}\left[S_x^2 + \sum_{i=1}^{k} n_i(\bar{x}_i - \mu_i)^2\right]}}{(\sigma^2)^{\frac{N}{2}}}.$$

Conditionally given $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k, S_x^2)$,

$$\bar{y}_i \sim N\left(\bar{x}_i, \frac{\hat{\sigma}^2}{n_i}\right), \text{ independently for } i = 1, 2, \ldots, k,$$

$$S_y^2 \sim \hat{\sigma}^2 \chi^2_{N-k}, \text{ independently of each } \bar{y}_i.$$

Therefore the conditional pdf of $(\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k, S_{\boldsymbol{y}}^2 | \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k, S_x^2)$ is given by,

$$f(\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k, S_{\boldsymbol{y}}^2 | \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k, S_x^2) \propto \frac{(S_y^2)^{\frac{N-k}{2}-1} e^{-\frac{1}{2\hat{\sigma}^2}\left[S_y^2 + \sum_{i=1}^k n_i(\bar{y}_i - \bar{x}_i)^2\right]}}{(\hat{\sigma}^2)^{\frac{N}{2}}}.$$

The joint pdf of $(\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k, S_{\boldsymbol{y}}^2)$ can be expressed as,

$$\begin{aligned} f(\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k, S_{\boldsymbol{y}}^2) &= \int_{\bar{x}_1} \int_{\bar{x}_2} \cdots \int_{\bar{x}_k} \int_0^\infty f(\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k, S_{\boldsymbol{y}}^2 | \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k, S_x^2) \\ &\qquad \times f(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k, S_x^2) d\bar{x}_1 d\bar{x}_2 \ldots d\bar{x}_k dS_x^2 \\ &\propto \frac{1}{(\sigma^2)^{\frac{N}{2}}} \int_{\bar{x}_1} \int_{\bar{x}_2} \cdots \int_{\bar{x}_k} \int_0^\infty (S_y^2)^{\frac{N-k}{2}-1} (S_x^2)^{-(\frac{k+2}{2})} \\ &\qquad \times e^{-\frac{1}{2}\left[\frac{S_x^2 + \sum_{i=1}^k n_i(\bar{x}_i - \mu_i)^2}{\sigma^2} + \frac{S_y^2 + \sum_{i=1}^k n_i(\bar{y}_i - \bar{x}_i)^2}{\hat{\sigma}^2}\right]} d\bar{x}_1 d\bar{x}_2 \ldots d\bar{x}_k dS_x^2. \end{aligned}$$

We see that

$$\begin{aligned} &\frac{\sum_{i=1}^k n_i(\bar{y}_i - \bar{x}_i)^2}{\hat{\sigma}^2} + \frac{\sum_{i=1}^k n_i(\bar{x}_i - \mu_i)^2}{\sigma^2} \\ &= \frac{\sum_{i=1}^k n_i(\bar{y}_i - \mu_i)^2}{\sigma^2 + \hat{\sigma}^2} + \left(\frac{1}{\sigma^2} + \frac{1}{\hat{\sigma}^2}\right) \sum_{i=1}^k n_i \left[\bar{x}_i - \mu_i - \frac{\bar{y}_i - \mu_i}{\hat{\sigma}^2(\frac{1}{\sigma^2} + \frac{1}{\hat{\sigma}^2})}\right]^2 \end{aligned}$$

Hence after integrating out $\bar{x}_i$'s the joint pdf reduces to

$$f(\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k, S_{\boldsymbol{y}}^2) \propto \frac{1}{(\sigma^2)^{\frac{N}{2}}} \int_0^\infty \frac{(S_y^2)^{\frac{N-k}{2}-1} (S_x^2)^{-(\frac{k+2}{2})}}{\left(\frac{1}{\sigma^2} + \frac{1}{\hat{\sigma}^2}\right)^{\frac{1}{2}}} e^{-\frac{1}{2}\left[\frac{S_x^2}{\sigma^2} + \frac{S_y^2}{\hat{\sigma}^2} + \frac{\sum_{i=1}^k n_i(\bar{y}_i - \mu_i)^2}{\sigma^2 + \hat{\sigma}^2}\right]} dS_x^2.$$

Finally we put $\hat{\sigma}^2 = \frac{S_x^2}{N-k}$ in the above form and hence the result follows. [Theorem (2.1) is proved]

The inferential results are discussed in the following remarks.

Remark 2.1. Following $F_{\boldsymbol{x}}$, we define the test statistic $F_{\boldsymbol{y}} = \left(\frac{N-k}{k-1}\right) \frac{\text{BSS}(\boldsymbol{y})}{\text{WSS}(\boldsymbol{y})}$, based on the PIS synthetic data, and a level γ test based on the synthetic data $\boldsymbol{y}$ for the testing problem (1) is given by $F_{\boldsymbol{y}} > C_{N,k,\gamma}$, where $C_{N,k,\gamma}$ is such that $P[F_{\boldsymbol{y}} > C_{N,k,\gamma} | H_0] = \gamma$. We obtain $C_{N,k,\gamma}$ by the following steps:

1. We consider the conditional distribution of $F_{\boldsymbol{y}} | \boldsymbol{x}$ which follows a non-central F-distribution with degrees of freedom $(k-1, N-k)$ and the non centrality parameter being $\lambda_{\boldsymbol{x}} = \frac{\sum_{i=1}^k n_i(\bar{x}_i - \bar{x}_w)^2}{\hat{\sigma}^2}$ where $\bar{x}_w = \frac{1}{N}\sum_{i=1}^k n_i \bar{x}_i$. Again $\frac{\sum_{i=1}^k n_i(\bar{x}_i - \bar{x}_w)^2}{\sigma^2}$ follows a non-central chi square distribution with $k-1$ degrees of freedom and the non centrality parameter being $\lambda = \frac{\sum_{i=1}^k n_i(\mu_i - \bar{\mu}_w)^2}{\sigma^2}$ where $\bar{\mu}_w = \frac{1}{N}\sum_{i=1}^k n_i \mu_i$. Note that $\lambda = 0$ under H_0 and hence $\frac{\sum_{i=1}^k n_i(\bar{x}_i - \bar{x}_w)^2}{\sigma^2} \sim \chi^2_{k-1}$ under H_0. Also $\hat{\sigma}^2 = \frac{S_x^2}{N-k} \sim \frac{\sigma^2}{N-k}\chi^2_{N-k}$ which is independently distributed with $\frac{\sum_{i=1}^k n_i(\bar{x}_i - \bar{x}_w)^2}{\sigma^2}$, therefore $\lambda_{\boldsymbol{x}}$ follows a $(k-1)$ times non-central F-distribution with parameters $(k-1, N-k)$ and the non centrality parameter be λ and $\lambda_{\boldsymbol{x}} \sim (k-1)F_{k-1,N-k}$ under H_0.

2. Note that $\gamma = P[F_{\boldsymbol{y}} > C_{N,k,\gamma} | H_0]$, which can be written as,

$$\begin{aligned} \gamma &= P[F_{\boldsymbol{y}} > C_{N,k,\gamma} | H_0] \\ &= E_{H_0}\left[P(F_{\boldsymbol{y}} > C_{N,k,\gamma} | \boldsymbol{x})\right] \\ &= E_{H_0}\left[P(F_{k-1,N-k}(\lambda_{\boldsymbol{x}}) > C_{N,k,\gamma} | \boldsymbol{x})\right] \end{aligned}$$

$$= E_{H_0}\left[e^{-\frac{\lambda_{\boldsymbol{x}}}{2}}\sum_{j=0}^{\infty}\frac{(\frac{\lambda_{\boldsymbol{x}}}{2})^j}{j!}\frac{k+2j-1}{k-1}P(F_{k+2j-1,N-k}>C_{N,k,\gamma})\right]$$

where $F_{k-1,N-k}(\lambda_{\boldsymbol{x}})$ is a non-central F-variate with degrees of freedom $(k-1,N-k)$ with non centrality parameter $\lambda_{\boldsymbol{x}}$ and $F_{k+2j-1,N-k}$ be a central F-variate with degrees of freedom $(k+2j-1,N-k)$.

3. For a fixed C we can compute the expectation under H_0 by generating a large number of $\lambda_{\boldsymbol{x}}$'s as $\lambda_{\boldsymbol{x}} \sim (k-1)F_{k-1,N-k}$ under H_0, and compute the the quantity inside the expectation for each of those $\lambda_{\boldsymbol{x}}$'s and take the arithmetic mean to get the expectation.

