Bias Testing for Big Data Analytics

Laura Borrajo,[1*] Ricardo Cao[1]

[1]Department of Mathematics, University of A Coruña,

Faculty of Computer Science, Campus de Elviña, A Coruña 15071, Spain

E-mail: laura.borrajo@udc.es

March 16, 2022

Keywords: bias testing; big data; large sample size; sampling bias; two-sample tests

Abstract:

A procedure for bias testing in a large-sized sample possibly subject to sampling bias is proposed in this paper. A small-sized unbiased simple random sample from the real population is assumed to be additionally observed. The method proposed consists of using an adaptation of two-sample existing methods to test the equality of distributions and the equality of means, but considering the distinctive feature of the context proposed in which two very different sample sizes are involved. For the equality of distributions, the two-sample Kolmogorov-Smirnov test, the Cramer-von Mises criterion and the Mann-Whitney U-test are considered. In the case of testing the mean, the Welch's adaptation of the Student's t-test is used. This two-sample mean test has been considered since different distributions do not necessary imply different means. Some bias

indices are also proposed in order to measure the amount of bias in the big data sample. A comparative analysis between the different methods proposed is performed. Simulation results show the good performance of these methods for bias testing. The proposed techniques are also applied to a real data set concerning airline on-time performance of US flights.

1 Introduction

Big Data Analytics offers many advantages nowadays but also presents new challenges, such as the importance of handling truthful and quality data. The false assumption that *with enough data, numbers speak for themselves*, often considered in this Big Data era, has been widely discussed. Massive data sets are not always totally objective, since on certain occasions, a large sample may not be completely representative of the population due to being biased: Big-But-Biased Data (B3D).[1,2] Some interesting examples are the data provided by the StreetBump smartphone app or the tweets generated by hurricane Sandy.[3]

In this context, when a large amount of data can be collected but the sampling mechanism is not controlled, even though the sample size is very large, the distribution from which this sample comes from does not necessarily coincide with that of the population of interest. This idea has been previously formalized.[2] Let us consider a continuous population with cumulative distribution function F (density f) and let us denote by

$$\mathbf{X} = (X_1, \ldots, X_n)$$

a simple random sample of size n from this population. Let us assume that we

are not able to observe this sample but we observe, instead, another sample,

$$\mathbf{Y} = (Y_1, \ldots, Y_N)$$

of a much larger size ($N >> n$) from a biased distribution G (with density g) different from F but with a common support, $\mathcal{D}$. This condition is formulated assuming a positive biasing function, $w(x), \forall x \in \mathcal{D}$, such that

$$g(x) = w(x)f(x) \ \forall x \in \mathcal{D}. \tag{1}$$

In order to reduce the significant bias that may appear in Big Data, bias correction methods have been already developed, proposing general nonparametric estimators for the mean of a transformation of a continuous random variable.[1,2] For this purpose, when ignoring the biasing weight function, a small-sized simple random sample (SRS) of the real population is assumed to be additionally observed.

However, when working with a large database, a logical first step, previous to bias correction, would be to check if we are in a context of biased data. Therefore, the main objective of this paper is to develop testing methods for bias detection.

The rest of the paper proceeds as follows. Section 2 presents the methods for bias testing and the bias indices introduced. Simulations results and a real data application are included in Section 3. The main conclusions are collected in Section 4. The mathematical proofs of the results presented in Section 2 are included in the Supplementary Material.

2 Methods

In order to detect if bias exist, we can use several existing methods that allow to test if the two distributions involved (F and G) come from the same distribution (unbiased situation) or not (biased situation). This is tantamount to using tests for the null hypothesis:

$$H_0 : F = G$$

against the alternative

$$H_1 : F \neq G,$$

like, for instance, the Kolmogorov-Smirnov test or the Cramer-von Mises criterion.

Despite the fact that the following tests for bias detection are widely known methods, it is important to consider the distinctive feature of our B3D context: we will assume that the ratio of both samples sizes involved does not tend to a constant but to infinity $N/n \to \infty$, i.e., the size of the B3D sample tends to infinity faster than that of the SRS.

