A new estimator Median of the distribution of the mean in robustness

Alfonso García-Pérez

Versión final publicada en ***Mathematics*, 2023, vol. 11 (12): 2694.**

https://www.mdpi.com/2227-7390/11/12/2694

Departamento de Estadística, I.O. y C.N., Universidad Nacional de Educación a Distancia (UNED),

Paseo Senda del Rey 9, 28040 Madrid, Spain; agar-per@ccia.uned.es

Abstract: In some statistical methods, the statistical information is provided in terms of the values used by classical estimators, such as the sample mean and sample variance. These estimations are used in a second stage, usually in a classical manner, to be combined into a single value, as a weighted mean. Moreover, in many applied studies, the results are given in these terms, i.e., as summary data. In all of these cases, the individual observations are unknown; therefore, computing the usual robustness estimators with them to replace classical non-robust estimations by robust ones is not possible. In this paper, the use of the median of the distribution Fx of the sample mean is proposed, assuming a location-scale contaminated normal model, where the parameters of Fx are estimated with the classical estimations provided in the first stage. The estimator so defined is called median of the distribution of the mean, MdM. This new estimator is applied in Mendelian randomization, defining the new robust inverse weighted estimator, RIVW.

Keywords: robust statistics; von Mises expansions; saddlepoint approximations; Mendelian randomization

MSC: 62F35; 62E17; 62P99

1. Introduction

In the application of some statistical methods, such as clinical trials, the results are, usually, described in terms of the values taken by classical estimators, such as the sample mean and sample variance. These results are combined, in a second stage, as a weighted mean in a meta analysis. The same occurs in its alternative, Mendelian Randomization, one of the main topics in causal inference.

Moreover, in many applied studies, their results have been described in these terms, i.e., as summary data, not knowing the individual observations, to compute robust estimators with them, replacing the classical non-robust estimations with robust ones.

In this paper, a solution to this problem is proposed, correcting, if necessary, the given classical estimations because, although the individual observations are unknown, the mechanism that generates the data is known because it is the model.

Focusing on the mean estimation problem, the optimal estimator (uniformly minimum variance unbiased estimator) is the sample mean, when no outliers exist in the sample, and the normal distribution N(μ, σ2) is assumed as the model, with μ and σ2 being the usual parameters of the normal distribution, population mean, and variance. Assume that a proportion e of outliers exists in the sample, i.e., a contaminated normal model (see [1], p. 2)

(1 − e)N(μ, σ2) + eN(g1μ, g22σ2)

where most of the data are from a N(μ, σ2), and a small part of them, , are from a normal model with more dispersion and a different location, N(g1μ, g22σ2), where g1 is a contamination parameter that affects the location, and g2 is a contamination parameter that affects the scale. The optimality of the sample mean is lost because the optimal procedure

and its properties heavily depend on the assumed probability model ([2], p. 2). This is the reason why classical statistics rests, basically, on the normal model and on the sample mean.

Additionally, under a contaminated normal model, the robustness of the sample mean is lost [1,3]. Under this model, the sample mean is not the maximum likelihood estimator [4], and even the normality of the sample mean is not guaranteed [5].

In this paper, a new estimator for a location–scale contaminated normal model is proposed, avoiding the extreme sensitivity of the sample mean but coinciding with it when no outliers are present in the data. The median of the distribution $F_{\overline{x}}$ of the sample mean is proposed as a new estimator, where the parameters of $F_{\overline{x}}$ are estimated with the classical estimations described in previous studies. This estimator is called the *median of the distribution of the mean*, *MdM*.

The two reasons why this new estimator relies on the distribution of the sample mean are that, first, the classical estimations are given in terms of the classical mean (and classical variance) and, second, this new procedure extends the classical one in the sense that if no outliers are present, this new estimator is the classical sample mean, i.e., with this method, the classical estimation is extended to the case in which outliers are present.

Another estimator somewhat related to *MdM* is the *median of the means* estimator *MoM*. However, this estimator is, finally, one of the sample means and, hence, is not robust (see [6]).

With the *MdM*, robustness and optimality are obtained if there are no outliers. Hence, with this approach, a new vision of the dilemma between optimality and robustness is provided.

Because the exact sample distribution of $\overline{x}$ under a mixture distribution is not known, here it is estimated in a closed form with the von Mises (VOM) plus saddlepoint (SAD) method, a technique used by the author in several studies (see, for instance, [7,8]) but in another context. With this approximation, the estimator introduced in this paper can also be extended to other more general models than the normal mixture considered here.

The rest of the paper is structured as follows. In Section 2, the VOM+SAD approximation for the distribution of the sample mean is obtained under a location–scale contaminated normal model. The definition and some properties of this new location estimator are considered in Section 3, and a scale estimator, based on these ideas, is defined in Section 4, and an example of the application of this new estimator is considered in Section 5. These ideas are applied to Mendelian randomization in Section 6. Some conclusions are outlined in Section 7.

2. VOM+SAD Approximation of Sample Mean Distribution

Because the new estimator depends on the distribution of the sample mean, the distribution of the sample mean must be very precise, especially when the considered sample sizes are very small. For this situation, using a von Mises expansion ([9], p. 215, or [10], p. 578) that depends on Hampel's influence function [11] is highly recommended.

Although, in the end, the obtained results are be applied to the mixture of normals model considered previously, these refer to more general models, F, G, and H, which indicate future extensions of this method.

The final approximation is called VOM+SAD and was previously obtained by the author in the context of spatial data (see [7,8]). Following the ideas developed in those two papers, considering the tail probability functional, initially, the approximation obtained is

$$P_F\{T_n > t\} \simeq P_G\{T_n > t\} + \int \text{TAIF}(x; t; T_n, G)\, dF(x)$$

which allows the approximation of the distribution of T_n when the observations follow model F by the distribution of T_n when the variables of T_n follow model G (pivotal distribution).

This approximation depends on the tail area influence function, TAIF, defined in [12].