4. Finally we numerically solve $E_{H_0}\left[e^{-\frac{\lambda_{\boldsymbol{x}}}{2}}\sum_{j=0}^{\infty}\frac{(\frac{\lambda_{\boldsymbol{x}}}{2})^j}{j!}\frac{k+2j-1}{k-1}P(F_{k+2j-1,N-k}>C)\right]-\gamma=0$ for C to get the cutoff $C_{N,k,\gamma}$.

Different cutoff values ($C_{N,k,\gamma}$'s) for different sets of sample sizes under fixed k and γ are given in section 4.

Remark 2.2. The power of the test proposed in remark (2.1) for a fixed alternative point $\boldsymbol{\mu}=(\mu_1,\mu_2,\mu_3,\mu_4,\mu_5)$ and $\sigma=1$ from H_1 is given by $\beta_{\text{PIS}}(\boldsymbol{\mu})=P[F_{\boldsymbol{y}}>C_{N,k,\gamma}|\boldsymbol{\mu}]$. $\beta_{\text{PIS}}(\boldsymbol{\mu})$ is calculated by Monte Carlo simulation technique. We generate a large number (10^6) of synthetic data sets and obtain the value of the test statistic for each of those data sets, then find the proportion of the values which are greater than $C_{N,k,\gamma}$. The powers for different choices of alternatives are given in section 4.

3 Posterior Predictive Sampling(PPS) Method

We assume a joint prior density of (μ_i,σ^2) as $\pi(\mu,\sigma^2)\propto(\sigma)^{-\alpha}$ for each $i=1,2,\ldots,k$, where $N+\alpha>7$. A synthetic data under this method can be obtained using the following steps.

1. First we draw $(\sigma^*)^2$ such that $\frac{S_x^2}{(\sigma^*)^2}\sim\chi^2_{N+\alpha-3}$.

2. Draw $\mu_i^*|(\sigma^*)^2\sim N(\bar{x}_i,\frac{(\sigma^*)^2}{n_i})$, independently for each $i=1,2,\ldots,k$.

3. Finally draw $\boldsymbol{z}_i=(z_{i1},z_{i2},\ldots,z_{in_i})$ where z_{ij}'s $(j=1,2,\ldots,n_i)$ are iid $N(\mu_i^*,(\sigma^*)^2)$, independently for each $i=1,2,\ldots,k$.

Here $\boldsymbol{z}=(\boldsymbol{z}_1,\boldsymbol{z}_2,\ldots,\boldsymbol{z}_k)$ constitutes the synthetic data obtained under PPS sampling method. Similar to the Plug-In sampling method which is discussed in section 2, here we define $\bar{z}_i=\frac{1}{n_i}\sum_{j=1}^{n_i}z_{ij}$, $\bar{z}_w=\frac{1}{N}\sum_{i=1}^{k}n_i\bar{z}_i$, Within Sum of Squares(WSS($\boldsymbol{z}$)) $=S_{\boldsymbol{z}}^2=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(z_{ij}-\bar{z}_i)^2$, and Between Sum of Squares(BSS($\boldsymbol{z}$)) $=\sum_{i=1}^{k}n_i(\bar{z}_i-\bar{z}_w)^2$. Likewise the case of PIS, here $\bar{z}_1,\bar{z}_2,\ldots,\bar{z}_k,S_{\boldsymbol{z}}^2$ are jointly sufficient for $(\mu_1,\mu_2,\ldots,\mu_k,\sigma^2)$ based on the synthetic data $\boldsymbol{z}=(\boldsymbol{z}_1,\boldsymbol{z}_2,\ldots,\boldsymbol{z}_k)$ obtained by the above method (see [14]).

Theorem 3.1. *The joint pdf of* $(\bar{z}_1,\bar{z}_2,\ldots,\bar{z}_k,S_{\boldsymbol{z}}^2)$ *is given by*

$$f(\bar{z}_1,\bar{z}_2,\ldots,\bar{z}_k,S_{\boldsymbol{z}}^2)\propto\frac{(S_{\boldsymbol{z}}^2)^{\frac{N-k}{2}-1}}{(\sigma^2)^{\frac{N}{2}}}\int_0^{\infty}\frac{\psi^{\frac{2N+\alpha-5}{2}}e^{-\frac{\psi}{2\sigma^2}\left[S_{\boldsymbol{z}}^2+\frac{1}{2+\psi}\sum_{i=1}^{k}n_i(\bar{z}_i-\mu_i)^2\right]}}{(1+\psi)^{\frac{2N+\alpha-k-3}{2}}(2+\psi)^{\frac{k}{2}}}d\psi$$

where $\psi=\frac{\sigma^2}{(\sigma^*)^2}$.

Proof. We start with the joint pdf of $(\bar{x}_1,\bar{x}_2,\ldots,\bar{x}_k,S_x^2)$ as given by,

$$f(\bar{x}_1,\bar{x}_2,\ldots,\bar{x}_k,S_x^2)\propto\frac{(S_x^2)^{\frac{N-k}{2}-1}e^{-\frac{1}{2\sigma^2}\left[S_x^2+\sum_{i=1}^{k}n_i(\bar{x}_i-\mu_i)^2\right]}}{(\sigma^2)^{\frac{N}{2}}}.$$

Next we note that μ_i^*'s $(i = 1, 2, \ldots, k)$ and $(\sigma^*)^2$ are generated as mentioned in Steps 1. and 2. of section 3, and therefore the joint pdf of $(\mu_1^*, \mu_2^*, \ldots, \mu_k^*, (\sigma^*)^2)$ conditionally for given $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k, S_x^2)$ is given by,

$$f(\mu_1^*, \mu_2^*, \ldots, \mu_k^*, (\sigma^*)^2 | \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k, S_x^2) \propto \frac{(S_x^2)^{\frac{N+\alpha-3}{2}} e^{-\frac{1}{2(\sigma^*)^2}\left[S_x^2 + \sum_{i=1}^k n_i(\mu_i^* - \bar{x}_i)^2\right]}}{\{(\sigma^*)^2\}^{\frac{N+\alpha+k-3}{2}+1}}.$$

Again conditionally given $(\mu_1^*, \mu_2^*, \ldots, \mu_k^*, (\sigma^*)^2)$,

$$\begin{aligned}
\bar{z}_i &\sim N\left(\mu_i^*, \frac{(\sigma^*)^2}{n_i}\right), \text{ independently for } i = 1, 2, \ldots, k, \\
S_z^2 &\sim (\sigma^*)^2 \chi_{N-k}^2, \text{ independently of each } \bar{z}_i.
\end{aligned}$$

Therefore the conditional pdf of $(\bar{z}_1, \bar{z}_2, \ldots, \bar{z}_k, S_{\boldsymbol{z}}^2 | \mu_1^*, \mu_2^*, \ldots, \mu_k^*, (\sigma^*)^2)$ is given by,

$$f(\bar{z}_1, \bar{z}_2, \ldots, \bar{z}_k, S_{\boldsymbol{z}}^2 | \mu_1^*, \mu_2^*, \ldots, \mu_k^*, (\sigma^*)^2) \propto \frac{1}{\{(\sigma^*)^2\}^{\frac{N}{2}}} (S_z^2)^{\frac{N-k}{2}-1} e^{-\frac{1}{2(\sigma^*)^2}\left[S_z^2 + \sum_{i=1}^k n_i(\bar{z}_i - \mu_i^*)^2\right]}$$

