Afrika Statistika
Vol. 18 (3), 2023, pages 3567 - 3602.
In : Special issue of the statistical applications of the fep and the *lrfep* to nonparametric estimation
DOI: https://dx.doi.org/10.16929/as/2023.3567.319.sFEP

Afrika Statistika

ISSN 2316-090X

Asymptotic Statistical Theory for the Samples Problems using the Functional Empirical Process.

Abdoulaye Camara (1,3,5,*) **Adja Mbarka Fall** (2,5), **Moumouni Diallo** (4,5)

(1) Université des Sciences, Techniques et des Technologies, Bamako, MALI
(2) Unversité Iba Der Thiam, Thiès, SENEGAL
(3) LERSTAD, Unversité Gaston Berger Saint-Louis, SENEGAL
(4) Université des Sciences Sociales et de Gestion, Bamako, MALI
(5) Imhotep International Mathematical Center (imho-imc), https://imhotepsciences.org

Received on January 01, 2023; Accepted on July 1, 2023; Published on September 01, 2024

Abstract In this paper we study the asymptotic theory for samples problem based on the functional empirical process (fep), this new method is called general samples problem. We suggest this method to develop the full theory of estimation of means, variances, ratios of variances and differences of means for independent samples. We compare the results of our new method to the Gaussian method using simulated and real data. The obtained results are almost equivalent to those in the Gaussian case for samples's size equal to 10 . It has been prove that the estimation of the mean difference is very precise regardless of the equality or inequality of variances for greater sizes of sample. This method is recommended when the sizes of samples is around or greater that 15 and it requires the finiteness of the second order moment.

Key words: asymptotic method; Gaussian method; independent samples problems; functional empirical process; nonparametric methods.
AMS 2010 Mathematics Subject Classification Objects : 62E20; 62G05; 62G10

(*) Correspondance Author:Abdoulaye Camara (camara.abdoulaye2@ugb.edu.sn)
Adja Mbarka Fall: adjambarka.fall@univ-thies.sn
Moumouni Diallo: moudiallo1@gmail.com

Résumé. (Abstract in French)

Dans cet article, nous étudions le problème de la théorie asymptotique pour les échantillons basé sur le processus empirique fonctionnel (fep), cette nouvelle méthode est appelée problème général des échantillons. Nous suggérons cette méthode pour développer la théorie complète de l'estimation des moyennes, des variances, des rapports de variances et des différences de moyennes pour des échantillons indépendants. Nous comparons les résultats de notre nouvelle méthode à la méthode gaussienne en utilisant des données simulées et réelles. Les résultats obtenus sont presque équivalents à ceux du cas gaussien pour une taille d'échantillon égale à 10. Il a été prouvé que l'estimation de la différence moyenne est très précise quelle que soit l'égalité ou l'inégalité des variances pour des échantillons de plus grande taille. Cette méthode est recommandée lorsque la taille des échantillons est autour de ou supérieure à 15 et nécessite la finitude du moment du second ordre.

Presentation of authors.

Abdoulaye Camara, M.Sc., is an assistant professor of Mathematics and Statistics at: Université des Sciences, Techniques et des Technologies, Bamako. He is preparing a PhD thesis at: LERSTAD, Unversité Gaston Berger Saint-louis, Senegal. He is also a junior researcher at: Mathematical Center (imho-imc), https://imhotepsciences.org.

Adja Mbarka Fall, PhD, is an Associated Professor of Mathematics and Statistics at: Université des Sciences, Techniques et des Technologies, Bamako. She is also a senior researcher at: Mathematical Center (imho-imc), https://imhotepsciences.org.

Moumouni Diallo, PhD, is a Professor of Mathematics and Statistics at:Université des Sciences Sociales et de Gestion, Bamako, MALI. He is a senior researcher at: Mathematical Center (imho-imc), https://imhotepsciences.org.

1. Introduction

In general, the statistical theory for linear methods in particular analysis of variance (ANOVA) and linear regression is mainly based on the Gaussian assumption. That theory of Gaussian linear models is generally well documend with a large number of teaching books and research books. At least we may cite a few including Seber (2003), Scheffé (1959), Tibshirani et al (2009), etc.

In that frame, the theory works fine. However in many situations the data are not Gaussian. For example accidental data are more likely to follow the gamma law. It happens that the drift from the Gaussian hypothesis can not lead to a so beautiful theory. Let us consider two types of sets.

(a) In Scheffé (1959), we pick the data as in page 3593. Therein, we have measurements of waterproof quality of sheets of material manufactured on five machines on nine days. Here, we have three batches of data. In each batch, lines refer to data from machine A, B and C. The study here is whether or not a machine gives more robust sheets that another in the sens of higher quality of waterproof. The normality Jarque-Bera test works fine for all variables (taken horizontally) and their differences, as we checked it in Subsection 5.1 of Section 5. For these data, the Gaussian methods are relevant and may be used.

(b) Income data from database ESAM1 (1996) and ESAM2 (2000) are considered. For regions Dakar and Diourbel, the income data are dakar1 and dakar2, diour1 and diour2. In this study, we consider a sampling of T observation from each set of data, just for illustration. We want to compare the the revenue between the two areas for each survey and eventually to try to see the evolution for one region, from one survey to the other. As we will see it just in the subsection mentioned above, these data are clearly non Gaussian.

In situations like (b), it is not reasonable to use the Gaussian method. As we already mentioned, there is a significant number of areas in which Gaussian data are not expected. A clear example concerns heavy tails data or mixing data in which they are involved.

In such situations, we propose to use asymptotic methods in the spirit of Lo et al. (2023) through asymptotic representations of the involved statistics. That method is a consequence of the whole theory of weak convergence. In the above cite article, a full list of related works and books on weak convergence can be found.

Our best achievement is the setting of full theory of estimation of means, variances, ratios of variances and differences of means for independent samples with sizes around or greater that $n \geq 15$. The asymptotic laws needed for building the confidence intervals are given in Theorems 1 (page 3574) and 2 (page 3579). Those confidence intervals are displayed Subsection 5.1 (page 3587) for the one sample problem and Subsection 5.5 (page 3589) for the two independent sample problem.

The rest of the paper is organized as follows. We begin by recalling the nice Gaussian theory in Section 2 while we give our results in Section 3 where a full theory of comparison based on the empirical process for general data. We focus on the implementation of the results Section 4. Next the results are simulated and applied to datadriven examples in Section 5 where a number of scenarii are presented and discussed. We close the paper with Section 6 as the paper conclusion.

2. The Gaussian Samples problems

2.1. Introduction

The samples problems for Gaussian data are treated in a parametric frame and use exact distribution theory. It is quite impossible to have a complete theory for other distributions. In order to able to highlight the results obtained from the use of the *fep*, we are going to recall the striking distribution facts for Gaussian samples.

This part introduces to the standard statistical methods related to Gaussian samples. It comes after the study of Gaussian vectors, of which it presents statistical applications. It is a preparation to the wider field of linear regression, analysis of variance (ANOVA) and analysis of covariance (ANCOVA). We will focus on the one sample and the two sample problems.

The techniques given here, in despite of their simplicity, yield very powerful methods in a number of fields, including Life Sciences, Medical and Pharmaceutical Sciences, Industrial Studies, etc., in short, in any discipline in which data may be modeled as Gaussian.

2.2. The Gaussian sample problem techniques

A Gaussian sample is simply of size n a sequence $X_1, X_2, ..., X_n$ of n independent and identically distributed (*iid*) random variables with a Gaussian common probability law of mean m and σ^2 denoted

$$(H) \quad X_1, X_2, ..., X_n \ \sim \mathcal{N}(m, \sigma^2).$$

We always suppose that $n \geq 2$, otherwise doing statistics is meaningless. The two main statistics are the sample mean:

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{1}$$

and the sample variance

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_n - \overline{X}_n\right)^2. \tag{2}$$

2.3. The one Gaussian sample problem

Here is the most important result concerning the distributional properties of the Gaussian sample

Proposition 1. *For all $n \geq 2$, we have :*

(a) S_n^2 is independent of $\overline{X}_n$

(b) Law of $\overline{X}_n$, when σ is known.

$$\frac{\sqrt{n}\left(\overline{X}_n - m\right)}{\sigma} \sim \mathcal{N}(0,1) \tag{3}$$

(c) Law of S_n^2.

$$(n-1)\, S_n^2/\sigma^2 \sim \chi^2_{n-1} \tag{4}$$

(d) Law of $\overline{X}_n$, when σ is unknown.

$$\frac{\sqrt{n}\left(\overline{X}_n - m\right)}{S_n} \sim t(n-1). \tag{5}$$

2.4. Two independent Gaussian samples problems

We have a first easy case where the two data to be compared are linked to the same statistical unit.

A - Two paired sample Gaussian problem.

Suppose that we have paired measures (X_i, Y_i) associated with the same individuals, $i = 1, ..., n$. For example, for the same person, X might represent blood sugar and Y the cholesterol measure.

Assume that the $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ are independent Gaussian random couples such that

$$(H0) \ \ \forall i \leq i \leq n, \ \mathbb{E}X_i = m_1 \textit{ and } \mathbb{E}Y_i = m_2$$

We want to test the hypothesis

$$(H) \ \ m_1 = m_2.$$

which is equivalent to test, $\Delta m = m_1 - m_2 = 0$.