2.1 Kolmogorov-Smirnov test

The Kolmogorov–Smirnov test[4,5] (KS test) is a nonparametric test of equality of distributions used to compare the population distribution with a reference probability distribution

$$H_0 : F = F_0,$$

or to compare two populations and conclude if they have the same distribution

$$H_0 : F = G.$$

Let F_n be the empirical cumulative distribution function (ecdf) for the sam-

ple $(X_1, X_2, \ldots, X_n)$ and G_N the ecdf for the sample $(Y_1, Y_2, \ldots, Y_N)$, defined as:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}}, \tag{2}$$

$$G_N(x) = \frac{1}{N}\sum_{j=1}^{N} \mathbf{1}_{\{Y_j \leq x\}}, \tag{3}$$

where $\mathbf{1}$ denotes de indicator function.

It is well known that F_n and G_N are the nonparametric maximum likelihood estimators of F and G, respectively. Therefore, the proximity of F_n and G_N will be indicative of the veracity of H_0, while a large distance between both ecdf will evidence that H_0 is probably false.

The two-sample Kolmogorov–Smirnov statistic quantifies the distance between the two ecdf involved:

$$D_{N,n} = \sup_{x \in \mathbb{R}} |G_N(x) - F_n(x)|,$$

where sup denotes the supremum over all $x \in \mathbb{R}$, while the one-sample test statistic computes the distance between the ecdf of the sample and the cdf of the reference distribution:

$$D_n^{F_0} = \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)|.$$

Many authors have studied the limiting distribution of $D_n^{F_0}$ under the assumption that $F_0(x)$ is continuous[4,5,6,7,8] and showed that the exact distribution of $D_n^{F_0}$ under H_0 is independent of F_0, if F_0 is continuous.[9]

As it happens with the one-sample test statistic with continuous $F = G$ under the null hypothesis, it can be proven that the exact distribution of the two-sample test statistic does not depend on the distributions involved, it is

distribution-free.

Proposition 1. *The two-sample Kolmogorov-Smirnov test is a distribution-free test under H_0 if $F = G$ is continuous.*

The proof of Proposition 1 is included in the Supplementary Material.

For n and N sufficiently large, under H_0 the statistic

$$\sqrt{\frac{N \cdot n}{N+n}} D_{N,n}$$

has the same asymptotic distribution that the Kolmogorov distribution:[5]

$$\begin{aligned} P\left(\sqrt{\frac{N \cdot n}{N+n}} D_{N,n} \leq t\right) \xrightarrow{d} K(t) &= 1 - 2\sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2 t^2} \\ &= \frac{\sqrt{2\pi}}{t} \sum_{i=1}^{\infty} e^{-(2i-1)^2 \pi^2/(8t^2)}, \end{aligned}$$

if $N/n \to c \in (0, \infty)$, where $K(t)$ denotes the Kolmogorov-Smirnov cdf.

Let's see now what happens when considering the distinctive feature of our B3D context, i.e., $N/n \to \infty$:

Proposition 2. *Assuming $F = G$ and $N/n \to \infty$, the statistic $\sqrt{\frac{N \cdot n}{N+n}} D_{N,n}$ has the same asymptotic distribution as the statistics $\sqrt{\frac{Nn}{N+n}} D_n^F$ and $\sqrt{n} D_n^F$ when $F = F_0$.*

As a consequence, when $N/n \to \infty$, the two-sample statistic, $D_{N,n}$, will be used but it has to be callibrated with the asymptotic distribution of the one-sample test.

Apart from this test, we could also consider other tests or criterions. In Subsections 2.2 and 2.3 the Cramer-von Mises criterion and the Mann-Whitney U test will be used for bias detection.

2.2 Cramer-von Mises criterion

The Cramer-von Mises criterion, like the Kolmogorov-Smirnov test, is used to judge the goodness of fit of a theoretical cumulative distribution function, F_0, compared to the empirical distribution function, F_n.[10,11] As F is unknown, the two following statistics could be used:

$$\tilde{\omega}^2 = \int_{-\infty}^{\infty} [F_n(x) - F_0(x)]^2 dF_n(x),$$

$$\tilde{\omega}^2 = \int_{-\infty}^{\infty} [F_n(x) - F_0(x)]^2 dF_0(x)$$

For comparing two empirical distributions, the generalization to the two-sample case is given by:

$$\omega^2 = \int_{-\infty}^{\infty} [F_n(x) - G_N(x)]^2 dH_{N+n}(x)$$

which compares the two empirical cdf.[12] In our context, F_n and G_N denote the empirical distribution functions of the SRS and the B3D sample respectively, H_{N+n} the empirical distribution function corresponding to the pooled sample, i.e., $(N+n)H_{N+n}(x) = nF_n(x) + NG_N(x)$.