We can write the joint pdf of $(\bar{z}_1, \bar{z}_2, \ldots, \bar{z}_k, S_{\boldsymbol{z}}^2)$ as,

$$\begin{aligned}
&f(\bar{z}_1, \bar{z}_2, \ldots, \bar{z}_k, S_{\boldsymbol{z}}^2) \\
&\propto \int_{\mu_1^*} \int_{\mu_2^*} \cdots \int_{\mu_k^*} \int_{\bar{x}_1} \int_{\bar{x}_2} \cdots \int_{\bar{x}_k} \int_0^\infty \int_0^\infty f(\bar{z}_1, \bar{z}_2, \ldots, \bar{z}_k, S_{\boldsymbol{z}}^2 | \mu_1^*, \mu_2^*, \ldots, \mu_k^*, (\sigma^*)^2) \\
&\times f(\mu_1^*, \mu_2^*, \ldots, \mu_k^*, (\sigma^*)^2 | \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k, S_x^2) f(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k, S_x^2) \left(\prod_{i=1}^k d\mu_i^*\right) \left(\prod_{i=1}^k d\bar{x}_i\right) dS_z^2 d(\sigma^*)^2 \\
&\propto \int_{\mu_1^*} \int_{\mu_2^*} \cdots \int_{\mu_k^*} \int_{\bar{x}_1} \int_{\bar{x}_2} \cdots \int_{\bar{x}_k} \int_0^\infty \int_0^\infty \frac{(S_z^2)^{\frac{N-k}{2}-1} (S_x^2)^{\frac{2N+\alpha-k-5}{2}}}{(\sigma^2)^{\frac{N}{2}} \{(\sigma^*)^2\}^{\frac{2N+\alpha+k-1}{2}}} e^{-\frac{S_z^2}{2(\sigma^*)^2}} e^{-\left[\frac{1}{\sigma^2} + \frac{1}{(\sigma^*)^2}\right]\frac{S_x^2}{2}} \\
&\times e^{-\frac{1}{2\sigma^2}\sum_{i=1}^k n_i(\bar{x}_i - \mu_i)^2} e^{-\frac{1}{2(\sigma^*)^2}\sum_{i=1}^k n_i\left[(\bar{z}_i - \mu_i^*)^2 + (\mu_i^* - \bar{x}_i)^2\right]} \left(\prod_{i=1}^k d\mu_i^*\right) \left(\prod_{i=1}^k d\bar{x}_i\right) dS_z^2 d(\sigma^*)^2.
\end{aligned}$$

Note that $(\bar{z}_i - \mu_i^*)^2 + (\mu_i^* - \bar{x}_i)^2 = \frac{(\bar{z}_i - \bar{x}_i)^2}{2} + 2\left[\mu_i^* - \frac{\bar{z}_i + \bar{x}_i}{2}\right]^2$, integrating out $\mu_1^*, \mu_2^*, \ldots, \mu_k^*$ we get,

$$\begin{aligned}
&f(\bar{z}_1, \bar{z}_2, \ldots, \bar{z}_k, S_{\boldsymbol{z}}^2) \\
&\propto \int_{\bar{x}_1} \int_{\bar{x}_2} \cdots \int_{\bar{x}_k} \int_0^\infty \int_0^\infty \frac{(S_z^2)^{\frac{N-k}{2}-1} (S_x^2)^{\frac{2N+\alpha-k-5}{2}}}{(\sigma^2)^{\frac{N}{2}} \{(\sigma^*)^2\}^{\frac{2N+\alpha-1}{2}}} e^{-\frac{S_z^2}{2(\sigma^*)^2}} e^{-\left[\frac{1}{\sigma^2} + \frac{1}{(\sigma^*)^2}\right]\frac{S_x^2}{2}} \\
&\times e^{-\frac{1}{2}\sum_{i=1}^k n_i\left[\frac{(\bar{z}_i - \bar{x}_i)^2}{2(\sigma^*)^2} + \frac{(\bar{x}_i - \mu_i)^2}{\sigma^2}\right]} \left(\prod_{i=1}^k d\bar{x}_i\right) dS_z^2 d(\sigma^*)^2.
\end{aligned}$$

Again we see that,

$$\begin{aligned}
&\frac{(\bar{z}_i - \bar{x}_i)^2}{2(\sigma^*)^2} + \frac{(\bar{x}_i - \mu_i)^2}{\sigma^2} \\
&= \frac{(\bar{z}_i - \mu_i)^2}{\sigma^2 + 2(\sigma^*)^2} + \left(\frac{1}{\sigma^2} + \frac{1}{2(\sigma^*)^2}\right) \left[\bar{x}_i - \frac{\frac{\mu_i}{\sigma^2} + \frac{\bar{z}_i}{2(\sigma^*)^2}}{\frac{1}{\sigma^2} + \frac{1}{2(\sigma^*)^2}}\right]^2.
\end{aligned}$$

Integrating out $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$, and after doing some simplifications we get,

$$\begin{aligned}
&f(\bar{z}_1, \bar{z}_2, \ldots, \bar{z}_k, S_{\boldsymbol{z}}^2) \\
&\propto \int_0^\infty \int_0^\infty \frac{(S_z^2)^{\frac{N-k}{2}-1} (S_x^2)^{\frac{2N+\alpha-k-5}{2}}}{(\sigma^2)^{\frac{N-k}{2}} \{(\sigma^*)^2\}^{\frac{2N+\alpha-k-1}{2}} (\sigma^2 + 2(\sigma^*)^2)^{\frac{k}{2}}} e^{-\frac{S_z^2}{2(\sigma^*)^2}} e^{-\left[\frac{1}{\sigma^2} + \frac{1}{(\sigma^*)^2}\right]\frac{S_x^2}{2}}
\end{aligned}$$

$$\times e^{-\frac{1}{2}\sum_{i=1}^{k}\frac{n_i(\bar{z}_i-\mu_i)^2}{\sigma^2+2(\sigma^*)^2}} dS_z^2 d(\sigma^*)^2,$$

and thereafter integrating over S_x^2 we get,

$$f(\bar{z}_1,\bar{z}_2,\ldots,\bar{z}_k,S_{\boldsymbol{z}}^2) \propto \int_0^\infty \frac{(S_z^2)^{\frac{N-k}{2}-1} e^{-\frac{S_z^2}{2(\sigma^*)^2}} e^{-\frac{1}{2}\sum_{i=1}^{k}\frac{n_i(\bar{z}_i-\mu_i)^2}{\sigma^2+2(\sigma^*)^2}}}{(\sigma^2)^{-\frac{N+\alpha-3}{2}}(\sigma^*)^2(\sigma^2+(\sigma^*)^2)^{\frac{2N+\alpha-k-3}{2}}(\sigma^2+2(\sigma^*)^2)^{\frac{k}{2}}} d(\sigma^*)^2.$$

Finally making the transformation $(\sigma^*)^2 \to \psi = \frac{\sigma^2}{(\sigma^*)^2}$ and after doing some simplifications we get,

$$f(\bar{z}_1,\bar{z}_2,\ldots,\bar{z}_k,S_{\boldsymbol{z}}^2) \propto \frac{(S_z^2)^{\frac{N-k}{2}-1}}{(\sigma^2)^{\frac{N}{2}}} \int_0^\infty \frac{\psi^{\frac{2N+\alpha-5}{2}} e^{-\frac{\psi}{2\sigma^2}\left[S_z^2+\frac{1}{2+\psi}\sum_{i=1}^{k} n_i(\bar{z}_i-\mu_i)^2\right]}}{(1+\psi)^{\frac{2N+\alpha-k-3}{2}}(2+\psi)^{\frac{k}{2}}} d\psi.$$

[Theorem (3.1) is proved]

The marginal distribution of ψ is given in the following theorem.

Theorem 3.2. *The marginal distribution of ψ is such that $\left(\frac{N-k}{N+\alpha-3}\right)\psi \sim F_{N+\alpha-3,N-k}$.*

Proof. The joint pdf of $(\bar{z}_1,\bar{z}_2,\ldots,\bar{z}_k,S_{\boldsymbol{z}}^2,\psi)$ which can be found from Theorem(3.1) as,

$$f(\bar{z}_1,\bar{z}_2,\ldots,\bar{z}_k,S_{\boldsymbol{z}}^2,\psi) \propto \frac{(S_z^2)^{\frac{N-k}{2}-1}\psi^{\frac{2N+\alpha-5}{2}} e^{-\frac{\psi}{2\sigma^2}\left[S_z^2+\frac{1}{2+\psi}\sum_{i=1}^{k} n_i(\bar{z}_i-\mu_i)^2\right]}}{(\sigma^2)^{\frac{N}{2}}(1+\psi)^{\frac{2N+\alpha-k-3}{2}}(2+\psi)^{\frac{k}{2}}}.$$

Integrating over $\bar{z}_1,\bar{z}_2,\ldots,\bar{z}_k$ we get,

$$f(S_{\boldsymbol{z}}^2,\psi) \propto \frac{\psi^{\frac{2N+\alpha-k-5}{2}}(S_z^2)^{\frac{N-k}{2}-1} e^{-\frac{\psi S_z^2}{2\sigma^2}}}{(\sigma^2)^{\frac{N-k}{2}}(1+\psi)^{\frac{2N+\alpha-k-3}{2}}},$$

and hence integrating over S_z^2 we get the marginal pdf of ψ as,

$$\begin{aligned} f(\psi) &\propto \frac{\psi^{\frac{N+\alpha-3}{2}-1}}{(1+\psi)^{\frac{N+\alpha-3}{2}+\frac{N-k}{2}}}, \quad 0<\psi<\infty \\ &= \frac{1}{B\left(\frac{N+\alpha-3}{2},\frac{N-k}{2}\right)} \frac{\psi^{\frac{N+\alpha-3}{2}-1}}{(1+\psi)^{\frac{N+\alpha-3}{2}+\frac{N-k}{2}}}, \quad 0<\psi<\infty, \end{aligned}$$

where $B(a,b)$ is the beta function. Therefore $\left(\frac{N-k}{N+\alpha-3}\right)\psi \sim F_{N+\alpha-3,N-k}$. [Theorem (3.2) is proved]

The inferential results are discussed in the following remarks.