Restricting this approximation to M estimators with a monotonic decreasing score function ψ (see [1], p. 46) and using the Lugannani and Rice formula ([13], or [14] p. 77,

or [1] p. 314) to obtain a saddlepoint approximation for the TAIF, as the approximation given in [15] (p. 94), for M estimators, the VOM+SAD approximation obtained is

$$P_F\{T_n > t\} \simeq P_G\{T_n > t\} + \int \frac{\phi(s)}{r_1}\, n^{1/2} \left(\frac{e^{z_0 \psi(x,t)}}{\int e^{z_0 \psi(y,t)} dG(y)} - 1 \right) dF(x).$$

In the case of a location–scale mixture normal model, the framework that it is considered in this paper, i.e., assuming that $Z_i \equiv (1-\epsilon)N(\mu, \sigma^2) + \epsilon N(g_1\mu, g_2^2\sigma^2)$, the VOM+SAD approximation is

$$P_F\{T_n > t\} \simeq P_G\{T_n > t\} + \epsilon\, \frac{\phi(s)}{r_1}\, \sqrt{n} \left(\frac{\int e^{z_0 \psi(x,t)} dH(x)}{\int e^{z_0 \psi(y,t)} dG(y)} - 1 \right) \tag{1}$$

where $G = N(\mu, \sigma^2)$, and $H = N(g_1\mu, g_2^2\sigma^2)$.

VOM-SAD Approximation for the Distribution of the Sample Mean

In the particular case of the sample mean, the score function is $\psi(x,t) = x - t$. Remember that in the VOM+SAD approximation, the saddlepoint is computed under $G = N(\mu, \sigma^2)$. Under this pivotal distribution, it is

$$K(\lambda, t) = \log \int e^{\lambda(y-t)} \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2\sigma^2}(y-\mu)^2}\, dy = \frac{\sigma^2\lambda^2}{2} + \lambda(\mu - t).$$

Hence, from the saddlepoint equation $K'(z_0, t) = 0$ the saddlepoint $z_0 = (t-\mu)/\sigma^2$ is obtained.

Additionally, $K(z_0, t) = -(t-\mu)^2/(2\sigma^2)$, $\phi(s) = \phi(\sqrt{n}(t-\mu)/\sigma)$, $r_1 = (t-\mu)/\sigma$, and $K''(\lambda, t) = \sigma^2$. The leading term is $P_G\{T_n > t\} = 1 - \Phi(\sqrt{n}(t-\mu)/\sigma)$, and the quotient in last term in the right side of (1) is

$$\frac{\int e^{z_0 \psi(x,t)} dH(x)}{\int e^{z_0 \psi(y,t)} dG(y)} = \exp\left\{ (g_1 - 1)\mu\, z_0 + \frac{1}{2}(g_2^2 - 1)\sigma\, z_0^2 \right\}.$$

Hence, the VOM+SAD approximation (1) is

$$P_F\{\overline{x} > t\} \simeq 1 - \Phi(\sqrt{n}\,\sigma\, z_0) + \epsilon\, \frac{\sqrt{n}}{\sigma\, z_0}\, \phi(\sqrt{n}\,\sigma\, z_0) \left(e^{(g_1-1)\mu\, z_0 + \frac{1}{2}(g_2^2-1)\sigma\, z_0^2} - 1 \right).$$

If distributions F and G are not close enough, intermediate distributions can be considered, as in [16–18], to obtain a more accurate approximation.

3. Estimator Median of the Distribution of the Mean

If the previous distribution of the mean is

$$F_{\overline{x}}(x) = 1 - P_F\{\overline{x} > x\}$$

the median of this distribution, i.e., $F_{\overline{x}}^{-1}(1/2)$, is called *the median of the distribution of the mean*, MdM, i.e., this estimator is the solution of

$$F_{\overline{x}}(MdM) = \frac{1}{2}.$$

The parameters of $F_{\overline{x}}$ are estimated with the classical estimations, the sample mean $\overline{x}$ and the sample variance s^2.

Figures 1–3 show that as contamination parameters ϵ, g_1, or g_2 increase, the difference between MdM and $\overline{x}$, i.e., z_0, increases.

The main reason for the definition of MdM is that the median is more robust than the sample mean and, hence, the influence of possible outliers, not knowing the individual

observations, as assumed here, should be lower with the median of the distribution of the mean than with the sample mean, used in this distribution as an estimator of the location parameter. Furthermore, in the case without outliers, this estimator is equal to the classical sample mean.

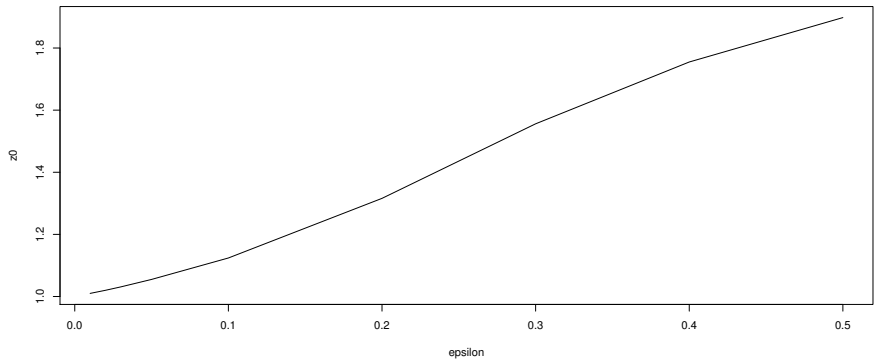

Figure 1. Differences between MdM and $\overline{x}$ as ϵ increases.