Put

$$Z_i = X_i - Y_i, \ \ i = 1, ..., n.$$

Given that the pair (X_1, Y_2) is Gaussian, we can and do transform the two samples problem a one sample problem of the data $Z_1, ..., Z_n$ of common probability law $\mathcal{N}\left(\Delta_m, \sigma^2\right)$, where $\sigma^2 = \mathbb{V}(X_1 - X_2)$.

Now, the most interesting case is the following.

B - The two Gaussian independent samples.

Suppose we have

$$\mathcal{E}_X : \ X_1, X_2, ..., X_n \ \text{ iid } \sim \mathcal{N}\left(m_1, \sigma_1^2\right)$$

and

$$\mathcal{E}_Y : \ Y_1, Y_2, ..., Y_m \ \text{ iid } \sim \mathcal{N}\left(m_2, \sigma_2^2\right)$$

B1 - The law of the ratio of variances

We denote by S_n^2 and S_m^2 the empirical variances for the samples $\mathcal{E}_X$ and $\mathcal{E}_Y$ respectively. We consider the ratio of variances $R = \sigma_1^2/\sigma_2^2$.

Proposition 2. *We have*

$$\frac{1}{R^2}\frac{S_n^2}{S_m^2} \sim F_{n,m},$$

where $F_{n,m}$ stands for the Fisher law of degrees of freedom n and m.

B2 - Estimation of Δm.

The estimation is done in two cases.

B2a - Case of equal variances.

Denote $\sigma_1^2 = \sigma_2^2 = \sigma^2$ and

$$S^2 = \frac{(n-1)\, S_n^2 + (m-1)\, S_m^2}{n+m-2}$$

Proposition 3. *We have*

$$\frac{\overline{X}_n - \overline{Y}_m - \Delta m}{s\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim t(n+m-2)$$

B2b - Case of unequal variances

In the literature, in such a case, the approximation of *Aspin-Welch test* and *Welch's t-test* (Welch, 1937) using the *Satterthwaite* method is used as follows:

Proposition 4.

$$\frac{\overline{X}_n - \overline{Y}_m - \Delta m}{\sqrt{\frac{S_n^2}{n} + \frac{S_m^2}{m}}} \sim t(f),$$

where

$$f = \frac{\left(\frac{S_n^2}{n} + \frac{S_m^2}{m}\right)^2}{\frac{1}{n-1}\left(\frac{S_n^2}{n}\right)^2 + \frac{1}{m-1}\left(\frac{S_m^2}{m}\right)^2}$$

Remark. This fascinated and most used tool for Gaussian data is justified in our simulation works below. We still think that a proof at the level of the modern weak convergence theory is welcome. We do think that the approached developed in this paper is the right one.

3. Alternative theory using the fep

In this section, we are concerned with asymptotic theory of samples problems based on the functional empirical process.

3.1. The one sample problem

Let X, X_1, X_2, ..., be a sequence of independent and identically distributed (iid) random variables defined on the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with the functional empirical process (fep) $\mathbb{G}_n$.

We assume that

$$\forall i, 1 \leq i \leq n, \quad \mathbb{E}(X_i) = m, \quad \mathbb{V}ar(X_i) = \sigma^2$$

$$m_k = \mathbb{E}(X^k), \quad \mu_k = \mathbb{E}\left(X - \mathbb{E}(X)\right)^k, \quad \mu_{k,n} = \frac{1}{n}\sum_{j=1}^{n}\left(X_j - \overline{X}_n\right)^k, \quad k \geq 1$$

We define the statistic T_n^2 given by

$$T_n^2 = \mu_{4,n} - S_n^4 \tag{6}$$

We have two main results relative to the mean m and the variance σ^2.

Theorem 1. *For all $n \geq 2$, we have :*

(a) $\frac{\sqrt{n}(\overline{X}_n - m)}{S_n} \sim \mathcal{N}(0,1)$

(b) $\frac{\sqrt{n}(S_n^2 - \sigma^2)}{T_n} \sim \mathcal{N}(0,1)$

Proof of Theorem 1

Proof of (a). Let $\alpha_n = \frac{\sqrt{n}(\overline{X}_n - m)}{S_n}$. By using the asymptotic representation, we have

$$\overline{X}_n = m + n^{-1/2}\mathbb{G}_n(h_1) \tag{7}$$

that implies that

$$\overline{X}_n - m = n^{-1/2}\mathbb{G}_n(h_1) \tag{8}$$

$$S_n^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2\right) = \frac{n}{n-1}\left(\overline{X_n^2} - \overline{X}_n^2\right) \tag{9}$$

By the delta method, we have

$$\overline{X_n}^2 = m^2 + n^{-1/2}\mathbb{G}_n(2mh_1) + o_{\mathbb{P}}\left(n^{-1/2}\right) \tag{10}$$

and

$$\overline{X_n^2} = m_2 + n^{-1/2}\mathbb{G}_n(h_2) \tag{11}$$

where $h_j(x) = x^j$. We know that

$$\frac{n}{n-1} = \left(\frac{n-1}{n}\right)^{-1} = (1 - 1/n)^{-1} = 1 + O_{\mathbb{P}}(1)$$

Hence, we have

$$S_n^2 = \sigma^2 + n^{-1/2}\mathbb{G}_n(H) + o_{\mathbb{P}}\left(n^{-1/2}\right) \tag{12}$$

where $H = h_2 - 2mh_1$. By the delta method for S_n^2, we get

$$S_n = \sigma + n^{-1/2}\mathbb{G}_n\left(\frac{1}{\sigma}H\right) + o_{\mathbb{P}}\left(n^{-1/2}\right) \tag{13}$$

By combining the asymptotic representation for S_n and $\overline{X}_n - m$, we obtain

$$\frac{\overline{X}_n - m}{S_n} = n^{-1/2}\mathbb{G}_n\left(\frac{1}{\sigma}h_1\right) + o_{\mathbb{P}}\left(n^{-1/2}\right) = n^{-1/2}\mathbb{G}_n\left(h\right) + o_{\mathbb{P}}\left(n^{-1/2}\right)$$

where $h = \frac{1}{\sigma}h_1$. Thus

$$\alpha_n = \frac{\sqrt{n}\left(\overline{X}_n - m\right)}{S_n} = \mathbb{G}_n\left(h\right) + o_{\mathbb{P}}(1) \tag{14}$$

Hence $\alpha_n \sim \mathcal{N}\left(0, \Gamma(h,h)\right)$,

$$\Gamma(h,h) = \mathbb{V}ar(h(X)) = \mathbb{V}ar\left(\frac{1}{\sigma}h_1(X)\right) = \frac{1}{\sigma^2}\mathbb{V}ar(h_1(X)) = 1$$

So $\alpha_n \sim \mathcal{N}\left(0,1\right)$.

Proof of (b).

Let $J_n = \frac{\sqrt{n}(S_n^2 - \sigma^2)}{\rho_n}$. We know that

$$S_n^2 = \sigma^2 + n^{-1/2}\mathbb{G}_n(H) + o_{\mathbb{P}}\left(n^{-1/2}\right), \tag{15}$$

where $H = h_2 - 2mh_1$. We have

$$S_n^2 - \sigma^2 = n^{-1/2}\mathbb{G}_n(H) + o_{\mathbb{P}}\left(n^{-1/2}\right) \tag{16}$$

The variance of H is given by

$$\mathbb{V}ar\left(H(X)\right) = \mu_4 - \sigma^4$$

We also know that

$$\mu_{4,n} = \mu_4 + n^{-1/2}\mathbb{G}_n(C_0) + o_{\mathbb{P}}\left(n^{-1/2}\right) \tag{17}$$

where $C_0 = h_4 + 4mh_3 + 6m^2h_2 + (-4m_3 + 12mm_2 - 12m^3)h_1$. By the delta method, we have

$$S_n^4 = \sigma^4 + n^{-1/2}\mathbb{G}_n\left(2\sigma^2 H\right) + o_{\mathbb{P}}\left(n^{-1/2}\right) \tag{18}$$

From the asymptotic representation of $\mu_{4,n}$ and S_n^4, we get

$$T_n^2 = T^2 + n^{-1/2}\mathbb{G}_n\left(C_0 - 2\sigma^2 H\right) + o_{\mathbb{P}}\left(n^{-1/2}\right)$$

where $T^2 = \mu_4 - \sigma^4$. By the delta method, we have

$$T_n = T + n^{-1/2}\mathbb{G}_n\left(L\right) + o_{\mathbb{P}}\left(n^{-1/2}\right) \tag{19}$$

where $L = C_0 - 2\sigma^2 H$. From the asymptotic representation of $S_n^2 - \sigma^2$ and T_n, we get

$$\frac{S_n^2 - \sigma^2}{T_n} = n^{-1/2}\mathbb{G}_n\left(\frac{1}{T}H\right) + o_{\mathbb{P}}\left(n^{-1/2}\right) \tag{20}$$

Therefore, we have

$$\frac{\sqrt{n}\left(S_n^2 - \sigma^2\right)}{T_n} = \mathbb{G}_n\left(D\right) + o_{\mathbb{P}}\left(1\right) \tag{21}$$

where $D = \frac{1}{T}H$. Thus

$$\frac{\sqrt{n}\left(S_n^2 - \sigma^2\right)}{T_n} \sim \mathcal{N}\left(0, \Gamma(D, D)\right)$$

$$\Gamma(D, D) = \mathbb{V}ar\left(D\right) = \frac{1}{T^2}\mathbb{V}ar\left(H\right) = \frac{T^2}{T^2} = 1$$

Hence

$$\frac{\sqrt{n}\left(S_n^2 - \sigma^2\right)}{T_n} \sim \mathcal{N}\left(0, 1\right)$$

We can also write T_n^2 as

$$T_n^2 = S_n^2\left(\frac{\mu_{4,n}}{S_n^2} - S_n^2\right) = S_n^2\left(K_n(X) - S_n^2\right)$$

Therefore

$$\frac{\sqrt{n}\left(S_n^2 - \sigma^2\right)}{S_n\left(K_n(X) - S_n^2\right)} \sim \mathcal{N}\left(0, 1\right)$$

3.2. The Two Independent Samples Problem

Let X_1, ..., X_{n_1} and Y_1, ..., Y_{n_2} be independent random variables, where X_1, ..., X_{n_1} are iid with the functional empirical process (fep) $\mathbb{G}_{n_1}$ and Y_1, ..., Y_{n_2} are iid with the functional empirical process (fep) $\mathbb{G}_{n_2}$. The functional empirical process $\mathbb{G}_{n_1}$ and $\mathbb{G}_{n_2}$ are assumed independent.