Remark 3.1. From theorem (3.1) it directly follows that conditionally given ψ,

i) $\bar{z}_i \sim N\left(\mu_i, \frac{(2+\psi)}{\psi}\frac{\sigma^2}{n_i}\right)$, independently for each $i=1,2,\ldots,k$,

ii) $\frac{\psi S_z^2}{\sigma^2} \sim \chi^2_{N-k}$, independently of all $\bar{z}_i$'s $(i=1,2,\ldots,k)$.

Defining $\sigma_i^2 = \frac{(2+\psi)}{\psi}\frac{\sigma^2}{n_i}$ and $\bar{\mu}_w = \frac{1}{N}\sum_{i=1}^{k} n_i\mu_i$ we get, for conditionally given ψ, $\sum_{i=1}^{k}\frac{(\bar{z}_i-\bar{z}_w)^2}{\sigma_i^2} = \frac{\mathrm{BSS}(\boldsymbol{z})}{\sigma^2}\left(\frac{\psi}{2+\psi}\right)$ will follow a non-central chi-square distribution with degrees of freedom $k-1$ and the non-centrality parameter be $\lambda = \sum_{i=1}^{k}\frac{(\mu_i-\bar{\mu}_w)^2}{\sigma_i^2}$ and $\frac{\psi}{\sigma^2}\mathrm{WSS}(\boldsymbol{z}) \sim \chi^2_{N-k}$ independently of $\mathrm{BSS}(\boldsymbol{z})$. To test the ANOVA problem given in (1) based on the synthetic data $\boldsymbol{z} = (\boldsymbol{z}_1,\boldsymbol{z}_2,\ldots,\boldsymbol{z}_k)$ as in the previous case, we define our test statistic under this method as $F_{\boldsymbol{z}} = \left(\frac{N-k}{k-1}\right)\frac{\mathrm{BSS}(\boldsymbol{z})}{\mathrm{WSS}(\boldsymbol{z})}$ which, conditionally given ψ, follows $(2+\psi)$ times a non-central F-distribution with degrees of freedom $(k-1,N-k)$ and with the non-centrality parameter λ. Note that $\lambda=0$ under H_0 and hence $F_{\boldsymbol{z}}|\psi \sim (2+\psi)F_{k-1,N-k}$ under H_0. A level γ test based on the synthetic data $\boldsymbol{z}$ for the testing problem (1) is given by $F_{\boldsymbol{z}} > D_{N,k,\alpha,\gamma}$, where $D_{N,k,\alpha,\gamma}$ is such that $P[F_{\boldsymbol{z}} > D_{N,k,\alpha,\gamma}|H_0] = \gamma$. We obtain $D_{N,k,\alpha,\gamma}$ by the following steps,

1. We can write,

$$\begin{aligned}\gamma &= P[F_{\boldsymbol{z}} > D_{N,k,\alpha,\gamma}|H_0] \\ &= E_{H_0}\left[P\left(F_{\boldsymbol{z}} > D_{N,k,\alpha,\gamma}|\psi\right)\right] \\ &= E_{H_0}\left[P\left(F_{k-1,N-k} > \frac{D_{N,k,\alpha,\gamma}}{2+\psi}|\psi\right)\right]\end{aligned}$$

2. For a fixed D, to compute the expectation $E_{H_0}\left[P\left(F_{k-1,N-k} > \frac{D}{2+\psi}|\psi\right)\right]$, we generate a large number of ψ's such that $\left(\frac{N-k}{N+\alpha-3}\right)\psi \sim F_{N+\alpha-3,N-k}$, then compute $P\left(F_{k-1,N-k} > \frac{D}{2+\psi}\right)$ for each of those ψ's and take their simple arithmetic mean.

3. Finally we numerically solve $E_{H_0}\left[P\left(F_{k-1,N-k} > \frac{D}{2+\psi}|\psi\right)\right] - \gamma = 0$ for D to obtain $D_{N,k,\alpha,\gamma}$.

Different $D_{N,k,\alpha,\gamma}$'s are obtained for different sets of sample sizes which are provided in section 4.

Remark 3.2. The power of the test based on $F_{\boldsymbol{z}}$, at a particular alternative point $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3, \mu_4, \mu_5)$ and $\sigma = 1$ is given by $\beta_{\text{PPS}}(\boldsymbol{\mu}) = P[F_{\boldsymbol{z}} > D_{N,k,\alpha,\gamma}|\boldsymbol{\mu}]$. Similar to the PIS case, here also we use Monte Carlo simulation to compute the powers. We generate a very large number (10^6) of synthetic data sets by PPS method under a fixed choice of alternative $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3, \mu_4, \mu_5)$ and obtain the value of $F_{\boldsymbol{z}}$ for each of these data sets. The estimated power will be the proportion of the cases that the value of $F_{\boldsymbol{z}}$ exceeds the cutoff $D_{N,k,\alpha,\gamma}$. The powers for different sets of alternatives are provided in section 4.

4 Simulation Studies

In this section we provide a simulation study for comparison of power of the two F-tests proposed in Sections 2 and 3. We take $k = 5$ and four choices of set of sample sizes as $n_i = 10, i = 1(1)5$, $n_i = 15, i = 1(1)5$, $n_i = 20, i = 1(1)5$, and $n_1 = 10, n_2 = 10, n_3 = 15, n_4 = 20, n_5 = 25$. In both PIS and PPS methods we take $\gamma = 0.05$ and in PPS method $\alpha = 4$. Five choices of set of alternatives $(\mu_1, \mu_2, \mu_3, \mu_4, \mu_5)$ are taken as $(0, 0, 0, 0, 0.5)$, $(0, 0, 0, -0.5, 0.5)$, $(0, -0.5, -0.5, 0.5, 0.5)$, $(0, -1, -0.5, 0.5, 1)$, and $(0, -1, -1, 1, 1)$. We use Monte Carlo simulation technique with $S = 100000$ iterations to compute the powers. Following are the tables showing the cut-off points of the test statistics for different sets of sample sizes for both PIS and PPS methods and also the tables showing the powers of the tests for different choices of alternatives. Clearly the powers under PIS method are higher than those under PPS method.

Table 1: CUTOFF POINTS FOR DIFFERENT CHOICES OF SAMPLE SIZES UNDER PIS ($k = 5, \gamma = 0.05, S = 100000$)

.

$(n_1, n_2, n_3, n_4, n_5)$	N	$C_{N,k,\gamma}$
(10,10,10,10,10)	50	5.33159
(15,15,15,15,15)	75	5.12243
(20,20,20,20,20)	100	5.02934
(10,10,15,20,25)	80	5.08072

Table 2: CUTOFF POINTS FOR DIFFERENT CHOICES OF SAMPLE SIZES UNDER PPS ($k = 5, \gamma = 0.05, \alpha = 4, S = 100000$)

.

$(n_1, n_2, n_3, n_4, n_5)$	N	$D_{N,k,\alpha,\gamma}$
(10,10,10,10,10)	50	8.20283
(15,15,15,15,15)	75	7.78576
(20,20,20,20,20)	100	7.59969
(10,10,15,20,25)	80	7.77348

Table 3: TABLE SHOWING THE POWER FOR DIFFERENT CHOICES OF ALTERNATIVES AND DIFFERENT CHOICES OF SAMPLE SIZES UNDER PIS ($k = 5, \gamma = 0.05, S = 100000$)

.