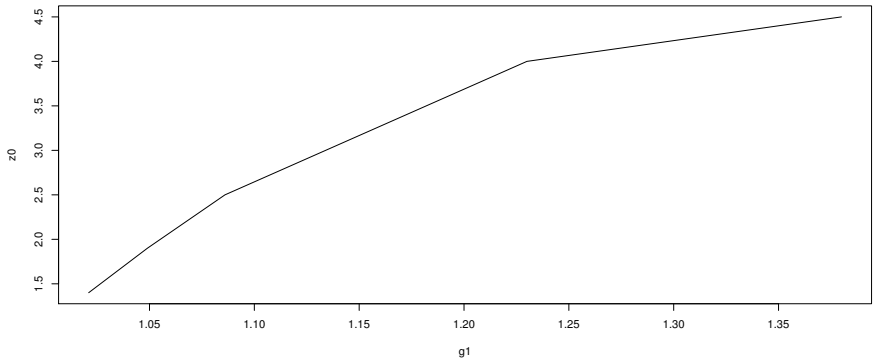

Figure 2. Differences between MdM and $\overline{x}$ as g_1 increases.

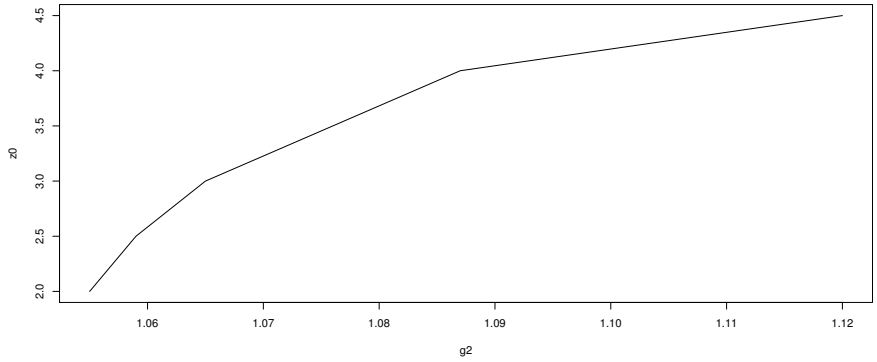

Figure 3. Differences between MdM and $\overline{x}$ as g_2 increases.

As a limitation, observe that MdM is also sensitive if outliers already affect the sample mean or sample variance used in the estimation of the location or scale parameter μ or σ^2. Nevertheless, with MdM, this sensitivity is lower.

One way to check the behavior of MdM with respect to $\overline{x}$ in a simple numerical example is to run the R sentences

```
> x<-0.80*rnorm(11,2,1)+0.2*rnorm(11,3*2,1)
> mean(x)
> median(x)
```

in which we consider a random sample of $n = 11$ sample data from a mixture normal $0.8N(2,1) + 0.2N(3 \cdot 2, 1)$, i.e., a sample where $\epsilon = 0.2, g_1 = 3$ and $g_2 = 1$.

Finally, in future research, other robust estimators could be considered, such as the trimmed mean of the distribution of the sample mean.

4. Dispersion Estimator

With the ideas developed in this paper, a dispersion estimator should be

$$F_{\overline{x}}^{-1}(3/4) - F_{\overline{x}}^{-1}(1/4).$$

5. Example

In most application papers, only the final values of the estimators used on them are given. Additionally, these estimators are usually the classical sample mean and sample variance and do not include the individual observations from which these estimators are obtained and, therefore, not providing the opportunity to robustify these values using robust techniques.

For this reason, a large number of examples could serve as an illustration of the estimator defined in this paper. Next, let us consider just one.

Example 1. *One of these studies is [19], where some vertebral column and thorax of Neanderthals fossils were re-evaluated using their vertebrae because, probably, as stated by the author, errors occurred in the reconstruction and the samples were wrongly classified. He mentions ([19], p. 23) a misclassification of 7/33, which can be considered as the value of the contamination parameter ϵ.*

Because modern humans and Neanderthals have very similar vertebrae, no difference in the mean is assumed, using, hence a distortion factor $g_1 = 1$. On the other hand, Neanderthals are slightly more stockier than modern humans, with the dispersion of the latter being larger, assuming that $g_2 = 1.5$.

In Table 2 in [19], classical acceptance confidence intervals are provided for several vertebrae of 28 modern humans. They are based on the classical mean and variance, as the author says in this table. From the table, with respect vertebra **T1**, *the remains of Kebara 2 and La Ferrassie can be considered as modern humans instead of Neanderthals, because they are inside of the confidence interval. The same happens with vertebra* **T7** *but not with vertebra* **T5**.

From these classical intervals, for vertebra **T1**, *the classical sample mean and standard deviation are $\overline{x} = 16.6$ and $S = 3.61$, respectively. In this case, the estimator median of the distribution of the mean takes the value $MdM = 15.034$, obtaining the new robust acceptance confidence interval equal to $[13.63\,,\,16.43]$, which does not contain the remains, concluding then, that these remains are Neanderthals and not modern humans, as they were wrongly considered with the classical estimators.*

With respect to vertebra **T5**, *$MdM = 17.54$, and the new robust acceptance confidence interval is $[15.84\,,\,19.24]$, with neither the classical nor the robust interval not including the remains, confirming that they are Neanderthals.*

Finally, for vertebra **T7**, *$MdM = 19.43$, and the new robust acceptance confidence interval $[17.73\,,\,21.13]$, both this robust and the previous classical confidence interval including the remains of the La Ferrassie, hence being modern humans and not Neanderthals.*

6. Robust Inverse-Weighted Estimator RIVW in Mendelian Randomization

Another field for the class of problems considered in this paper is randomized clinical trials (CTs). In each of these CTs, the sample mean and sample variance are the usual final result. These are usually combined, in a classical way, as a weighted mean in a meta-analysis. In CTs, the relationship of a variable X (called cause) with another variable Y (called effect) is analyzed, but reverse causality may exist or a lack complete randomization or, more importantly, confounders may be present.

Moreover, CTs are expensive and take a long time. With Mendelian randomization (MR), a method that has received a renewed interest in recent years, CTs are imitated

because, in any person, all genetic material is randomized allocated from their parents, including DNA markers. Randomly, some people receive more DNA markers related with variable X and, for others, fewer. MR uses genetic variants (usually single-nucleotide polymorphisms (SNPs)) as instrumental variables Z.