We have the notation:

$$\forall \ \ i, \ \ 1 \leq i \leq n_1, \ \ \mathbb{E}X_i = \mu_x; \ \ \mathbb{V}arX_i = \sigma_1^2, \quad m_{k,x} = \mathbb{E}X^k, \ \ \mu_{k,x} = \mathbb{E}\left(X - \mu_x\right)^k, \ \ k \geq 1$$

$$\forall \ \ j, \ \ 1 \leq j \leq n_2, \ \ \mathbb{E}Y_j = \mu_y; \ \ \mathbb{V}arY_j = \sigma_2^2, \quad m_{k,y} = \mathbb{E}Y^k, \ \ \mu_{k,y} = \mathbb{E}\left(Y - \mu_y\right)^k, \ \ k \geq 1$$

We always suppose that $n_1 \geq 2$ and $n_2 \geq 2$. The main statistics are the sample mean and the sample variance for each sample. For the sample with size n_1, we have

$$\overline{X}_{n_1} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i \tag{22}$$

and

$$S_{n_1}^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}\left(X_i - \overline{X}_{n_1}\right)^2 \tag{23}$$

For the sample with size n_2, we have

$$\overline{Y}_{n_2} = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_j \tag{24}$$

and

$$S_{n_2}^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_1}\left(Y_j - \overline{Y}_{n_2}\right)^2 \tag{25}$$

We have

$$\overline{X}_{n_1} = \mu_x + n_1^{-1/2}\mathbb{G}_{n_1}\left(\pi_1\right), \quad \pi_1(x,y) = h_1(x)$$

$$\overline{Y}_{n_2} = \mu_y + n_2^{-1/2}\mathbb{G}_{n_2}\left(\pi_2\right), \quad \pi_2(x,y) = h_1(y)$$

$$\overline{X^2}_{n_1} = m_{2,x} + n_1^{-1/2}\mathbb{G}_{n_1}\left(\pi_3\right), \quad \pi_3(x,y) = h_2(x)$$

$$\overline{Y^2}_{n_2} = m_{2,y} + n_2^{-1/2}\mathbb{G}_{n_2}(\pi_4), \quad \pi_4(x,y) = h_2(y)$$

By the delta method, we have

$$\overline{X}^2_{n_1} = \mu_x^2 + n_1^{-1/2}\mathbb{G}_{n_1}(2\mu_x\pi_1) + o_{\mathbb{P}}\left(n_1^{-1/2}\right)$$

$$\overline{Y}^2_{n_2} = \mu_y^2 + n_2^{-1/2}\mathbb{G}_{n_2}(2\mu_y\pi_2) + o_{\mathbb{P}}\left(n_2^{-1/2}\right)$$

$$S^2_{n_1} = \sigma_x^2 + n_1^{-1/2}\mathbb{G}_{n_1}(H_1) + o_{\mathbb{P}}\left(n_1^{-1/2}\right) \tag{26}$$

where $H_1 = \pi_3 - 2\mu_x\pi_1 = h_2 - 2\mu_x h_1$

$$S^2_{n_2} = \sigma_y^2 + n_2^{-1/2}\mathbb{G}_{n_2}(H_2) + o_{\mathbb{P}}\left(n_2^{-1/2}\right) \tag{27}$$

where $H_2 = \pi_4 - 2\mu_y\pi_2 = h_2 - 2\mu_y h_1$. Now we will study the law of $\left(\frac{S^2_{n_1}}{S^2_{n_2}}\right)$.

Let $n = n_1 \wedge n_2$.

We easily check $O_{\mathbb{P}}(n_j^{-1/2}) = O_{\mathbb{P}}(n^{-1/2}), j = 1, 2$. We also assume that $n_1 \longrightarrow +\infty$; $n_2 \longrightarrow +\infty$ and (n_1) and (n_2) are balanced between them

(H) $\overline{\lim}(\frac{n_1}{n_2}) < +\infty$ and $\overline{\lim}(\frac{n_2}{n_1}) < +\infty$

We have

$$\frac{S^2_{n_1}}{S^2_{n_2}} = \frac{\sigma_1^2 + n_1^{-1/2}\mathbb{G}_{n_1}(H_1) + o_{\mathbb{P}}\left(n_1^{-1/2}\right)}{\sigma_2^2 + n_2^{-1/2}\mathbb{G}_{n_2}(H_2) + o_{\mathbb{P}}\left(n_2^{-1/2}\right)}$$

By the delta method, we have

$$\left(\frac{S^2_{n_1}}{S^2_{n_2}} - \frac{\sigma_1^2}{\sigma_2^2}\right) = n_1^{-1/2}\mathbb{G}_{n_1}\left(\frac{1}{\sigma_2^2}H_1\right) - n_2^{-1/2}\mathbb{G}_{n_2}\left(\frac{\sigma_1^2}{\sigma_2^4}H_2\right) + o_{\mathbb{P}}\left(n^{-1/2}\right) \tag{28}$$

Let

$$N_{1,n} = \mathbb{G}_{n_1}\left(\frac{1}{\sigma_2^2}H_1\right) = \mathcal{N}_1\left(0, T_1^2\right) + o_{\mathbb{P}}(1)$$

$$N_{2,n} = \mathbb{G}_{n_2}\left(\frac{2\sigma_1^2}{\sigma_2^2}H_2\right) = -\mathcal{N}_2\left(0, T_2^2\right) + o_{\mathbb{P}}(1)$$

$\mathcal{N}_1$ and $\mathcal{N}_2$ are Gaussian and independent.

$$T_1^2 = \mathbb{V}ar\left(\frac{1}{\sigma_2^2}H_1\right) = \frac{\mu_{4,x} - \sigma_1^4}{\sigma_2^4}$$

$$T_2^2 = \mathbb{V}ar\left(\frac{\sigma_1^2}{\sigma_2^2}H_1\right) = \frac{\sigma_1^4\left(\mu_{4,y} - \sigma_2^4\right)}{\sigma_2^8}$$

Let take $a_n = \sqrt{\frac{n_1 n_2}{n_1 T_2^2 + n_2 T_1^2}}$

Theorem 2. *We have*
(A) For all $n_1, n_2 \geq 2$

$$a_n\left(\frac{S_{n_1}^2}{S_{n_2}^2} - \frac{\sigma_1^2}{\sigma_2^2}\right) \sim \mathcal{N}(0,1)$$

(B) For all $n_1 \geq 2$ *and For all* $n_2 \geq 2$

$$\sqrt{\frac{n_1 n_2}{n_2 T_{1,n}^2 + n_1 T_{2,n}^2}}\left(\frac{S_{n_1}^2}{S_{n_2}^2} - \frac{\sigma_1^2}{\sigma_2^2}\right) \sim \mathcal{N}(0,1)$$

Proof of Part (A) of Theorem 2. We recall that

$$\left(\frac{S_{n_1}^2}{S_{n_2}^2} - \frac{\sigma_1^2}{\sigma_2^2}\right) = n_1^{-1/2}\mathcal{N}_1\left(0, T_1^2\right) + n_1^{-1/2}o_{\mathbb{P}}(1) + n_2^{-1/2}\mathcal{N}_2\left(0, T_2^2\right) + n_2^{-1/2}o_{\mathbb{P}}(1) + o_{\mathbb{P}}\left(n^{-1/2}\right)$$