$(\mu_1, \mu_2, \mu_3, \mu_4, \mu_5)$	$(n_1, n_2, n_3, n_4, n_5)$			
	(10,10,10,10,10)	(15,15,15,15,15)	(20,20,20,20,20)	(10,10,15,20,25)
(0,0,0,0,0.5)	0.09913	0.13102	0.16348	0.17448
(0,0,0,-0.5,0.5)	0.18898	0.27959	0.37456	0.41313
(0,-0.5,-0.5,0.5,0.5)	0.35557	0.53521	0.68064	0.57680
(0,-1,-0.5,0.5,1)	0.75660	0.92609	0.98118	0.94345
(0,-1,-1,1,1)	0.92926	0.99272	0.99930	0.99609

Table 4: TABLE SHOWING THE POWER FOR DIFFERENT CHOICES OF ALTERNATIVES AND DIFFERENT CHOICES OF SAMPLE SIZES UNDER PPS ($k = 5, \gamma = 0.05, \alpha = 4, S = 100000$)

.

$(\mu_1, \mu_2, \mu_3, \mu_4, \mu_5)$	$(n_1, n_2, n_3, n_4, n_5)$			
	(10,10,10,10,10)	(15,15,15,15,15)	(20,20,20,20,20)	(10,10,15,20,25)
(0,0,0,0,0.5)	0.08755	0.10629	0.12831	0.13636
(0,0,0,-0.5,0.5)	0.15121	0.21016	0.26993	0.29953
(0,-0.5,-0.5,0.5,0.5)	0.26855	0.39275	0.51015	0.42202
(0,-1,-0.5,0.5,1)	0.60137	0.79649	0.90669	0.82283
(0,-1,-1,1,1)	0.80836	0.94558	0.98722	0.96096

5 Disclosure Risk Evaluation

When the original (unit level) microdata is considered to be sensitive and thus hidden through the use of a masked version, it is natural to examine the extent to which sensitivity of a data point has been protected. A slight variation of a popular measure to study the disclosure risk of a single value x_i, given by Klein and Sinha ([16]), is taken as

$$P[|\hat{x}_i - x_i| < \epsilon | \mathbf{X}] = \theta_i$$

where $\mathbf{X}$ is the entire original data, and $\hat{x}_i$ is an intruder's prediction of x_i based upon seeing the released (artificial/synthetic) data, ϵ be any small positive quantity. Naturally, a high value of the above probability indicates a low level of protection and vice versa. This measure is computed based on the random mechanism producing the masked data, given the original data $\mathbf{X}$.

Returning to our specific problem, the j^{th} observation from the i^{th} experiment, namely, x_{ij}, has been perturbed and replaced by y_{ij} under PIS and z_{ij} under PPS. We consider two cases: Case (1) - the label (unit) which produced the j^{th} item is identifiable and Case (2) - identity is lost. In Case (1), intruder's best guess about x_{ij} can be taken as y_{ij} (PIS) or z_{ij} (PPS). In Case (2), on the other hand, the intruder being unable to identify the j^{th} unit in i^{th} experiment, makes a guess $\bar{y}_i$ (PIS) or $\bar{z}_i$ (PPS) for the missing x_{ij} value.

The following theorem gives upper bounds to the disclosure risk both under PIS and PPS. Although it is quite possible to exactly compute the required disclosure risk probabilities for any unit (identifiable or not), the usefulness of the upper bounds lies in the merit that they provide the best case scenarios and also these bounds are independent of any specific responder, thus providing a uniform comparison under PIS and PPS.

Theorem 5.1. *Suppose θ_{ij} be the disclosure risk for the j^{th} unit in the i^{th} experiment then,*
Case (1): (Units are identifiable)

(a) $\theta ij \leq 2\Phi(\frac{\epsilon}{s_x}) - 1$ *(PIS)*

(b) $\theta_{ij} \leq 2G_\nu\left(\frac{\epsilon}{s_x\sqrt{1+\frac{1}{n_i}}}\right) - 1$ *(PPS)*

Case (2): (Units are unidentifiable)

(a) $\theta ij \leq 2\Phi\left(\frac{\sqrt{n_i}\epsilon}{s_x}\right) - 1$ *(PIS)*

(b) $\theta_{ij} \leq 2G_\nu\left(\frac{\epsilon}{s_x\sqrt{\frac{2}{n_i}}}\right) - 1$ *(PPS)*

where $\phi(.)$ is the pdf of a $N(0,1)$ distribution and $g_\nu(.)$ is the pdf of a t-distribution with $\nu = N + \alpha - 3$ degrees of freedom and $s_x^2 = \hat{\sigma}^2 = \frac{WSS}{(N-k)} = \frac{S_x^2}{N-k}$.

Proof. Case (1): Here we assume that all the units of each experiments are identifiable. The disclosure risk for the j^{th} unit corresponding to the i^{th} experiment is given by

$$\theta_{ij} = P[|\hat{x}_{ij} - x_{ij}| < \epsilon|\mathbf{X}].$$

(a) Under PIS method, intruder's best guess about x_{ij} will be y_{ij}, that is $\hat{x}_{ij} = y_{ij}, i = 1, 2, \ldots, k; j = 1, 2, \ldots, n_i$. Now $y_{ij}|\mathbf{X} \sim N(\bar{x}_i, s_x^2)$ for each $j = 1, 2, \ldots, n_i$ and $i = 1, 2, \ldots, k$, therefore we can write,

$$\begin{aligned}
\theta_{ij} &= P[|\hat{x}_{ij} - x_{ij}| < \epsilon|\mathbf{X}] \\
&= P[|y_{ij} - x_{ij}| < \epsilon|\mathbf{X}] \\
&= P[-\epsilon < y_{ij} - x_{ij} < \epsilon|\mathbf{X}] \\
&= P\left[\frac{-\epsilon - (\bar{x}_i - x_{ij})}{s_x} < Z < \frac{\epsilon - (\bar{x}_i - x_{ij})}{s_x}\right] \\
&\text{(where } Z \text{ is a } N(0,1) \text{ variate.)} \\
&= P[-\eta + \delta_{ij} < Z < \eta + \delta_{ij}] \\
&\text{(writing } \eta = \tfrac{\epsilon}{s_x} \text{ and } \delta_{ij} = \tfrac{x_{ij} - \bar{x}_i}{s_x}) \\
&= \Phi(\eta + \delta) - \Phi(-\eta + \delta) \\
&(\Phi \text{ be the CDF of } N(0,1) \text{ distribution.}) \\
&\leq \Phi(\eta) - \Phi(-\eta) \\
&= 2\Phi(\eta) - 1.
\end{aligned}$$

(b) Under PPS method, intruder's best guess about x_{ij} will be z_{ij}, that is $\hat{x}_{ij} = z_{ij}, i = 1, 2, \ldots, k; j = 1, 2, \ldots, n_i$. Now $z_{ij}|\mu_i^*, (\sigma^*)^2 \sim N(\mu_i^*, (\sigma^*)^2)$ for each $j = 1, 2, \ldots, n_i$ and $i = 1, 2, \ldots, k$, where $\mu_i^*|(\sigma^*)^2 \sim N(\bar{x}_i, \frac{(\sigma^*)^2}{n_i})$, independently for each $i = 1, 2, \ldots, k$ and $\frac{S_x^2}{(\sigma^*)^2} \sim \chi_\nu^2$ $(\nu = N + \alpha - 3)$.

In order to find the disclosure risk $\theta_{ij} = P[|z_{ij} - x_{ij}| < \epsilon|\mathbf{X}]$, we need to find the marginal distribution of $z_{ij}|\mathbf{X}$. The joint pdf of $(z_{ij}, \mu_i^*, (\sigma^*)^2)$ has the pdf of the form,

$$\begin{aligned}
f(z_{ij}, \mu_i^*, (\sigma^*)^2) &= f(z_{ij}|\mu_i^*, (\sigma^*)^2) \times f(\mu_i^*|(\sigma^*)^2) \times f((\sigma^*)^2) \\
&\propto \frac{(S_x^2)^{\frac{\nu}{2}}}{[(\sigma^*)^2]^{\frac{\nu}{2}+2}} e^{-\frac{S_x^2}{2(\sigma^*)^2}} e^{-\frac{1}{2(\sigma^*)^2}\left[(z_{ij} - \mu_i^*)^2 + n_i(\mu_i^* - \bar{x}_i)^2\right]}.
\end{aligned}$$

Note that,

$$(z_{ij} - \mu_i^*)^2 + n_i(\mu_i^* - \bar{x}_i)^2 = (n_i + 1)\left[\mu_i^* - \frac{z_{ij} + n_i\bar{x}_i}{n_i + 1}\right]^2 + \frac{n_i(z_{ij} - \bar{x}_i)^2}{n_i + 1}.$$

Therefore integrating the joint pdf over μ_i^* and then over $(\sigma^*)^2$ we get the marginal pdf of z_{ij} given the data as,

$$f(z_{ij}|\mathbf{X}) \propto \int_0^\infty \frac{(S_x^2)^{\frac{\nu}{2}} e^{-\frac{S_x^2}{2(\sigma^*)^2}}}{[(\sigma^*)^2]^{\frac{\nu}{2}+1}} \times \frac{e^{-\frac{n_i}{2(n_i+1)(\sigma^*)^2}(z_{ij}-\bar{x}_i)^2}}{\sigma^*} d(\sigma^*)^2.$$

From the above expression of the marginal pdf we can write,

1. $z_{ij}|(\sigma^*)^2, \mathbf{X} \sim N\left(\bar{x}_i, \frac{(n_i+1)(\sigma^*)^2}{n_i}\right) \Rightarrow \frac{z_{ij}-\bar{x}_i}{\sigma^*\sqrt{1+\frac{1}{n_i}}}|(\sigma^*)^2, \mathbf{X} \sim N(0,1)$,

2. $\frac{S_x^2}{(\sigma^*)^2}|\mathbf{X} \sim \chi_\nu^2$.