Mathematically, MR is used to avoid possible biases in the regression of Y on X due to these three causes just mentioned. Formally, MR leads us to a two-step linear regression process; first, for every genetic variant $Z_j, j = 1, \ldots, L$, a linear regression of X on Z_j is performed, where, for individuals, $i = 1, \ldots, n_j$ is

$$X_i \mid Z_{ij} = \beta_{X_0} + \beta_{X_j} Z_{ij} + e_{X_{ij}}$$

from which the fitted values $\hat{X}_i$ are obtained and used in a second regression of Y on these $\hat{X}$, obtaining finally [20]

$$Y_i \mid Z_{ij} = \beta_{Y_0} + (\beta \cdot \beta_{X_j} + \alpha_j) Z_{ij} + e_{Y_{ij}} = \beta_{Y_0} + \beta_{Y_j} Z_{ij} + e_{Y_{ij}}$$

where β_{X_j} and β_{Y_j} represent the association of Z_j with the exposure and the outcome (only through X), respectively. The parameter $\beta \cdot \beta_{X_j}$ represents the effect of Z_j on Y through X, where β is the causal effect of X on Y that is being estimated. Moreover, α_j represents the association between Z_j and Y not through the exposure of interest. Finally, the errors terms $e_{X_{ij}}$ and $e_{Y_{ij}}$ are assumed to be independent because independent samples are assumed to be used to fit the two previous regression models.

In MR, the standard estimator of the parameter of interest β, the slope in the linear regression of Y on X, is the classical *two-stage least squares estimator*

$$\hat{\beta}_{R_j} = \frac{\hat{\beta}_{Y_j}}{\hat{\beta}_{X_j}}$$

which is the quotient of the slope of the regression of Y on Z_j, $\hat{\beta}_{Y_j}$, and the slope estimator of the regression of X on Z_j, $\hat{\beta}_{X_j}$. These classical estimations, one for each value of the instrumental variable Z_j, are combined with the classical inverse-variance weighted (IVW) estimator

$$\text{IVW} = \sum_{j=1}^{L} \omega_j \hat{\beta}_{R_j} / (\sum_{j=1}^{L} \omega_j)$$

where $\omega_j = 1/var(\hat{\beta}_{R_j})$, which is used to weight the $\hat{\beta}_{R_j}$ estimators, assuming that the L genetic variants are mutually independent. In this way, a single causal effect estimate from L genetic instruments is obtained.

This classic and widely used estimator is not robust because it has a 0% breakdown point because it is a weighted mean, see, for instance, [21].

In this section, the robustification of the classical estimator IVW is obtained, first, by replacing estimators $\hat{\beta}_{R_j}$ with the *median of the distribution of the mean*, MdM_j estimators and, second, by replacing the weights ω_j with v_j, the inverse of the new dispersion estimator,

$$v_j = \frac{1}{F_{\bar{x}}^{-1}(3/4) - F_{\bar{x}}^{-1}(1/4)}$$

defining the new estimator, based on the $\hat{\beta}_{R_j}$ distribution, as

$$\text{RIVW} = \frac{\sum_{j=1}^{L} v_j \, MdM_j}{\sum_{j=1}^{L} v_j}.$$

6.1. Distribution of $\hat{\beta}_{R_j}$ Estimator

In this section, an approximation for the distribution of $\hat{\beta}_{R_j}$ is obtained for each genetic variant $Z_j, j = 1, \ldots, L$, i.e, j is fixed. Moreover, because of the usual regression assumptions, Z_{ij} is not random in the two previous linear regressions, i.e., in the estimator $\hat{\beta}_{R_j}$.

Hence, with μ_{X_i} denoting the constant

$$\mu_{X_i} = \beta_{X_0} + \beta_{X_j} Z_{ij}$$

avoiding the j in the notation of μ_{X_i} to simplify it, and with μ_{Y_i} being the constant

$$\mu_{Y_i} = \beta_{Y_0} + \beta_{Y_j} Z_{ij},$$

and assuming no outliers in the sample, the variable $X_i \mid Z_{ij}$ follows a normal distribution

$$X_i \mid Z_{ij} \equiv N(\mu_{X_i}, \sigma^2_{X_i})$$

and

$$Y_i \mid Z_{ij} \equiv N(\mu_{Y_i}, \sigma^2_{Y_i}).$$

The estimator $\hat{\beta}_{R_j}$ is equal to

$$\hat{\beta}_{R_j} = \frac{\hat{\beta}_{Y_j}}{\hat{\beta}_{X_j}}$$

and, considering standardized data, i.e., that $\hat{\beta}_{R_j}$ is computed as a correlations quotient,

$$\hat{\beta}_{R_j} = \frac{\sum_{i=1}^{n_j} Y_i Z_{ij}}{\sum_{i=1}^{n_j} X_i Z_{ij}}$$

its tail distribution is

$$\begin{aligned} P\left\{\hat{\beta}_{R_j} > a\right\} &= P\left\{\frac{\sum_{i=1}^{n_j} Y_i Z_{ij}}{\sum_{i=1}^{n_j} X_i Z_{ij}} > a\right\} \\ &= P\left\{\sum_{i=1}^{n_j} Y_i Z_{ij} - a \sum_{i=1}^{n_j} X_i Z_{ij} > 0\right\} \\ &= P\left\{\sum_{i=1}^{n_j} (Y_i - aX_i) Z_{ij} > 0\right\}. \end{aligned}$$

Letting W_i (removing the j if there is no risk of confusion) denote the random variable $W_i = W_{ij} = (Y_i - aX_i)Z_{ij}$, $i = 1, \ldots, n_j$, where a and Z_{ij} are not random, the aim is to compute the distribution of the sample mean of the variables W_i at 0, i.e.,

$$P\left\{\hat{\beta}_{R_j} > a\right\} = P\left\{\sum_{i=1}^{n_j} W_i > 0\right\} = P\{\overline{W} > 0\}$$

where W_i is independent but not identically distributed because

$$W_i \mid Z_{ij} \equiv N(\mu_i, \sigma_i^2), \;\; i = 1, \ldots, n_j$$

where

$$\mu_i = (\mu_{Y_i} - a \cdot \mu_{X_i}) Z_{ij}$$

and

$$\sigma_i^2 = V((Y_i - a \cdot X_i)Z_{ij})$$

which depends on $\sigma_{X_i}^2$ and $\sigma_{Y_i}^2$. The values of these parameters are given from previous studies following the median of the distribution the mean method.