Thus we have xyz

$$\begin{aligned}
a_n\left(\frac{S_{n_1}^2}{S_{n_2}^2} - \frac{\sigma_1^2}{\sigma_2^2}\right) &= \frac{a_n}{\sqrt{n_1}}\mathcal{N}_1\left(0, T_1^2\right) + \frac{a_n}{\sqrt{n_1}}o_{\mathbb{P}}(1) + \frac{a_n}{\sqrt{n_2}}\mathcal{N}_2\left(0, T_2^2\right) + \frac{a_n}{\sqrt{n_2}}o_{\mathbb{P}}(1) + a_n o_{\mathbb{P}}\left(n^{-1/2}\right) \\
&= \left(\frac{a_n}{\sqrt{n_1}}\mathcal{N}_1\left(0, T_1^2\right) + \frac{a_n}{\sqrt{n_2}}\mathcal{N}_2\left(0, T_2^2\right)\right) \\
&+ \left(\frac{a_n}{\sqrt{n_1}}o_{\mathbb{P}}(1) + \frac{a_n}{\sqrt{n_2}}o_{\mathbb{P}}(1) + a_n o_{\mathbb{P}}\left(n^{-1/2}\right)\right) \\
&= \mathcal{N}_n + R_n
\end{aligned}$$

where

$$\mathcal{N}_n = \frac{a_n}{\sqrt{n_1}}\mathcal{N}_1\left(0, T_1^2\right) + \frac{a_n}{\sqrt{n_2}}\mathcal{N}_2\left(0, T_2^2\right)$$

$$R_n = \frac{a_n}{\sqrt{n_1}}o_{\mathbb{P}}(1) + \frac{a_n}{\sqrt{n_2}}o_{\mathbb{P}}(1) + a_n o_{\mathbb{P}}\left(n^{-1/2}\right) = o_{\mathbb{P}}(1).$$

But $\mathcal{N}_n \sim \mathcal{N}\left(0, A^2\right)$ with

$$A^2 = \frac{a_n^2}{n_1}T_1^2 + \frac{a_n^2}{n_2}T_2^2 = \frac{n_2 T_1^2}{n_1 T_2^2 + n_2 T_1^2} + \frac{n_1 T_2^2}{n_1 T_2^2 + n_2 T_1^2} = 1$$

Thus

$$a_n\left(\frac{S_{n_1}^2}{S_{n_2}^2} - \frac{\sigma_1^2}{\sigma_2^2}\right) = \mathcal{N}_n + o_{\mathbb{P}}(1) \sim \mathcal{N}\left(0, A^2\right).$$

Therefore

$$a_n\left(\frac{S_{n_1}^2}{S_{n_2}^2} - \frac{\sigma_1^2}{\sigma_2^2}\right) \sim \mathcal{N}\left(0, 1\right)$$

Let

$$T_{1,n}^2 = \frac{\mu_{4,x} - S_{n_1}^4}{S_{n_2}^4}; \quad T_{2,n}^2 = \frac{S_{n_1}^4\left(\mu_{4,y} - S_{n_2}^4\right)}{S_{n_2}^8}.$$

Proof of Part (B) of Theorem 2. We have

$$T_{1,n}^2 = T_1^2 + n_1^{-1/2}\mathbb{G}_{n_1}\left(D_{11}\right) + n_2^{-1/2}\mathbb{G}_{n_2}\left(D_{12}\right) + o_{\mathbb{P}}\left(n^{-1/2}\right)$$

$$T_{2,n}^2 = T_2^2 + n_1^{-1/2}\mathbb{G}_{n_1}\left(D_{21}\right) + n_2^{-1/2}\mathbb{G}_{n_2}\left(D_{22}\right) + o_{\mathbb{P}}\left(n^{-1/2}\right)$$

We do not need to know explicit forms of the D_{ij}. Thus we have

$$\frac{T_{1,n}^2}{n_1} = \frac{T_1^2}{n_1} + O_{\mathbb{P}}\left(n^{-3/2}\right) = \frac{T_1^2}{n_1}\left(1 + O_{\mathbb{P}}\left(n^{-1/2}\right)\right)$$

$$\frac{T_{2,n}^2}{n_2} = \frac{T_2^2}{n_2} + O_{\mathbb{P}}\left(n^{-3/2}\right) = \frac{T_2^2}{n_2}\left(1 + O_{\mathbb{P}}\left(n^{-1/2}\right)\right)$$

Let

$$\hat{a}_n = \left(\frac{1}{\frac{T_1^2}{n_1}\left(1 + O_{\mathbb{P}}\left(n^{-1/2}\right)\right) + \frac{T_2^2}{n_2}\left(1 + O_{\mathbb{P}}\left(n^{-1/2}\right)\right)}\right)^{1/2}$$

We recall that $a_n = \left(\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2}\right)^{-1/2}$

$$\hat{a}_n - a_n = \left(\hat{b}_n\right)^{-1/2} - (b_n)^{-1/2} = \left(\hat{b}_n - b_n\right)\left(\bar{b}_n\right)^{-3/2}$$

where $\bar{b}_n \in \left[\hat{b}_n \wedge b_n, \hat{b}_n \vee b_n\right]$

Let $b_n = \frac{T_1^2}{n_1} + \frac{T_2^2}{n_2}$ *and* $\hat{b}_n = \frac{T_1^2}{n_1}\left(1 + O_{\mathbb{P}}\left(n^{-1/2}\right)\right) + \frac{T_2^2}{n_2}\left(1 + O_{\mathbb{P}}\left(n^{-1/2}\right)\right)$

We have

$$\hat{b}_n - b_n = O_{\mathbb{P}}\left(n^{-1/2}\right)$$

$$\frac{b_n}{\hat{b}_n} = \frac{\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2}}{\frac{T_1^2}{n_1}\left(1 + O_{\mathbb{P}}(1)\right) + \frac{T_2^2}{n_2}\left(1 + O_{\mathbb{P}}(1)\right)}$$

By multiplying the numerator and denominator by $n_1 n_2$, we have

$$\frac{b_n}{\hat{b}_n} = \frac{n_2 T_1^2 + n_1 T_2^2}{n_2 T_1^2\left(1 + O_{\mathbb{P}}(1)\right) + n_1 T_2^2\left(1 + O_{\mathbb{P}}(1)\right)}$$

$$\frac{b_n}{\hat{b}_n} = \frac{T_1^2 + \left(\frac{n_1}{n_2}\right) T_2^2}{T_1^2\left(1 + O_{\mathbb{P}}(1)\right) + \left(\frac{n_1}{n_2}\right) T_2^2} = O_{\mathbb{P}}(1)$$

That means $b_n \asymp \hat{b}_n$. Hence

$$\hat{a}_n = \left(\frac{1}{\frac{T_1^2}{n_1}\left(1 + O_{\mathbb{P}}\left(n^{-1/2}\right)\right) + \frac{T_2^2}{n_2}\left(1 + O_{\mathbb{P}}\left(n^{-1/2}\right)\right)}\right)^{1/2} = \left(\frac{n_1 n_2}{\frac{n_2}{n_1} T_1^2 + \frac{n_1}{n_2} T_2^2 + O_{\mathbb{P}}\left(n^{-1/2}\right)}\right)^{1/2}$$

Since $\frac{n_1}{n_2} \asymp 1$, so we have

$$\hat{a}_n = \left(\frac{n_1 n_2}{\frac{n_2}{n_1} T_1^2 + \frac{n_1}{n_2} T_2^2}\right)^{1/2} + O_{\mathbb{P}}\left(n^{-1/2}\right)$$

$$\begin{aligned}\hat{a}_n\left(\frac{S_{n_1}^2}{S_{n_2}^2} - \frac{\sigma_1^2}{\sigma_2^2}\right) &= \sqrt{\frac{n_1 n_2}{\frac{n_2}{n_1} T_1^2 + \frac{n_1}{n_2} T_2^2}}\,\mathbb{G}_{n_1}\left(\frac{1}{\sigma_2^2} H_1\right) + \sqrt{\frac{n_1 n_2}{\frac{n_2}{n_1} T_1^2 + \frac{n_1}{n_2} T_2^2}}\,\mathbb{G}_{n_2}\left(\frac{2\sigma_1^2}{\sigma_2^2} H_2\right) + O_{\mathbb{P}}\left(n^{-1/2}\right) \\ &= \mathcal{N}_n(1) + \mathcal{N}_n(2)\end{aligned}$$

i.e., due to *(H)*, we get

$$\hat{a}_n = a_n + O_{\mathbb{P}}\left(n^{-1/2}\right) = a_n + o_{\mathbb{P}}\left(1\right),$$

and can conclude that

$$\begin{aligned}\hat{a}_n\left(\frac{S_{n_1}^2}{S_{n_2}^2} - \frac{\sigma_1^2}{\sigma_2^2}\right) &= a_n\left(\frac{S_{n_1}^2}{S_{n_2}^2} - \frac{\sigma_1^2}{\sigma_2^2}\right) + o_{\mathbb{P}}(1)\left(\frac{S_{n_1}^2}{S_{n_2}^2} - \frac{\sigma_1^2}{\sigma_2^2}\right)\\ &= a_n\left(\frac{S_{n_1}^2}{S_{n_2}^2} - \frac{\sigma_1^2}{\sigma_2^2}\right) + o_{\mathbb{P}}(1)\end{aligned}$$

Since $a_n\left(\frac{S_{n_1}^2}{S_{n_2}^2} - \frac{\sigma_1^2}{\sigma_2^2}\right) \sim \mathcal{N}\left(0,1\right)$, so we can conclude that

$$\hat{a}_n\left(\frac{S_{n_1}^2}{S_{n_2}^2} - \frac{\sigma_1^2}{\sigma_2^2}\right) \sim \mathcal{N}\left(0,1\right)$$