Defining $s_x^2 = \frac{S_x^2}{\nu}$, we can write

$$\frac{z_{ij}-\bar{x}_i}{s_x\sqrt{1+\frac{1}{n_i}}}|\mathbf{X} \sim t_\nu.$$

Finally the disclosure risk can be written as,

$$\begin{aligned}
\theta_{ij} &= P[|z_{ij}-x_{ij}| < \epsilon|\mathbf{X}] \\
&= P[-\epsilon + x_{ij} < z_{ij} < \epsilon + x_{ij}|\mathbf{X}] \\
&= P\left[\frac{-\epsilon + x_{ij} - \bar{x}_i}{s_x\sqrt{1+\frac{1}{n_i}}} < t_\nu < \frac{\epsilon + x_{ij} - \bar{x}_i}{s_x\sqrt{1+\frac{1}{n_i}}}\right] \\
&\text{where } t_\nu \text{ follows a } t\text{-distribution with } \nu \text{ degrees of freedom.} \\
&= G_\nu(\eta_i + \zeta_{ij}) - G_\nu(-\eta_i + \zeta_{ij}) \\
&(\text{writing } \eta_i = \tfrac{\epsilon}{s_x\sqrt{1+\frac{1}{n_i}}} \text{ and } \zeta_{ij} = \tfrac{x_{ij}-\bar{x}_i}{s_x\sqrt{1+\frac{1}{n_i}}}) \\
&\leq G_\nu(\eta_i) - G_\nu(-\eta_i) \\
&[G_\nu(.) \text{ is the CDF of a } t\text{-distribution with } \nu \text{ degrees of freedom}] \\
&= 2G_\nu(\eta_i) - 1.
\end{aligned}$$

Case (2): Here the identities of j^{th} unit $(j = 1, 2, \ldots, n_i)$ corresponding to the i^{th} experiment $(i = 1, 2, \ldots, k)$ are lost, hence the intruder's best guess about x_{ij} will be taken as $\bar{y}_i$ for PIS and $\bar{z}_i$ for PPS. The disclosure risk for the j^{th} unit corresponding to the i^{th} experiment is given by

$$\begin{aligned}
\theta_{ij} &= P[|\hat{x}_{ij} - x_{ij}| < \epsilon|\mathbf{X}] \\
&= \begin{cases} P[|\bar{y}_i - x_{ij}| < \epsilon|\mathbf{X}] & \text{(PIS)} \\ P[|\bar{z}_i - x_{ij}| < \epsilon|\mathbf{X}] & \text{(PPS).} \end{cases}
\end{aligned}$$

(a) Note that, $\bar{y}_i|\mathbf{X} \sim N(\bar{x}_i, \frac{s_x^2}{n_i})$ for $i = 1, 2, \ldots, k$, and hence the disclosure risk under PIS is given by

$$\begin{aligned}
\theta_{ij} &= P[|\bar{y}_i - x_{ij}| < \epsilon|\mathbf{X}] \\
&= P[-\epsilon < \bar{y}_i - x_{ij} < \epsilon|\mathbf{X}] \\
&= P\left[\frac{-\epsilon + (x_{ij} - \bar{x}_i)}{\frac{s_x}{\sqrt{n_i}}} < Z < \frac{-\epsilon + (x_{ij} - \bar{x}_i)}{\frac{s_x}{\sqrt{n_i}}}\right] \\
&(\text{where } Z \text{ is a } N(0,1) \text{ variate.}) \\
&= P[-\eta_i + \delta_{ij} < Z < \eta_i + \delta_{ij}] \\
&(\text{writing } \eta_i = \tfrac{\sqrt{n_i}\epsilon}{s_x} \text{ and } \delta_{ij} = \tfrac{\sqrt{n_i}(x_{ij}-\bar{x}_i)}{s_x}) \\
&= \Phi(\eta_i + \delta_{ij}) - \Phi(-\eta_i + \delta_{ij}) \\
&(\Phi \text{ be the CDF of } N(0,1) \text{ distribution.})
\end{aligned}$$

$$\leq \Phi(\eta_i) - \Phi(-\eta_i)$$
$$= 2\Phi(\eta_i) - 1.$$

(b) To derive the disclosure risk under PPS, we need to find the marginal distribution of $\bar{z}_i|\mathbf{X}$. Now $\bar{z}_i|\mu_i^*, (\sigma^*)^2 \sim N(\mu_i^*, \frac{(\sigma^*)^2}{n_i})$ for $i = 1, 2, \ldots, k$, where $\mu_i^*|(\sigma^*)^2 \sim N(\bar{x}_i, \frac{(\sigma^*)^2}{n_i})$, independently for each $i = 1, 2, \ldots, k$ and $\frac{S_x^2}{(\sigma^*)^2} \sim \chi^2_\nu$ ($\nu = N + \alpha - 3$). The joint pdf of $(\bar{z}_i, \mu_i^*, (\sigma^*)^2)$ has the pdf of the form,

$$f(\bar{z}_i, \mu_i^*, (\sigma^*)^2) = f(\bar{z}_i|\mu_i^*, (\sigma^*)^2) \times f(\mu_i^*|(\sigma^*)^2) \times f((\sigma^*)^2)$$
$$\propto \frac{(S_x^2)^{\frac{\nu}{2}}}{[(\sigma^*)^2]^{\frac{\nu}{2}+2}} e^{-\frac{S_x^2}{2(\sigma^*)^2}} e^{-\frac{n_i}{2(\sigma^*)^2}\left[(\bar{z}_i - \mu_i^*)^2 + (\mu_i^* - \bar{x}_i)^2\right]}.$$

Note that,

$$(z_{ij} - \mu_i^*)^2 + (\mu_i^* - \bar{x}_i)^2 = 2\left[\mu_i^* - \frac{\bar{z}_i + \bar{x}_i}{2}\right]^2 + \frac{(\bar{z}_i - \bar{x}_i)^2}{2}.$$

Integrating the joint pdf over μ_i^* and then over $(\sigma^*)^2$ we get the marginal pdf of $\bar{z}_i$ given the data as,

$$f(\bar{z}_i|\mathbf{X}) \propto \int_0^\infty \frac{(S_x^2)^{\frac{\nu}{2}} e^{-\frac{S_x^2}{2(\sigma^*)^2}}}{[(\sigma^*)^2]^{\frac{\nu}{2}+1}} \times \frac{e^{-\frac{n_i}{4(\sigma^*)^2}(\bar{z}_i - \bar{x}_i)^2}}{\sigma^*} d(\sigma^*)^2.$$

From the above expression of the marginal pdf we can write,

1. $\bar{z}_i|(\sigma^*)^2, \mathbf{X} \sim N\left(\bar{x}_i, \frac{2(\sigma^*)^2}{n_i}\right) \Rightarrow \frac{\sqrt{n_i}(\bar{z}_i - \bar{x}_i)}{\sigma^*\sqrt{2}}|(\sigma^*)^2, \mathbf{X} \sim N(0, 1)$,
2. $\frac{S_x^2}{(\sigma^*)^2}|\mathbf{X} \sim \chi^2_\nu$.