If the data contain no outliers, it will be

$$W_i \mid Z_{ij} \equiv N(\mu_i, \sigma_i^2)$$

but, as usual, a proportion ϵ of outliers in the data is assumed, i.e., as a model for the observations W_i the following

$$F_i = (1-\epsilon)N(\mu_i, \sigma_i^2) + \epsilon N(g_{i1}\,\mu_i, g_{i2}^2\,\sigma_i^2)$$

where the *contamination constants* g_{i1} and g_{i2} are assumed to depend on $i = 1, \ldots, n_j$.

To compute the distribution of $\overline{W}$ under models F_i, assuming that the sample sizes n_j are small, a von Mises approximation (VOM), based on a von Mises expansion, is used to obtain an accurate approximation with small sample sizes.

6.2. VOM Approximation of the Distribution

In general, to approximate the tail probability of statistic T_n under a vector of model distributions $\mathbf{F} = (F_1, \ldots, F_n)$, knowing its tail distribution under the vector of model distributions $\mathbf{G} = (G_1, \ldots, G_n)$ (called *pivotal distributions*), the von Mises expansion of the tail probability of $T_n(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ at $\mathbf{F}$ is used ([10], Section 2, or [22], Theorem 2.1, or [17], Corollary 2),

$$P_{\mathbf{F}}\{T_n(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n) > t\} = P_{F_1,\ldots,F_n}\{T_n(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n) > t\}$$

$$= P_{\mathbf{G}}\{T_n(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n) > t\} + \sum_{i=1}^{n} \int_{\mathcal{X}} \mathrm{TAIF}_i(\mathbf{x}; t; T_n, \mathbf{G})\, dF_i(\mathbf{x}) + Rem$$

where the sample space $\mathcal{X} \subset \mathbb{R}^m$,

$$Rem = \frac{1}{2} \int_{\mathcal{X}} \int_{\mathcal{X}} T_{\mathbf{G_F}}^{(2)}(\mathbf{x}_1, \mathbf{x}_2) d[\mathbf{F}(\mathbf{x_1}) - \mathbf{G}(\mathbf{x_1})] d[\mathbf{F}(\mathbf{x_2}) - \mathbf{G}(\mathbf{x_2})]$$

where $T_{\mathbf{G_F}}^{(2)}$ is the second derivative of the tail probability functional at the mixture distribution $\mathbf{G_F} = (1-\lambda)\mathbf{G} + \lambda\mathbf{F}$, for some $\lambda \in [0, 1]$; and TAIF_i is the ith (multivariate) *partial tail area influence function* of T_n at $\mathbf{G} = (G_1, \ldots, G_n)$ in relation to G_i, $i = 1, \ldots, n$, introduced in [17], Definition 1,

$$\mathrm{TAIF}_i(\mathbf{x}; t; T_n, \mathbf{G}) = \frac{\partial}{\partial \epsilon} P_{G_i^{\epsilon,x}}\{T_n(\mathbf{X}_1, \ldots, \mathbf{X}_n) > t\} \mid_{\epsilon=0}$$

in those $\mathbf{x} \in \mathcal{X}$ where the right-hand side exists. In the computation of TAIF_i, only G_i is contaminated; the other distributions remain fixed, $i = 1, \ldots, n$.

In general, *Rem* is close to 0, and the *von Mises approximation* (VOM) is defined as

$$\begin{aligned} P_{\mathbf{F}}\{T_n(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n) > t\} \quad &\simeq \quad P_{\mathbf{G}}\{T_n(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n) > t\} \\ &\quad + \sum_{i=1}^{n} \int_{\mathcal{X}} \mathrm{TAIF}_i(\mathbf{x}; t; T_n, \mathbf{G})\, dF_i(\mathbf{x}). \end{aligned} \tag{2}$$

Moreover, if $\mathbf{F}$ is a mixture distribution, $\mathbf{F} = (1-\epsilon)\mathbf{G} + \epsilon\mathbf{H}$, $Rem = O(\epsilon^2)$ ([23], p. 77). Additionally, because of the partial influence functions properties ([22], p. 3) that are valid for the partial tail area influence functions defined in [17], for any T_n it will be

$$\int_{\mathcal{X}} \mathrm{TAIF}_i(\mathbf{x}; t; T_n, \mathbf{G}) dG_i(x) = 0, \tag{3}$$

i.e., the integral with respect a given model of the TAIF_i that depends on this model is equal to 0. Hence,

$$\begin{aligned} P_{\mathbf{F}}\{T_n(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n) > t\} &= P_{\mathbf{G}}\{T_n(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n) > t\} \\ &+(1-\epsilon) \sum_{i=1}^{n} \int_{\mathcal{X}} \mathrm{TAIF}_i(\mathbf{x}; t; T_n, \mathbf{G})\, dG_i(\mathbf{x}) \\ &+\epsilon \sum_{i=1}^{n} \int_{\mathcal{X}} \mathrm{TAIF}_i(\mathbf{x}; t; T_n, \mathbf{G})\, dH_i(\mathbf{x}) + O(\epsilon^2) \\ &= P_{\mathbf{G}}\{T_n(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n) > t\} + 0 + \epsilon \sum_{i=1}^{n} \int_{\mathcal{X}} \mathrm{TAIF}_i(\mathbf{x}; t; T_n, \mathbf{G})\, dH_i(\mathbf{x}) + O(\epsilon^2) \end{aligned}$$

i.e., the VOM approximation is

$$\begin{aligned} P_{\mathbf{F}}\{T_n(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n) > t\} &\simeq P_{\mathbf{G}}\{T_n(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n) > t\} \\ &+\epsilon \sum_{i=1}^{n} \int_{\mathcal{X}} \mathrm{TAIF}_i(\mathbf{x}; t; T_n, \mathbf{G})\, dH_i(\mathbf{x}). \end{aligned}$$