Estimation of Δm

We have

$$\begin{aligned}\overline{X}_{n_1} - \overline{Y}_{n_2} - \Delta m &= n_1^{-1/2}\mathbb{G}_{n_1}\left(h_1\right) - n_2^{-1/2}\mathbb{G}_{n_2}\left(h_1\right) + o_{\mathbb{P}}\left(n^{-1/2}\right)\\ &= n_1^{-1/2}\left[\mathcal{N}_1\left(0,\sigma_1^2\right) + o_{\mathbb{P}}\left(1\right)\right] + n_2^{-1/2}\left[\mathcal{N}_2\left(0,\sigma_2^2\right) + o_{\mathbb{P}}\left(1\right)\right] + o_{\mathbb{P}}\left(n^{-1/2}\right)\\ &= \mathcal{N}_n + o_{\mathbb{P}}\left(n_1^{-1/2}\right) + o_{\mathbb{P}}\left(n_2^{-1/2}\right) + o_{\mathbb{P}}\left(n^{-1/2}\right)\\ &= \mathcal{N}_n + o_{\mathbb{P}}\left(n^{-1/2}\right)\end{aligned}$$

where

$$\mathcal{N}_n = n_1^{-1/2}\left[\mathcal{N}_1\left(0,\sigma_1^2\right)\right] + n_2^{-1/2}\left[\mathcal{N}_2\left(0,\sigma_2^2\right)\right] \sim \mathcal{N}\left(0, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

$\mathcal{N}_1$ independent with $\mathcal{N}_2$. For

$$\tau_1^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{n_2\sigma_1^2 + n_1\sigma_2^2}{n_1 n_2}$$

Then

$$\sqrt{\frac{n_1 n_2}{n_2\sigma_1^2 + n_1\sigma_2^2}}\left(\overline{X}_{n_1} - \overline{Y}_{n_2} - \Delta m\right) \sim \mathcal{N}\left(0,1\right)$$

Let us replace $b_n = \sqrt{\frac{n_1 n_2}{n_2\sigma_1^2 + n_1\sigma_2^2}}$ by $\hat{b}_n = \sqrt{\frac{n_1 n_2}{n_2 S_{n_1}^2 + n_1 S_{n_2}^2}}$

Now

$$\hat{b}_n = \sqrt{\frac{n_1 n_2}{n_2\left(\sigma_1^2 + o_{\mathbb{P}}(1)\right) + n_1\left(\sigma_2^2 + o_{\mathbb{P}}(1)\right)}}$$

We can easily prove, as in the letter proof, that

$$\hat{b}_n = b_n + O_{\mathbb{P}}\left(n^{-1/2}\right)$$

Hence

$$\hat{b}_n\left(\overline{X}_{n_1} - \overline{Y}_{n_2} - \Delta m\right) = b_n\left(n_1^{-1/2}\mathbb{G}_{n_1}(h_1) - n_2^{-1/2}\mathbb{G}_{n_2}(h_2)\right) + o_{\mathbb{P}}(1) \sim \mathcal{N}(0,1)$$

Therefore

$$\sqrt{\frac{n_1 n_2}{n_2 S_{n_1}^2 + n_1 S_{n_2}^2}}\left(\overline{X}_{n_1} - \overline{Y}_{n_2} - \Delta m\right) \sim \mathcal{N}(0,1)$$

4. Implementations

Of course the Gaussian method, as an exact one, is preferable if we have the right type of data. Otherwise, we use the *fep* theory as we did above. A striking fact is that we do not need to test the equality of variances in the two independent samples and by this we avoid using the *Satterthwaite*.

From results in both situations, we draw confidence interval for estimating the means, the differences of means, the variances and the variances ratios. The results are summarized in tables below. Here, we harmonize the notation regarding the notation in the two independent samples problem. In the Gaussian case, we denoted the sample means and variances by $\{\overline{X}_n,\ S_n^2\}$ and $\{\overline{Y}_m,\ S_m^2\}$ with sizes n and m while in the general case, we used $\{\overline{X}_{n_1},\ S_{n_1}^2\}$ and $\{\overline{Y}_{n_2},\ S_{n_2}^2\}$ with sizes n_1 and n_2. In the sequel, the notation in the general case will prevail and in general n_1 and n_2 are used only in the two sample problem.

Let us recall some important definitions.

For some sample of size n

$\overline{X}_n = \sum_{j=1}^{n} X_j$	$S_n^2 = \sum_{j=1}^{n}(X_j - \overline{X}_n)^2$
$\mu_{4,n} = \sum_{j=1}^{n}(X_j - \overline{X}_n)^2$	$T_n^2 = \mu_{4,n} - S_n^4$

For tow samples of size n_1 and n_2

$T_{n_1}^2 = \mu_{4,n_1} - S_{n_1}^4$	$T_{n_2}^2 = \mu_{4,n_2} - S_{n_2}^4$
$\hat{a}_{n_1,n_2}^2 = \frac{n_1 n_2}{n_1 T_{n_2}^2 + n_2 T_{n_1}^2}$	$\hat{b}_{n_1,n_2}^2 = \frac{n_1 n_2}{n_1 S_{n_2}^2 + n_2 S_{n_1}^2}$
$f = \frac{\left(\frac{S_n^2}{n} + \frac{S_m^2}{m}\right)^2}{\frac{1}{n-1}\left(\frac{S_n^2}{n}\right)^2 + \frac{1}{m-1}\left(\frac{S_m^2}{m}\right)^2}$	

As well, we have the following quantiles

Z_α	quantile of the standard normal law
$t_\alpha(n)$	quantile of the Student law with n as $d.f$
$\chi_\alpha(n)$	quantile of the Chi-square law with n as $d.f$
$f_\alpha(n_1, n_2)$	quantile of the Fisher law with n_1 and n_2 as $d.f$

Here are the related formulas for the

4.1. One sample problem

Confidence interval of m	Inferior bound	Superior bound
Gaussian Data	$\overline{X}_n - \frac{S_n \ t_{1-\alpha/2}(n-1)}{\sqrt{n}}$	$\overline{X}_n + \frac{S_n \ t_{1-\alpha/2}(n-1)}{\sqrt{n}}$
General case	$\overline{X}_n - \frac{S_n \ z_{1-\alpha/2}}{\sqrt{n}}$	$\overline{X}_n + \frac{S_n \ z_{1-\alpha/2}}{\sqrt{n}}$
Confidence interval of σ^2	**Inferior bound**	**Superior bound**
Gaussian Data	$\frac{(n-1)S_n^2}{\chi_{\alpha/2}(n-1)}$	$\frac{(n-1)S_n^2}{\chi_{1-\alpha/2}(n-1)}$
General case	$S_n^2 - \frac{T_n \ z_{1-\alpha/2}}{\sqrt{n}}$	$S_n^2 + \frac{T_n \ z_{1-\alpha/2}}{\sqrt{n}}$

Here are the related formulas for the

4.2. Two independent sample problem

Confidence interval of $(\sigma_1/\sigma_2)^2$	Inferior bound	Superior bound
Gaussian Data	$\frac{S_{n_1}^2}{S_{n_2}^2 f_{1-\alpha/2}(n_1-1,n_2-1)}$	$\frac{S_{n_1}^2}{S_{n_2}^2 f_{\alpha/2}(n_1-1,n_2-1)}$
General case	$\frac{S_{n_1}^2}{S_{n_2}^2} - z_{1-\alpha/2}/\hat{a}_n$	$\frac{S_{n_1}^2}{S_{n_2}^2} + z_{1-\alpha/2}/\hat{a}_n$
Confidence interval of Δm	Inferior bound	Superior bound
Gaussian Data (equal variances)	$\Delta_{n_1,n_2}(X,Y) - V_{n_1,n_2}$	$\Delta_{n_1,n_2}(X,Y) + V_{n_1,n_2}$
$\Delta_{n_1,n_2}(X,Y) = \overline{X}_n - \overline{Y}_n$	$V_{n_1,n_2} = S\ t_{1-\alpha/2}(n+m-2)\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$	
unequal of variances	$\Delta_{n_1,n_2}(X,Y) - W_{n_1,n_2}$	$\Delta_{n_1,n_2}(X,Y) + W_{n_1,n_2}$
	$W_{n_1,n_2} = S\ t_{1-\alpha/2}(f)\sqrt{\frac{S_{n_1}^2}{n_1}+\frac{S_{n_2}^2}{n_2}}$	
General case	$\Delta_{n_1,n_2}(X,Y) - P_{n_1,n_2}$	$\Delta_{n_1,n_2}(X,Y) + P_{n_1,n_2}$
	$P_{n_1,n_2} = z_{1-\alpha/2}/\hat{b}_n$	

4.3. R packages

(a) Testing the exact normality. we use the Jarque-Berra test, based on the Jarque-Berra Statistic (JBS) on data X_j, $1 \leq j \leq n$, $n \geq 1$,

$$J_n = n\left(\frac{(a_n-3)^2}{24} + \frac{b_n^2}{6}\right) \approx \chi_2^2$$

where a_n are b_n the empirical kurtosis and the skewness define by

$$\overline{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j, \quad a_n = \frac{\frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X}_n)^4}{(\frac{1}{n}\sum_{j=1}^{n}(X_j-\overline{X}_n)^2)^2} \quad and \quad b_n = \frac{\frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X}_n)^3}{(\frac{1}{n}\sum_{j=1}^{n}(X_j-\overline{X}_n)^2)^{3/2}}.$$

So, the normality accepted if the p-value $p = \mathbb{P}(\chi_2^2 > J_n)$ is equal to or exceeds the level of 5%. Our simple R function for computing p is given by the function $imhJarqueBerra(X,n)$ using the data X and its size n (see page 3596).