Defining $s_x^2 = \frac{S_x^2}{\nu}$, we can write

$$\frac{\sqrt{n_i}(\bar{z}_i - \bar{x}_i)}{s_x\sqrt{2}}|\mathbf{X} \sim t_\nu.$$

Therefore the disclosure risk under PPS can be written as,

$$\theta_{ij} = P[|\bar{z}_i - x_{ij}| < \epsilon|\mathbf{X}]$$
$$= P[-\epsilon + x_{ij} < \bar{z}_i < \epsilon + x_{ij}|\mathbf{X}]$$
$$= P\left[\frac{-\epsilon + x_{ij} - \bar{x}_i}{s_x\sqrt{\frac{2}{n_i}}} < t_\nu < \frac{\epsilon + x_{ij} - \bar{x}_i}{s_x\sqrt{\frac{2}{n_i}}}\right]$$
where t_ν follows a t-distribution with ν degrees of freedom.
$$= G_\nu(\eta_i + \zeta_{ij}) - G_\nu(-\eta_i + \zeta_{ij})$$
(writing $\eta_i = \frac{\epsilon}{s_x\sqrt{\frac{2}{n_i}}}$ and $\zeta_{ij} = \frac{x_{ij} - \bar{x}_i}{s_x\sqrt{\frac{2}{n_i}}}$)
$$\leq G_\nu(\eta_i) - G_\nu(-\eta_i).$$
[$G_\nu(.)$ is the CDF of a t-distribution with ν degrees of freedom]
$$= 2G_\nu(\eta_i) - 1.$$

[Theorem (5.1) is proved]

Next we compute the upper bounds to the disclosure risks under PIS and PPS methods for both case (1) and (2) by taking suitable choices of ϵ, s_x and different sample sizes for different experiments. Here we have taken $\epsilon = 0.1$, $s_x = 5, 10, 15, 20$ and the sample sizes for three independent experiments as $n_1 = 10, n_2 = 15$ and $n_3 = 20$. The tables are given below. We observe that the disclosure risk under PIS is more than that under PPS in all the cases.

Table 5: TABLE SHOWING THE UPPER BOUND TO THE DISCLOSURE RISKS UNDER CASE (1) [ALL UNITS ARE IDENTIFIABLE] ($\epsilon = 0.1, \alpha = 4, k = 3$)

.

EXPERIMENTS ($k = 3$)	s_x							
	5		10		15		20	
	PIS	PPS	PIS	PPS	PIS	PPS	PIS	PPS
EXPERIMENT – 1 ($n_1 = 10$)	0.01595	0.01513	0.00798	0.00756	0.00532	0.00504	0.00399	0.00378
EXPERIMENT – 2 ($n_2 = 15$)	0.01595	0.01536	0.00798	0.00768	0.00532	0.00512	0.00399	0.00384
EXPERIMENT – 3 ($n_3 = 20$)	0.01595	0.01549	0.00798	0.00774	0.00532	0.00516	0.00399	0.00387

Table 6: TABLE SHOWING THE UPPER BOUND TO THE DISCLOSURE RISKS UNDER CASE (2) [UNITS ARE UNIDENTIFIABLE] ($\epsilon = 0.1, \alpha = 4, k = 3$)

.

EXPERIMENTS ($k = 3$)	s_x							
	5		10		15		20	
	PIS	PPS	PIS	PPS	PIS	PPS	PIS	PPS
EXPERIMENT – 1 ($n_1 = 10$)	0.05043	0.03548	0.02523	0.01774	0.01682	0.01183	0.01261	0.00887
EXPERIMENT – 2 ($n_2 = 15$)	0.06174	0.04344	0.03089	0.02173	0.02060	0.01449	0.01545	0.01087
EXPERIMENT – 3 ($n_3 = 20$)	0.07127	0.05015	0.03567	0.02509	0.02378	0.01673	0.01784	0.01255

6 Robustness Study

In this section we address the dual issues of 1) what happens to the standard ANOVA F-test based on original normal data when normality is violated?, and 2) what happens to the (modified) F-test based on synthetic data under violations of the assumption of normal samples?

We consider a few scenarios - samples are drawn from i) Double Exponential (DE) distribution with location 0 and scale 1, ii) Student's t distribution with degrees of freedom $\nu(\nu = 1, 5)$, iii) Exponential distribution with mean 1, and iv) Lognormal distribution with mean of $\log(X) = 0$ and variance of $\log(X) = 1$. Choice of the parameters of the above distributions is without any loss of generality due to the nature of the underlying null hypothesis.

Under 1), we compute the Type I error of the standard F-test with its natural cut-off point based on the F distribution under the above types of deviations from normality and check to what extent it changes from original stipulated size. We have taken the case of equal sample size $n = 15, 20, 25$ and $k = 5, 10, 15$. Results appear in Tables (7-11). Our finding is that the Type 1 errors are significantly affected for the extreme t_1 distribution, moderately affected for log-normal and exponential distributions, and there is hardly any change for DE and t_5 distributions. Some of these findings were also recorded in [23], [24], [25].

Under 2), we again consider the same four scenarios as under 1) with the same choice of n and k. Once samples have been drawn, we pretend that the data is from normal and proceed to generate synthetic data under PIS/PPS and carry out the data analysis based on the F-statistic and its (modified) F cut-off points. Simulated Type I errors are then computed and used to check to what extent it changes from original level 0.05. Results appear in Tables (7-11). Our finding is that, strangely enough, our proposed tests under both PIS/PPS show great robustness under deviations from normality for the two symmetric distributions (DE and t_5), while for Exponential, Log normal and t_1, simulated Type I error rate is also

very close to the nominal level.
All the calculations have been done using Monte Carlo simulation technique with $S = 10^6$ iterations.

Table 7: TABLE SHOWING THE TYPE – 1 ERRORS UNDER DIFFERENT CHOICES OF k AND EQUAL SAMPLE SIZES n UNDER *DOUBLE EXPONENTIAL*(0,1). $(\gamma = 0.05, \alpha = 4)$

.

k	n	*Type-I Error*		
		ORIGINAL	PIS	PPS
5	15	0.04755	0.04887	0.04948
	20	0.04873	0.04956	0.05168
	25	0.04932	0.04983	0.05005
10	15	0.04839	0.05008	0.04903
	20	0.04945	0.05014	0.04963
	25	0.04984	0.05077	0.04925
15	15	0.04816	0.04891	0.04994
	20	0.04951	0.05051	0.04876
	25	0.04944	0.04990	0.05086

Table 8: TABLE SHOWING TYPE – 1 ERRORS UNDER DIFFERENT CHOICES OF k AND EQUAL SAMPLE SIZES n UNDER t - DISTRIBUTION $(df = 1)$. $(\gamma = 0.05, \alpha = 4)$

.

k	n	*Type-I Error*		
		ORIGINAL	PIS	PPS
5	15	0.01609	0.04069	0.04607
	20	0.01555	0.03904	0.04692
	25	0.01600	0.04190	0.04557
10	15	0.01679	0.03996	0.04455
	20	0.01752	0.04049	0.04527
	25	0.01701	0.04123	0.04289
15	15	0.01900	0.03984	0.04567
	20	0.01886	0.03960	0.04408
	25	0.01849	0.04090	0.04590

Table 9: TABLE SHOWING TYPE – 1 ERRORS UNDER DIFFERENT CHOICES OF k AND EQUAL SAMPLE SIZES n UNDER t - DISTRIBUTION $(df = 5)$. $(\gamma = 0.05, \alpha = 4)$

.

k	n	*Type-I Error*		
		ORIGINAL	PIS	PPS
5	15	0.04822	0.04982	0.05063
	20	0.04851	0.04974	0.05218
	25	0.04862	0.05143	0.04872
10	15	0.05054	0.05112	0.05020
	20	0.04921	0.05010	0.05000
	25	0.05005	0.05048	0.04838
15	15	0.04992	0.04955	0.05024
	20	0.04968	0.04988	0.04972
	25	0.05054	0.05079	0.05199

Table 10: TABLE SHOWING TYPE – 1 ERRORS UNDER DIFFERENT CHOICES OF k AND EQUAL SAMPLE SIZES n UNDER *EXPONENTIAL* (Mean=1). ($\gamma = 0.05, \alpha = 4$)

.

k	n	*Type-I Error*		
		ORIGINAL	PIS	PPS
5	15	0.04528	0.04954	0.05070
	20	0.04631	0.04745	0.05102
	25	0.04727	0.04845	0.04849
10	15	0.04822	0.04975	0.04992
	20	0.04837	0.05003	0.05048
	25	0.04871	0.05130	0.04775
15	15	0.04923	0.04965	0.05132
	20	0.04950	0.04944	0.04809
	25	0.04910	0.05015	0.05143