Moreover, because of Proposition 1 in [17],

$$\begin{aligned} \mathrm{TAIF}_i(\mathbf{x}; t; T_n, \mathbf{G}) &= -P_{G_1,\ldots,G_n}\{T_n(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n) > t\} \\ &+ P_{G_1,\ldots,G_{i-1},G_{i+1},\ldots,G_n}\{T_n(\mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{x}, \mathbf{X}_{i+1}, \ldots, \mathbf{X}_n) > t\} \end{aligned}$$

and the VOM approximation of the tail probability $P_{\mathbf{F}}\{T_n(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n) > t\}$ can also be expressed as

$$\begin{aligned} P_{\mathbf{F}}\{T_n(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n) > t\} &\simeq (1-n) P_{\mathbf{G}}\{T_n(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n) > t\} \\ &+ \int_{\mathcal{X}} P_{G_2,\ldots,G_n}\{T_n(\mathbf{x}, \mathbf{X}_2, \ldots, \mathbf{X}_n) > t\}\, dF_1(\mathbf{x}) \\ &+ \int_{\mathcal{X}} P_{G_1,G_3,\ldots,G_n}\{T_n(\mathbf{X}_1, \mathbf{x}, \ldots, \mathbf{X}_n) > t\}\, dF_2(\mathbf{x}) + \cdots \\ &+ \int_{\mathcal{X}} P_{G_1,\ldots,G_{n-1}}\{T_n(\mathbf{X}_1, \ldots, \mathbf{X}_{n-1}, \mathbf{x}) > t\}\, dF_n(\mathbf{x}) \end{aligned} \tag{4}$$

which allows an approximation of the tail probability $P_{\mathbf{F}}\{T_n(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n) > t\}$ under models $\mathbf{F} = (F_1, \ldots, F_n)$, knowing the value of this tail probability under near models $\mathbf{G} = (G_1, \ldots, G_n)$.

In the particular case that $T_n(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n) = \overline{W}$, the VOM approximation for the tail of $\overline{W}$ can be expressed as (see (2) with $n = n_j, j = 1, \ldots, L$ and $t = 0$ now)

$$P_{\mathbf{F}}\{\overline{W} > 0\} \simeq P_{\mathbf{G}}\{W_1 + \ldots + W_{n_j} > 0\} + \sum_{i=1}^{n_j} \int_{\mathbb{R}} \mathrm{TAIF}_i(x; 0; \overline{W}, \mathbf{G})\, dF_i(x)$$

or, see (4),

$$\begin{aligned} P_{\mathbf{F}}\{\overline{W} > 0\} &\simeq P_{\mathbf{G}}\{W_1 + \ldots + W_{n_j} > 0\} \\ &+ \sum_{i=1}^{n_j} \int_{\mathbb{R}} \Big[-P_{\mathbf{G}}\{W_1 + \ldots + W_{n_j} > 0\} \end{aligned}$$

$$+P_{G_1,\ldots,G_{i-1},G_{i+1},\ldots,G_{n_j}}\{W_1+\ldots+W_{i-1}+x+W_{i+1}+\ldots+W_{n_j}>0\}\,dF_i(x)\Big]$$

or

$$\begin{aligned}P_{\mathbf{F}}\{\overline{W}>0\} \;\simeq\;& (1-n_j)P_{\mathbf{G}}\{W_1+\ldots+W_{n_j}>0\}\\ &+\int_{\mathbb{R}}P_{G_2,\ldots,G_{n_j}}\{x+W_2+\ldots+W_{n_j}>0\}\,dF_1(x)\\ &+\int_{\mathbb{R}}P_{G_1,G_3,\ldots,G_{n_j}}\{W_1+x+W_3+\ldots+W_{n_j}>0\}\,dF_2(x)+\ldots\\ &+\int_{\mathbb{R}}P_{G_1,\ldots,G_{i-1},G_{i+1},\ldots,G_{n_j}}\{W_1+\ldots+W_{i-1}+x+W_{i+1}+\ldots+W_{n_j}>0\}\,dF_i(x)\\ &+\ldots+\int_{\mathbb{R}}P_{G_1,\ldots,G_{n_j-1}}\{W_1+\ldots+W_{n_j-1}+x>0\}\,dF_{n_j}(x).\end{aligned}$$

If it is assumed as model for the observations W_i

$$F_i=(1-\epsilon)N(\mu_i,\sigma_i^2)+\epsilon N(g_{i1}\,\mu_i,g_{i2}^2\,\sigma_i^2)$$

and it is denoted by $G_i^{g_{i1},g_{i2}}\equiv N(g_{i1}\,\mu_i,g_{i2}^2\,\sigma_i^2)$, and by $G_i\equiv N(\mu_i,\sigma_i^2)$ the pivotal distribution, $i=1,\ldots,n_j$, i.e.,

$$F_i=(1-\epsilon)G_i+\epsilon G_i^{g_{i1},g_{i2}}.$$

the generic component of this last equation is

$$\int_{\mathbb{R}}P_{G_1,\ldots,G_{i-1},G_{i+1},\ldots,G_{n_j}}\{W_1+\ldots+W_{i-1}+x+W_{i+1}+\ldots+W_{n_j}>0\}\,dF_i(x)=$$

$$\int_{\mathbb{R}}P_{G_1,\ldots,G_{i-1},G_{i+1},\ldots,G_{n_j}}\{W_1+\ldots+W_{i-1}+W_{i+1}+\ldots+W_{n_j}>-x\}\,dF_i(x)$$

$$=\int_{\mathbb{R}}\left[1-\Phi\left(\frac{-x-\mu_{-i}}{\sigma_{-i}}\right)\right]dF_i(x)$$

where Φ is the cumulative distribution function of a standard normal distribution,

$$\mu_{-i}=\mu_1+\ldots+\mu_{i-1}+\mu_{i+1}+\ldots+\mu_{n_j}$$

and

$$\sigma_{-i}^2=\sigma_1^2+\ldots+\sigma_{i-1}^2+\sigma_{i+1}^2+\ldots+\sigma_{n_j}^2.$$