(b) Tests of one sample.

Our function *imhSimpleTestZ,nn,alpha1,alpha2,dg)* (see 3597) provides all results in the table in Subsection 4.1, page 3585 above.

(b) Two samples Problem.

Our function *imhTwoSamplesTest(Z1,Z2,n1,n2,alpha1,alpha2,dg)* (see 3597) provides all results in in the tablein Subsection 4.2, page 3586 above
these function are explained as well as their application in the above mentioned pages.

5. Data-driven applications and simulation

We are going to apply our both methods and the Gaussian method to a set of real-life data or to simulated data. Studies are called *scenarios*. Each scenario commented and general comments and conclusions are proposed at the end.

5.1. Simulated Gaussian data of small size (n=9)

Let us begin by comparing the two methods on Gaussian data with small sizes around $n = 9$. We generate $X = rnorm(n, 3, 2)$ with $m = 3$ and $\sigma^2 = 4$. We apply our R function *imhSimpleTest(X,n,*$\alpha 1$,$\alpha 2$*, dg)* [with $\alpha 1 = 0.05$, $\alpha 2 = 0.1$, $dg = 2$] to get the following results:

Confidence interval of m	Inferior bound	Superior bound
Gaussian Data	1.9	4.3
General case	2.1	4.1
Confidence interval of σ^2	Inferior bound	Superior bound
Gaussian Data	1.3	7.3
General case	0.54	4.5

After a significant number replications, we get similar outputs with the following regards:

(1) Generally the mean estimation is more precise for the General case.
(2) Generally the variance estimation is more precise for the General case,. However the inferior can be negative in the general case and we get an confidence interval of the form $[0,\ c]$ with $c > 0$ always less than it counterpart for the Gaussian Case.

Partial conclusion [1]. Even with weak sizes as for $n = 9$, the general case beats the Gaussian case for the mean estimation. As to the variance estimation, the general has the drawback to possibly present a negative inferior bound.

5.2. Data-driven Gaussian Data for small size (n=9)

We may remake the same application to the data-driven application with the *Sheffé data*. Let us consider $A1$ as the data in Batch 1 for machine A (first in the the data) and $B1$ as the data in Batch 1 for machine B (fourth line in the data). Comparing the means $A1$ and (B1) is performed by studying the the variable $AB = A1 - B1$. The Jarque-Bera test for AB gives: $JBS = 0.13$ and $p-value : 93.647\%$. The normality of the data is hence assumed. The results of the one sample study are summarized in the table below. Here the confidence intervals for the means use a 5%-level of significance As for the variance a 10%-level of significance is used in the Gaussian case and a 5%-level of significance is used for the general case. We apply our R function *imhSimpleTest(AB,nn,$\alpha1$,$\alpha2$, dg)* [with $nn = 9$, $\alpha 1 = 0.05$, $\alpha 2 = 0.1$, $dg = 2$]. Here is an output for one replication.

Confidence interval of m	Inferior bound	Superior bound
Gaussian Data	0.11	0.47
General case	-1.9	2
Confidence interval of σ^2	Inferior bound	Superior bound
Gaussian Data	0.028	0.16
General case	-1.9	2

In this present case, the better performance of Machine $A1$ over Machine $B1$ is validated by the marge $\Delta m \geq 0.11$ trough the Gaussian exact distribution method while the general method fails to do it. We already recommended the application of the method from $n \approx 15$.

Partial conclusion [2]. Here we are in the wrong 5%. The results for the general case present not good inferior bounds in the mean and the variance estimation. We get the following recommendation: (R1) For sizes equal to less 9, we recommend to be careful for the application of the general case because not efficient inferior bounds.

5.3. What size make the application of General method reasonable?

Form several experiments, we recommend the application of the general case for sizes from $n = 15$. Here are outputs for $X = rnorm(n, 3, 2)$ with $n = 29$, $m = 3$ and $\sigma^2 = 4$, *imhSimpleTest(X,n,$\alpha1$,$\alpha2$, dg)* [with $\alpha 1 = 0.05$, $\alpha 2 = 0.1$, $dg = 2$] to get the following results:

Confidence interval of m	Inferior bound	Superior bound
Gaussian Data	2.2	4
General case	2.3	3.9
Confidence interval of σ^2	Inferior bound	Superior bound
Gaussian Data	2.2	6.5
General case	2	4.9

Partial conclusion [3]. All results for the General case are . Still, the inferior bound for the variance estimation, evecn it is correct, is a little underestimated compared to the Gaussian case.

5.4. Application to Senegalese data

We apply the methods to randomly picked data of sizs $n = 50$ from *dakar1*, *dakar2*, *diour1*, *diour2*. The Jarque-Berra test gives almost null p-values ($< 10^{-9}$) for all of them. SInce the sizes are big enough (nn=50), the *QQ-plots* are efficient to check the nor normality. Ces *QQ-plots* can be found at page 3595.

We have the following outputs for *dakar1* then for *dakar2*:

For *dakar1*:

Confidence interval of m	Inferior bound	Superior bound
Gaussian Data	190746	786611
General case	200000	779256
Confidence interval of σ^2	Inferior bound	Superior bound
Gaussian Data	8.110^{11}	1.610^{12}
General case	1.110^{12}	$1.1.10^{11}$

For *dakar2*.

Confidence interval of m	Inferior bound	Superior bound
Gaussian Data	542666	990509
General case	548194	984981
Confidence interval of σ^2	Inferior bound	Superior bound
Gaussian Data	4.610^{11}	910^{11}
General case	6.210^{11}	6.210^{12}

Partial conclusion [4]. In this non-Gaussian analysis, the confidence intervals for the mean have smaller amplitudes. The results are thus more precise, making the methods we introduced reliable.

5.5. The two samples problem

Here we expected to see the same conclusion as in the general case. We will see how things really are after some scenarii: We apply our function *imhTwoSamplesTest(Z1,Z2,n1,n2,*$\alpha 1$*,*$\alpha 2$*, dg)* [with $\alpha 1 = 0.05$, $\alpha 2 = 0.1$, $dg = 2$ to study the estimation of Δm.

(a) Comparison between performances of $A1$ and $B1$ in Scheffé's data.

Here, we use $nn = 9$, $Z1 = A1$, $Z2 = B1$, $dg = 4$. We get

Confidence interval of $(\sigma_1/\sigma_2)^2$	Inferior bound	Superior bound
Gaussian Data	0.34	4.4
General case	1.254	1.285
Confidence interval of Δm	Inferior bound	Superior bound
Gaussian Data		
(equal variances)	0.093	0.48
unequal of variances	0.093	0.48
General case	0.11	0.46

(b) Comparison between Gaussian Data for $n1$ and $n2$ small.

Here, we use $nn = 9$, $Z1 = rnorm(nn, 4, \sqrt{6})$, $Z2 = rnorm(nn, 1, \sqrt{2})$, $dg = 4$. We get

Confidence interval of $(\sigma_1/\sigma_2)^2$	Inferior bound	Superior bound
Gaussian Data	0.33	4.32
General case	-0.57	3.034
Confidence interval of Δm	Inferior bound	Superior bound
Gaussian Data		
(equal variances)	0.76	4.26
unequal of variances	0.76	4.26
General case	0.89	4.13

(c) Comparison between Gaussian Data for $n1 \geq 15$ and $n2 \geq 15$ small.

We only change $nn = 15$ in the precedent case.

Confidence interval of $(\sigma_1/\sigma_2)^2$	Inferior bound	Superior bound
Gaussian Data	1.35	8.69
General case	-0.91	7.82
Confidence interval of Δm	Inferior bound	Superior bound
Gaussian Data		
(equal variances)	$1, 63$	4.25
unequal of variances	1.63	4.25
General case	1.23	4.18

Partial conclusion [5]. From scenarii (a), (b) and (c) above, we have three striking remarks about the general case, two of them being positive and the other negative. (i) The estimation of Δm is always more precise with the general method. (ii) The estimation through the general case does not care about the equality of the variances. (iii) However, the inferior bound for the estimation of the variance ratio can take the value zero. At the same time, the superior bound is less than in the case of Gaussian data. Since, the estimation of Δm does not depend of the variances equality, this does not affect the efficiency in the method.

(d) Comparison between non-Gaussian data in Esam Senegalese data.

(d1) Evolution of area's income between two periods.

We use nn=50, imhTwoSamplesTest(dak2E,dak1E,nn,nn,0.05,0.1,4).

Confidence interval of Δm	Inferior bound	Superior bound
Gaussian Data		
(equal variances)	-291309	164352
unequal of variances	-291309	164352
General case	-288496	161354

(d2) Geographical discrepancy income between Dakar and Diourbel in 1996

Confidence interval of Δm	Inferior bound	Superior bound
Gaussian Data		
(equal variances)	214102	587024
unequal of variances	214102	587024
General case	216405	$584v722$

(d3) Geographical discrepancy income income between in 2000

Confidence interval of Δm	Inferior bound	Superior bound
Gaussian Data		
(equal variances)	157976	438435
unequal of variances	157976	438435
General case	159708	436705

Partial conclusion [6]. In scenarii (d1), (d2) and (d3), we applied our method to income Senegalese data that are not Gaussian. In general, as already noticed, the confidence intervals for estimated Δm are more precise with smaller amplitude. By using them, say that:

(i) the income does not change from 1996 to 2000. We see this for Dakar. We did not report the corresponding outputs for Diourbel since the conclusion is still the same.