Table 11: TABLE SHOWING TYPE – 1 ERRORS UNDER DIFFERENT CHOICES OF k AND EQUAL SAMPLE SIZES n UNDER *LOG NORMAL*(0,1). ($\gamma = 0.05, \alpha = 4$)

.

k	n	*Type-I Error*		
		ORIGINAL	PIS	PPS
5	15	0.03729	0.04628	0.04878
	20	0.03777	0.04587	0.04857
	25	0.04018	0.04768	0.04782
10	15	0.04146	0.04731	0.04906
	20	0.04213	0.04796	0.04902
	25	0.04314	0.04914	0.04730
15	15	0.04463	0.04821	0.05002
	20	0.04419	0.04889	0.04864
	25	0.04393	0.04883	0.04955

7 Concluding Remarks

From the tables given in section 5 we can see that, larger the value of WSS [WSS $= (N-k)s_x^2$], lower the disclosure risk, on the other hand as WSS becomes larger the inference will be less efficient. From table (5) and table (6) we can conclude that PPS method gives a better privacy protection than the PIS method throughout for each choice of s_x and each experiment. On the other hand, from table (3) and table (4) it is clear that the powers at different choices of alternatives are larger for the test under PIS method than the PPS method. Therefore accuracy of inference and privacy protection work in opposite direction. One can raise a good point regarding a suitable adaptive sampling strategy which will offer privacy protection as well as valid inference, both at an acceptable level. This has been a longstanding issue with any privacy preserving mechanism and there is no satisfactory approach because while valid inference addresses how best to use the privacy-preserving data, the privacy part deals only with the mechanism of how data is perturbed and not how it is used afterwards. In our case, comparison of PIS and PPS procedures on the basis of power reported in Tables 3 and 4 naturally depends on the choice of the alternative hypothesis, and we find that PIS demonstrates more power than PPS uniformly for all alternatives. On the other hand, the privacy measures under PIS and PPS, reported in Tables 5 and 6 based on unequal sample sizes, show that PPS offers better privacy protection than PIS uniformly for all

the cases studied. Since a direct comparison of power and privacy measure is not apparent, unfortunately we are not able to come up with an adaptive sampling scheme. It is of course possible to compute power and privacy measures separately for PIS and PPS, as we have done in our paper, and select the right sampling scheme depending on the desirable acceptable levels. It is quite interesting to observe from our findings in Section 6 the robustness of the proposed F tests based on synthetic data under deviations from the assumption of normality although it does not quite hold for the F test based on the original data. In this paper we have used the usual F - statistic based on the synthetic data to carry out the tests for both PIS and PPS, but it is desirable to derive the Likelihood Ratio Test (LRT) for each of them. We wish to take it up in the future.

Acknowledgements

Our sincere thanks are due to an anonymous reviewer for some excellent comments which led to an improved version of the paper. The first author is thankful to Sister Nivedita University for providing the research facilities and the second author is thankful to Dr. Tommy Wright (Chief, Center for Statistical Research and Methodology at the Census Bureau) for his kind support and encouragement.

Declarations

Funding

No funding was received to assist with the preparation of this manuscript.

Conflicts of interest

The authors have no conflicts of interest to declare that are relevant to the content of this manuscript.

References

[1] Nayak, T., Zayatz, L. and Sinha, B. (2011) : 'Statistical Properties of Multiplicative Noise Masking for Confidentiality Protection'. *Journal of Official Statistics, 27, 527-544.*

[2] Nayak, T., Zayatz, L. and Sinha, B. (2011) : 'Privacy Protection and Quantile Estimation from Noise Multiplied Data'. *Sankhya, Series B, 73, 297-315.*

[3] Klein, M. and Sinha, B. (2013) : 'Statistical Analysis of Noise Multiplied Data Using Multiple Imputation'. *Journal of Official Statistics, 29, 425-465.*

[4] Klein, M., Mathew, T. and Sinha, B. (2014) : 'Noise Multiplication for Disclosure Limitation of Extreme Values in Log-normal Samples'. *Journal of Privacy and Confidentiality, 6, 77-125.*

[5] Drechsler, J. (2011) : Synthetic Datasets for Statistical Disclosure Control. *New York : Springer.*

[6] Kinney, S.K., Reiter, J.P., Reznek, A.P., Miranda, J., Jarmin, R.S., and Abowd, J.M. (2011) : Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database. *International Statistical Review, 79, 362-384.*

[7] Raghunathan, T.E., Reiter, J.P., and Rubin, D.B. (2003) : Multiple Imputation for Statistical Disclosure Limitation. *Journal of Official Statistics, 19, 1-16.*

[8] Reiter, J.P. (2003) : Inference for Partially Synthetic, Public Use Microdata Sets. *Survey Methodology, 29, 181-188.*

[9] Reiter, J.P., and Kinney, S.K. (2012) : Inferentially Valid, Partially Synthetic Data: Generating From Posterior Predictive Distributions Not Necessary. *Journal of Official Statistics, 28.* 583-590.

[10] Reiter, J.P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology, 30, 235-242.*

[11] Reiter, J.P. (2005a). Releasing multiply-imputed synthetic public use microdata: An illustration and empirical study. *Journal of Royal Statistical Society, Series A, 168, 185-205.*

[12] Reiter, J.P. (2005b). Significance tests for multi-component estimands from multiply imputed, synthetic microdata. *Journal of Statistical Planning and Inference, 131, 365-377.*

[13] Reiter, J.P. (2005c). Using CART to generate partially synthetic public use microdata. *Journal of Official Statistics, 21, 441-462.*

[14] Klein, M. and Sinha, B. (2015) : 'Likelihood-based inference for synthetic data based on a normal model'. *Statistics and Probability Letters, 168-175.*

[15] Klein, M. and Sinha, B. (2015) : 'Inference for Singly Imputed Synthetic Data Based on Posterior Predictive Sampling under Multivariate Normal and Multiple Linear Regression Models'. *Sankhya, Series B, 293-311.*

[16] Klein, M. and Sinha, B. (2016) : Likelihood-Based Finite Sample Inference for Singly Imputed Synthetic Data under Multivariate Normal and Multiple Linear Regression Models. *Journal of Privacy and Confidentiality, Volume 7, Number 1, 43-98.*

[17] Klein, M., Coelho, C., Moura, R. and Sinha, B. (2017) : Inference for Multivariate Regression Model based on Synthetic Data generated under Posterior and Fixed Posterior Predictive Sampling: Comparison with Plug-in Sampling. *REVSTAT.*

[18] Klein, M., Zylstra, J. and Sinha, B. (2019) : Finite Sample Inference for Multiply Imputed Synthetic Data under a Multiple Linear Regression Model. *Calcutta Statistical Association Bulletin, Volume 71, Issue 2, pp. 63-82.*

[19] Klein, M. and Sinha, B. (2019) : 'Multiple imputation for parametric inference under a differentially private Laplace mechanism'. *Technical report, Department of Math and Statistics, UMBC.*

[20] Kifle, Y. and Sinha, B. (2021) : 'Comparison of Some Exact Tests for a Common Location Parameter of Several Truncated Exponential Distributions with Different Scale Parameters'. *Sankhya B, Volume 83.1, 36 - 64.*

[21] Klein, M., Moura, R. and Sinha, B. (2021) : 'Multivariate normal inference based on singly imputed synthetci data under plug-in sampling'. *Sankhya B, Volume 83-B, Part 1, 273 - 287.*

[22] Moura, R., Klein, M., Zylstra, J., Coelho, C. and Sinha, B. (2021) : 'Inference for Multivariate Regression Model Based on Synthetic Data Generated Under Plug-In Sampling'. *Journal of American Statistical Association, pp. 1 - 41.*

[23] Tanweer Ul Islam and Erum Abbas (2022) : Validity of ANOVA under Non-normality & Heterogeneity. *Research Article*

[24] María J Blanca 1, Rafael Alarcón, Jaume Arnau, Roser Bono and Rebecca Bendayan (2017) : Non-normal data: Is ANOVA still a valid option? *Psicothema . 2017 Nov;29(4):552-557. PMID: 29048317.*

[25] Mukhtar M. Ali and Subhash C. Sharma (1996) : Robustness to nonnormality of regression F-tests. *Journal of Econometrics Volume 71, Issues 1–2, March–April 1996, Pages 175-205.*