If

$$\mu_s=\mu_1+\ldots+\mu_{n_j}=\mu_{-i}+\mu_i$$

and

$$\sigma_s^2=\sigma_1^2+\ldots+\sigma_{n_j}^2=\sigma_{-i}^2+\sigma_i^2$$

then,

$$P_{\mathbf{G}}\{W_1+\ldots+W_{n_j}>0\}=1-\Phi\left(\frac{-\mu_s}{\sigma_s}\right)$$

and

$$P_{\mathbf{F}}\left\{\hat{\beta}_{R_j}>a\right\}=P_{\mathbf{F}}\left\{\sum_{i=1}^{n_j}W_i>0\right\}=P_{\mathbf{F}}\{\overline{W}>0\}$$

$$\simeq 1-\Phi\left(\frac{-\mu_s}{\sigma_s}\right)+\sum_{i=1}^{n_j}\int_{\mathbb{R}}\left[\Phi\left(\frac{-\mu_s}{\sigma_s}\right)-\Phi\left(\frac{-x-\mu_{-i}}{\sigma_{-i}}\right)\right]dF_i(x). \tag{5}$$

Because F_i is a normal mixture

$$F_i=(1-\epsilon)G_i+\epsilon G_i^{g_{i1},g_{i2}}$$

the VOM approximation (5) is

$$= 1 - \Phi\left(\frac{-\mu_s}{\sigma_s}\right) + \epsilon \sum_{i=1}^{n_j} \int_{\mathbb{R}} \left[\Phi\left(\frac{-\mu_s}{\sigma_s}\right) - \Phi\left(\frac{-x-\mu_{-i}}{\sigma_{-i}}\right)\right] dG_i^{g_{i1},g_{i2}}(x).$$

Moreover, because of property (3) for the partial influence functions mentioned before, it is

$$\int_{\mathbb{R}} \left[\Phi\left(\frac{-\mu_s}{\sigma_s}\right) - \Phi\left(\frac{-x-\mu_{-i}}{\sigma_{-i}}\right)\right] dG_i(x) = 0$$

or

$$\int_{\mathbb{R}} \Phi\left(\frac{-x-\mu_{-i}}{\sigma_{-i}}\right) dG_i(x) = \Phi\left(\frac{-\mu_s}{\sigma_s}\right).$$

Hence, making the change of variable $(x+\mu_{-i})/\sigma_{-i} = y$, it is

$$\int_{\mathbb{R}} \Phi\left(\frac{-x-\mu_{-i}}{\sigma_{-i}}\right) dG_i^{g_{i1},g_{i2}}(x) = \Phi\left(\frac{-\mu_s^{g_{i1}}}{\sigma_s^{g_{i2}}}\right)$$

where

$$\mu_s^{g_{i1}} = \mu_1 + \ldots + \mu_{i-1} + g_{i1}\,\mu_i + \mu_{i+1} + \ldots + \mu_{n_j}$$

and

$$\sigma_s^{g_{i2}} = \sqrt{\sigma_1^2 + \ldots + \sigma_{i-1}^2 + g_{i2}\,\sigma_i^2 + \sigma_{i+1}^2 + \ldots + \sigma_{n_j}^2}.$$

Then, the VOM approximation to the distribution of $\hat{\beta}_{R_j}$ is

$$P\left\{\hat{\beta}_{R_j} > a\right\} = 1 - \Phi\left(\frac{-\mu_s}{\sigma_s}\right) + \epsilon \sum_{i=1}^{n_j} \left[\Phi\left(\frac{-\mu_s}{\sigma_s}\right) - \Phi\left(\frac{-\mu_s^{g_{i1}}}{\sigma_s^{g_{i2}}}\right)\right].$$

Example 2. *In a study [24], whether low-density lipoprotein cholesterol (LDL-C) is a cause of coronary artery disease (CAD) was analyzed considering 28 DNA markers*

```
DNA markers X Y
SNP exposure.beta exposure.se outcome.beta outcome.se
1 snp_1 0.0260 0.004 0.0677 0.0286
2 snp_2 -0.0440 0.004 -0.1625 0.0300
.....................................................
27 snp_27 0.0090 0.003 0.0000 0.0255
28 snp_28 -0.0360 0.007 0.0198 0.0647
```

Usually, $Z_i \equiv B(2, 0.5)$ is assumed to be an instrumental variable to mimic biallelic SNPs in Hardy–Weinberg equilibrium. A value

$$IVW = 2.834214$$

was obtained.