(ii) For each period, the income is far higher in Dakar than in Diourbel with significant thresholds of 216405 in 1996 and 159708. The good thing is that the discrepancy diminished by $26,19\%$.

6. Conclusion

The new method, called the General Samples Problem, is based on asymptotic methods, specially on the striking results of the *functional empirical process (fep)*. Almost for samples' sizes of $n = 10$, the results are almost equivalent to those in the Gaussian case. For greater sizes, the estimation of the mean difference is very precise regardless of the equality or inequality of variances. We applied the results to Senegalese data which are not Gaussian as checked by the use of Jarque-Bera test. We recommend that method for sizes around or greater that $n = 15$. However, the method requires the finiteness of the second order moment. We have to be cautious about this when dealing with heavy tails.

Acknowledgment. The authors are indebted to Professor Gane Samb Lo, who made significant contributions to this paper, despite not being listed as a co-author. We also extend our appreciation to the scientific team of the project for their insightful remarks, editing, and critiques, which have notably enhanced the quality of this paper. Finally, we would like to thank the editorial team of the journal for their excellent work in producing the final version of this paper.

APPENDIX DATA

Data from Scheffé (1959), page. 290

Machine	D1	D2	D3	D4	D5	D6	D7	D8	D9
A	1.4	1.45	1.91	1.89	1.77	1.66	1.92	1.84	1.54
A	1.35	1.57	1.48	1.48	1.73	1.54	1.93	1.79	1.43
A	1.62	1.82	1.89	1.39	1.54	1.68	2.13	2.04	1.7
B	1.31	1.24	1.51	1.67	1.23	1.4	1.23	1.58	1.64
B	1.63	1.18	1.58	1.37	1.4	1.45	1.51	1.63	1.07
B	1.41	1.52	1.65	1.11	1.53	1.63	1.44	1.28	1.38
C	1.93	1.43	1.38	1.72	1.32	1.63	1.33	1.69	1.7
C	1.67	1.77	1.69	1.53	1.49	1.61	1.8	2.25	1.37
C	1.4	1.86	1.36	1.37	1.34	1.36	1.38	1.8	1.84

Income Data of Dakar 1.

212860.9	267010.7	258080.4	578540.3	244696.2	215546.2
290591.1	304817.6	291419.1	440473.7	138962.8	145774.7
261535.4	1182688.3	231771.8	108884.2	583703.4	340638.7
143220.2	361730.6	488192.8	232069.2	599704.8	194193.3
320658.9	167081.5	236670.5	1027139.7	226607.0	567213.9
714605.4	107189.7	160302.3	363402.8	761425.0	466297.9
179790.5	179206.9	7591873.8	257280.9	247732.1	271506.2
242673.6	495594.4	204820.5	182907.2	314469.1	390925.7
388434.4	251026.1				

Income Data of Dakar 2.

2326347.1	202446.1	497640.6	1505369.5	206145.4	670007.1
688580.6	3800902.5	456837.8	772329.3	1268005.4	944911.9
579565.7	345986.3	605220.4	275933.6	664973.9	345986.3
1443230.0	3111786.7	190088.7	400339.4	1440574.8	456007.6
204802.6	400339.4	901053.5	184492.8	900889.4	324556.9
612592.7	603949.7	705670.0	255847.5	711869.7	2385253.6
129207.5	354822.7	175583.9	206145.4	532207.0	224469.6
364105.7	609470.4	751249.5	129207.5 2	441979.0	328765.1
374895.0	316734.3				

Income Data of Diourbel 1.

165087.11	191031.56	123808.57	396346.90	82781.90	519556.29
146485.27	144961.70	122684.94	146399.19	349398.28	153086.53
292284.34	160460.77	194542.64	82656.72	214976.94	151839.94
160460.77	104743.29	662028.92	132762.82	91387.36	211374.42
135755.33	144137.22	75249.22	87559.66	163498.36	89995.23
125835.76	300345.97	257382.04	215629.97	109558.28	214976.94
87289.32	233952.60	104693.64	197164.40	343203.01	385182.29
82781.90	658761.83	215069.65	117829.49	132192.84	95234.29
216132.15	396346.90				

Income Data of Diourbel 2.

339293.85	199794.97	167231.82	317826.88	473868.40	440738.61
150857.01	205082.61	480722.21	160956.28	313255.46	267165.79
205322.67	268027.99	78512.00	242960.37	152158.57	182687.95
396531.52	156383.34	205082.61	393941.15	810344.85	99643.59
107577.92	242895.53	299596.63	138594.52	205852.70	94937.99
190404.38	317826.88	97794.50	192905.57	167724.34	777178.29
190377.06	366065.69	152322.68	186260.20	474140.29	222936.34
325813.46	138305.38	514067.87	77420.32	211870.89	189376.45
223896.62	345868.96				

qq-plots against normal law for data from ESAM.

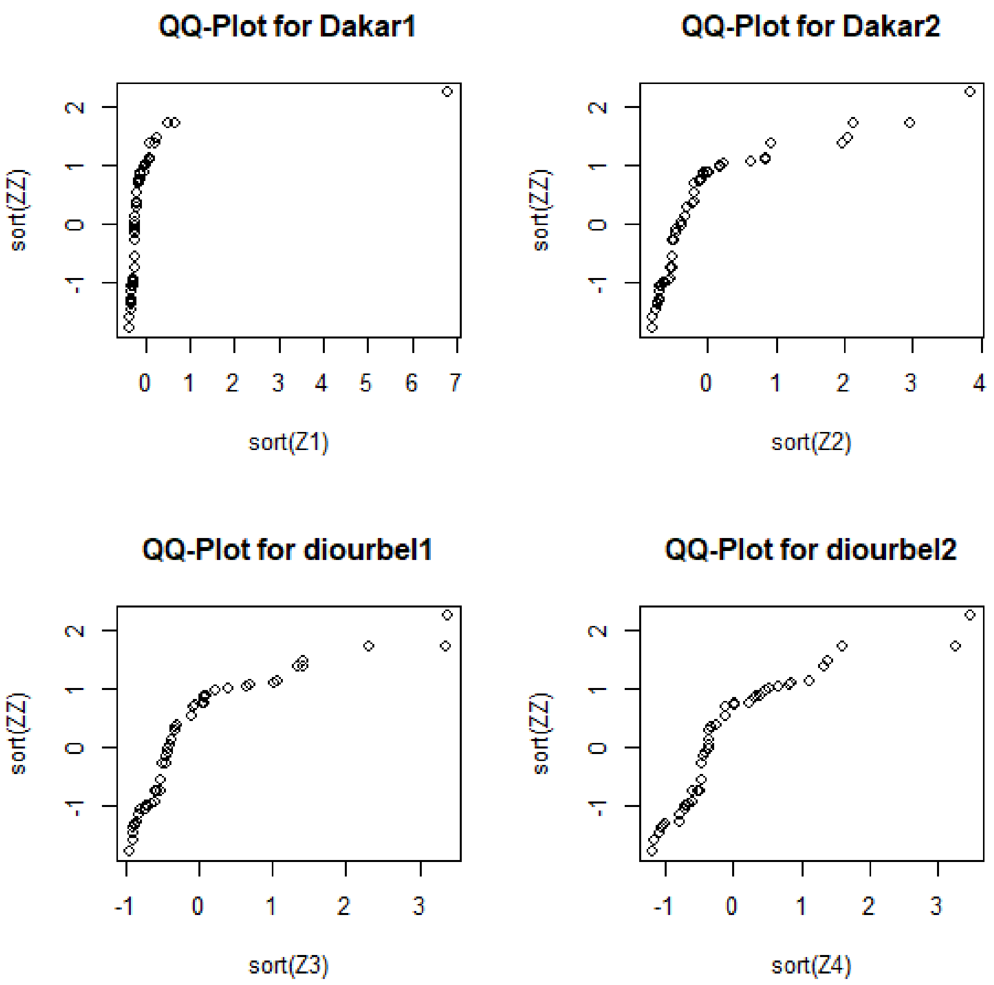

Figure 1. QQ-plots of data from ESAM against Normal law

APPENDIX PACKAGES

1. Normality Test of Jarque-Berra

```
# Ce package requires the data vector V of size at least one
# the size TT of the data
# Sortie: display skewness, kurtosis, jarque-Bera Statistics
# and pv-alue of the test.
# At he first running, initialize as follows
TT=15
V <- rnorm(TT,0,1)

imhJarqueBerra <- function(V, XX){
        va=sqrt(mean((V-mean(V))^2))
        va
        a0n=(mean((V-mean(V))^4)/(va^4))
        an=(a0n-3)^2
        b0n=(mean((V-mean(V))^3)/(va^3))
        bn=b0n^2
        jbt=taille*((an/24)+(bn/6))
        ret=numeric(4)
        ret[1]=paste("skewness:", b0n)
        ret[2]=paste("kurtosis:", a0n)
        ret[3]=paste("Jarque Berra Statistic:", jbt)
        ret[4]=paste("pa-value: ", format(100*(1-pchisq(jbt,2)),digits=5),"%
        return(noquote(ret))
}
# Apply the function as : imhJarqueBerra(X,nn)
#Thar gives the p-values for data X of size nn
```

2. A display function

```
# This function display the the number dg of digits of string zz
# with  no quotes. The limitation of the number of digits
# works only of zz is originally a number
(zz=12.2213)
imhAff2 <- function(zz,dg){
         aff=noquote(format(zz,digits=dg))
                 return(aff)
}
 #Apply the function as : imhAff2(335.98673,5)
```

2. One sample problem

```
# This function has as arguments: a data set Z of size nn,
# alpha1: seuil of significance of estimating means or ther differences
# alpha2: seuil of significance of testing variances of their rations
# dg=2: number of digits for the display of the results
# Initialize as follows:
nn=30
m=2
s=4
Z <- rnorm(nn,m,s)
alpha1=0.05
alpha2=.1

imhSimpleTest <- function(Z,nn,alpha1,alpha2,dg){
guenne=c(1:14)
(tc=qt(1-alpha1/2,nn-1))
(uc1=qnorm(1-alpha1/2,0,1))
(uc2=qnorm(1-alpha2/2,0,1))
moy=mean(Z)
sigma=sd(Z)
amp=sigma*tc/sqrt(nn)
ampGm=sigma*uc1/sqrt(nn)
minf=moy-amp
msup=moy+amp
mginf=moy-ampGm
mgsup=moy+ampGm
mu4= mean((Z-mean(Z))^4)
(TnC=mu4-(sigma^4))
ampGsigma=(sigma^2)*uc1/sqrt(TnC)
## Case general : alpha=0.05
sigInfG=(sigma^2)-ampGsigma
sigSupG=(sigma^2)+ampGsigma

(sigInf=(nn-1)*(sigma^2)/qchisq(1-alpha2/2,nn-1))
(sigSup=(nn-1)*(sigma^2)/qchisq(alpha2/2,nn-1))


guenne[1]="======================"
guenne[2]=paste('Estimation of the mean')
guenne[3]=paste('Gaussian case')
guenne[4]=paste(imhAff2('Mean:',dg),imhAff2(moy,dg),imhAff2('/alpha',dg),
imhAff2(alpha1,dg),imhAff2('/BiCI:',dg),imhAff2(minf,dg),
imhAff2('/BsCI:',dg),imhAff2(msup,dg))

guenne[5]="======================"
```

```
guenne[6]=paste('General Case')
guenne[7]=paste(imhAff2('Mean:',dg),imhAff2(moy,dg),
imhAff2('/alpha',dg),imhAff2(alpha1,dg),imhAff2('/BiCI:',dg),
imhAff2(mginf,dg), imhAff2('/BsCI:',dg),imhAff2(mgsup,dg))

guenne[8]="======================"
guenne[9]=paste('Estimation of the variance')
guenne[10]=paste('Gaussian case')
guenne[11]=paste(imhAff2('Variance:',dg),imhAff2(sigma^2,dg),
imhAff2('/alpha',dg), imhAff2(alpha2,dg),imhAff2('/BiCI:',dg),
imhAff2(sigInf,dg),imhAff2('/BsCI:',dg),imhAff2(sigSup,dg))

guenne[12]="======================"
guenne[13]=paste('General Case')
guenne[14]=paste(imhAff2('Variance:',dg),imhAff2(sigma^2,dg),
imhAff2('/alpha',dg),imhAff2(alpha1,dg),imhAff2('/BiCI:',dg),
imhAff2(sigInfG,dg),imhAff2('/BsCI:',dg),imhAff2(sigSupG,dg))

return(noquote(guenne))
}

#Finally, apply the function as:
# imhSimpleTestZ,nn,alpha1,alpha2,dg)
```

2. Two samples problem

```
# This function has as arguments: two data sets Z1 and Z2
#of size n1 and n2,
# alpha1: seuil of significance of estimating means
#or ther differences
# alpha2: seuil of significance of testing variances
# of their rations
# dg=2: number of digits for the display of the results
# Initialize as follows:
#======
n1=25
n2=25
m1=2
m2=5
sigma1C=3
sigma2C=7
Z1=rnorm(n1, m1, sigma1C)
Z2=rnorm(n2, m2, sigma1C)
Dm=m1-m2

imhTwoSamplesTest <- function(Z1,Z2,n1,n2,alpha1,alpha2,dg){
guenne=c(1:16)
Dmoy=mean(Z1)-mean(Z2)
sigma1EC=var(Z1)
sigma2EC=var(Z2)
RV=(sigma1C/sigma2C)
RVE=(sigma1EC/sigma2EC)
de=((sigma1EC/n1)+(sigma2EC/n2))^2
num=((sigma1EC/n1)^2/(n1-1))+((sigma2EC/n2)^2/(n2-1))
(fs=de/num)
(tcf=qt(1-alpha1/2,fs))
(uc1=qnorm(1-alpha1/2,0,1))
mu4E1= mean((Z1-mean(Z1))^4)
mu4E2= mean((Z2-mean(Z2))^4)
T1C=mu4E1-(sigma1EC^2)
T2C=mu4E2-(sigma2EC^2)
achap=sqrt((n1*n2)/((n1*T2C)+(n2*T1C)))
bchap=sqrt(n1*n2/((n1*sigma2EC)+(n2*sigma1EC)))

#sigmaRatio
srInf=(sigma1EC/sigma2EC)/qf(1-alpha2/2,n1-1,n2-2)
srSup=(sigma1EC/sigma2EC)/qf(alpha2/2,n1-1,n2-2)
srInfG=(sigma1EC/sigma2EC) - (uc1/achap)
srSupG=(sigma1EC/sigma2EC) + (uc1/achap)
(sigma1EC/sigma2EC)
```

```
#sqrt(((T1C/n1)+(T2C/n2)))
(uc1/achap)
uc1*achap
#Global S
gsc=((n1-1)*sigma1EC+(n2-1)*sigma2EC)/(n1+n2-2)
gs=sqrt(gsc)

quotEV= sqrt((1/n1)+(1/n2))
quotUV= sqrt((sigma1EC/n1)+(sigma2EC/n2))

DiffAmpN= gs*qt(1-alpha1/2,n1+n2-2)*quotEV
biN=Dmoy-DiffAmpN
bsN=Dmoy+DiffAmpN

DiffAmpAp= qt(1-alpha1/2,n1+n2-2)*quotUV
biA=Dmoy-DiffAmpAp
bsA=Dmoy+DiffAmpAp

DiffAmpG= uc1/bchap
biG=Dmoy-DiffAmpG
bsG=Dmoy+DiffAmpG

guenne[1]="======================"
guenne[2]=paste('Estimation of the ratio of variance')
guenne[3]=paste('Gaussian case')
guenne[4]=paste(imhAff2('Ratio:',dg),imhAff2('/alpha',dg),
imhAff2(alpha2,dg),
imhAff2('/BiCI:',dg),imhAff2(srInf,dg),imhAff2('/BsCI:',dg),
imhAff2(srSup,dg))

guenne[5]="======================"
guenne[6]=paste('General Case')
guenne[7]=paste(imhAff2('Ratio:',dg),imhAff2('/alpha',dg),
imhAff2(alpha1,dg),imhAff2('/BiCI:',dg),imhAff2(srInfG,dg),
imhAff2('/BsCI:',dg), imhAff2(srSupG,dg))

guenne[8]="======================"
guenne[9]=paste('Estimation of the mean difference')
guenne[10]=paste('Gaussian case with equality of variances')
guenne[11]=paste(imhAff2('Diff:',dg),imhAff2(Dm,dg),
imhAff2('/alpha',dg), imhAff2(alpha1,dg),imhAff2('/BiCI:',dg),
imhAff2(biN,dg), imhAff2('/BsCI:',dg),imhAff2(bsN,dg))

guenne[12]="======================"
guenne[13]=paste('Gaussian case with unequality of variances')
guenne[14]=paste(imhAff2('Diff:',dg),imhAff2('/alpha',dg),
```

```
imhAff2(alpha1,dg),  imhAff2('/BiCI:',dg),imhAff2(biA,dg),
imhAff2('/BsCI:',dg),imhAff2(bsA,dg))

guenne[15]="======================"
guenne[16]=paste('General Case')
guenne[17]=paste(imhAff2('Diff:',dg),imhAff2('/alpha',dg),
imhAff2(alpha1,dg),  imhAff2('/BiCI:',dg),imhAff2(biG,dg),
imhAff2('/BsCI:',dg),imhAff2(bsG,dg))
         return(noquote(guenne))
}

# Use the function as :
# imhTwoSamplesTest(Z1,Z2,n1,n2,alpha1,alpha2,dg)
```

References

Lo G.S., Kpanzou T.A. and Da B.O.G. (2023). General asymptotic representations of indexes based on the functional empirical process and the residual functional empirical process and applications. *Afrika Statistika*, Vol 18, (3), pp. 3521-3550. Doi:http://dx.doi.org/10.16929/as/3521.2023.317.ssFEPLREP.

Scheffé H. (1959) The Analysis of Variance. Wiley. ISBN 0 471 75834 5

Seber G.A.F. and Lee A.J (2003). *Introduction to Linear Regression*. Wiley

T. Hastie T., Tibshirani R. and Friedman J. (221). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd Edition, Springer