With the method proposed in this paper, considering sample sizes of $n = 37$, $n_1 = 17$, $n_2 = 10$, and $n_3 = 10$, and contamination parameters $\epsilon = 0.05$, $g_{i1} = 1$, and $g_{i2} = 1.5$, for the first DNA marker is obtained

$$\mu_s = \mu_1 + \ldots + \mu_{n_j} = 30 \times (0.0677 - a \times 0.0260)$$

$$\sigma_i^2 = 0.0286 \times 37 = 1.0582$$

$$\sigma_s = \sqrt{1.0582 \times 37} = 6.257268$$

$$\mu_s^{g_{i1}} = \mu_s$$

and

$$\sigma_s^{g_{i2}} = \sqrt{\sigma_1^2 + \ldots + \sigma_{i-1}^2 + g_{i2}\,\sigma_i^2 + \sigma_{i+1}^2 + \ldots + \sigma_{n_j}^2}$$

$$= \sqrt{1.0582 \times 36 + 1.5 \times 1.0582} = 6.299405$$

$$MdM_1 = 2.59$$

$$v_1 = \frac{1}{F_{\bar{x}}^{-1}(3/4) - F_{\bar{x}}^{-1}(1/4)} = \frac{1}{8.08 - (-2.877)} = 0.0912.$$

For all the 28 DNA markers, we have

MdM_j	2.59	3.70	2.78	2.71	4.93	...
v_j	0.091	0.151	0.128	0.088	0.068	...

which are combined in the new robust estimate

$$RIVW = 2.042703.$$

7. Conclusions

In this paper, a new method for estimating the parameters in a location–scale contamination model is introduced, in the case where individual observations are not available and, therefore, applying the usual robust methods is not possible, i.e., in summary data problems.

For the location problem, a new estimator was defined that is equal to the usual sample mean when no outliers exist and correcting classical estimations when outliers exist.

This new estimator was applied to one of the most used estimators in Mendelian randomization, the inverse-variance weighted estimator (IVW), defining a new estimator robust inverse weighted estimator (RIVW).

Funding: This study was partially supported by grant PID2021-124933NB-I00 from the Ministerio de Ciencia e Innovación (Spain).

Data Availability Statement: Not applicable.

Acknowledgments: The author is very grateful to the referees and to the assistant editor for their kind and professional remarks.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Huber, P.J.; Ronchetti, E.M. *Robust Statistics*, 2nd ed.; John Wiley & Sons: New York, NY, USA, 2009.
2. Lehmann, E.L. *Theory of Point Estimation*; John Wiley & Sons: New York, NY, USA, 1983.
3. Hampel, F.R.; Ronchetti, E.M.; Rousseeuw, P.J.; Syahel, W.A. *Robust Statistics. The Approach Based on Influence Functions*; John Wiley & Sons: New York, NY, USA, 1986.
4. Basford, K.E.; McLachlan, G.J. Likelihood estimation with normal mixture models. *Appl. Statist.* **1985**, *34*, 282–289. [CrossRef]
5. Berckmoes, B.; Molenberghs, G. On the asymptotic behavior of the contaminated sample mean. *Math. Methods Stat.* **2018**, *27*, 312–323. [CrossRef]
6. Rodríguez, D.; Valdora, M. The breakdown point of the median of means tournament. *Stat. Probab. Lett.* **2019**, *153*, 108–112. [CrossRef]
7. García-Pérez, A. Saddlepoint approximations for the distribution of some robust estimators of the variogram. *Metrika* **2020**, *83*, 69–91. [CrossRef]
8. García-Pérez, A. New robust cross-variogram estimators and approximations for their distributions based on saddlepoint techniques. *Mathematics* **2021**, *9*, 762. [CrossRef]
9. Serfling, R.J. *Approximation Theorems of Mathematical Statistics*; John Wiley & Sons: New York, NY, USA, 1980.
10. Withers, C.S. Expansions for the distribution and quantiles of a regular functional of the empirical distribution with applications to nonparametric confidence intervals. *Ann. Stat.* **1983**, *11*, 577–587. [CrossRef]
11. Hampel, F.R. The Influence Curve and its role in robust estimation. *J. Am. Statist. Assoc.* **1974**, *69*, 383–393. [CrossRef]
12. Field, C.A.; Ronchetti, E. A tail area influence function and its application to testing. *Seq. Anal.* **1985**, *4*, 19–41. [CrossRef]

13. Lugannani, R.; Rice, S. Saddle point approximation for the distribution of the sum of independent random variables. *Adv. Appl. Probab.* **1980**, *12*, 475–490. [CrossRef]
14. Jensen, J.L. *Saddlepoint Approximations*; Clarendon Press: Oxford, UK, 1995.
15. Daniels, H.E. Saddlepoint approximations for estimating equations. *Biometrika* **1983**, *70*, 89–96. [CrossRef]
16. García-Pérez, A. Another look at the Tail Area Influence Function. *Metrika* **2011**, *73*, 77–92. [CrossRef]
17. García-Pérez, A. A linear approximation to the power function of a test. *Metrika* **2012**, *75*, 855–875. [CrossRef]
18. García-Pérez, A. A von Mises approximation to the small sample distribution of the trimmed mean. *Metrika* **2016**, *79*, 369–388. [CrossRef]
19. Gómez-Olivencia, A. The presacral spine of the La Ferrassie 1 Neandertal: A revised inventory. *Bull. Mém. Soc. Anthropol. Paris* **2013**, *25*, 19–38. [CrossRef]
20. Pires, H.F.; Smith, G.D.; Bowden, J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *Int. J. Epidemiol.* **2017**, *46*, 1985–1998. [CrossRef]
21. Slob, E.A.W.; Burgess, S. A comparison of robust Mendelian randomization methods using summary data. *Genet. Epidemiol.* **2020**, *44*, 313–329. [CrossRef] [PubMed]
22. Pires, A.M.; Branco, J.A. Partial influence functions. *J. Multivar. Anal.* **2002**, *83*, 451–468. [CrossRef]
23. Ronchetti, E. Accurate and robust inference. *Econom. Stat.* **2020**, *14*, 74–88. [CrossRef]
24. Waterworth, D.M.; Ricketts, S.L.; Song, K.; Chen, L.; Zhao, J.H.; Ripatti, S.; Aulchenko, Y.S.; Zhang, W.; Yuan, X.; Lim, N.; et al. Genetic Variants Influencing Circulating Lipid Levels and Risk of Coronary Artery Disease. *Arterioscler. Thromb. Vasc. Biol.* **2011**, *30*, 2264–2276. [CrossRef] [PubMed]