The statistic for the one-sample case is

$$T_n = n\tilde{\omega}^2 = n\int_{-\infty}^{\infty} [F_n(x) - F_0(x)]^2 dF_0(x) = \frac{1}{12n} + \sum_{i=1}^{n} \left[\frac{2i-1}{2n} - F_0\left(X_{(i)}\right) \right]^2,$$

where $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ is the ordered sample and for the two-sample case:

$$\begin{aligned} T_{N,n} &= \frac{Nn}{N+n}\omega^2 = \frac{Nn}{N+n}\int_{-\infty}^{\infty} [F_n(x) - G_N(x)]^2 dH_{N+n}(x) \\ &= \frac{V}{Nn(N+n)} - \frac{4Nn-1}{6(N+n)}, \end{aligned}$$

where V is defined by

$$V = n\sum_{i=1}^{n}(r_i - i)^2 + N\sum_{j=1}^{N}(s_j - j)^2$$

being $r_i, i = 1, 2, \ldots, n$, the ranks of the ordered SRS, $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$, in the combined sample and $s_j, j = 1, 2, \ldots, N$, the ranks of the ordered B3D sample, $Y_{(1)} \leq Y_{(2)} \leq \ldots \leq Y_{(n)}$, in the pooled sample.

It has been proved that, under the null hypothesis ($F = G$), $T_{N,n}$ has the same limiting null distribution ($F = F_0$) as T_n when $n \to \infty$, $N \to \infty$ and $N/n \to \lambda$, being λ a positive constant.[13] For moderate sample sizes, the limiting distribution is a good approximation to the exact distribution.

Proposition 3. *Assuming $F = G$ and $N/n \to \infty$, the statistic $T_{N,n}$ has the same asymptotic distribution that the statistic T_n under $F = F_0$.*

2.3 Mann–Whitney U test

The Mann–Whitney U test, also called the Mann–Whitney–Wilcoxon (MWW) or Wilcoxon rank-sum test,[14,15] is a nonparametric test of the null hypothesis that, for randomly selected values X and Y from two populations, the probability of X being greater than Y is equal to the probability of Y being greater than X:

$$H_0 : P(X > Y) = P(Y > X).$$

The U test is weaker than that of Kolmogorov-Smirnov, since it does not test the equality of distributions, but a condition that is verified in that case.

To compute the statistic, U, the pooled sample is ranked and each of the values of the two samples is assigned to its rank (i.e., rank 1 is assigned to the smallest observation, rank 2 to the second smallest observation, and so on). If two or more observations are equal, the mean rank is assigned to the tied

observations. Finally, R_X and R_Y, the adjusted rank-sums, (i.e. the sum of the ranks in each of the samples X and Y, respectively), are computed. This allows to construct:

$$U_X = Nn + \frac{n(n+1)}{2} - R_X$$

$$U_Y = Nn + \frac{N(N+1)}{2} - R_Y.$$

Knowing that

$$R_X + R_Y = \frac{(N+n)(N+n+1)}{2},$$

the sum of two values is given by

$$U_X + U_Y = Nn.$$

The U statistic is defined as the minimum between U_X and U_Y:

$$U = \min\{U_X, U_Y\}.$$

For large sized samples, U is approximately normally distributed under the null hypothesis. In that case, the standardized value, $z = (U - m_U)/\sigma_U$, has a standard normal asymptotic distribution, where m_U and σ_U are the mean and the standard deviation of U, under H_0, which are given by $m_U = (Nn)/2$ and $\sigma_U = \sqrt{[Nn(N+n+1)]/12}$.

When the two populations have very different distributions, the Mann-Whitney U-test can lead to a misinterpretation of the results.[16] In that case, it is recommended to use the unequal variances version of the t-test (Welch's t-test), which provides more reliable results.

2.4 Bias detection for mean estimation

Since different distributions may have equal means, it would be reasonable to use a specific two-sample test for the means when studying bias testing in a mean estimation problem.

To see the effect of bias on mean estimation, the Student's t-test for equal means will be used. In particular, in our context, the Welch's adaptation of the two sample t-test[17,18] will be considered. Welch's t-test is a more reliable version of the test when the two populations have unequal variances and/or the samples have unequal sizes.

Welch's t-test defines the statistic as follows:

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{N}}},$$

where $\frac{S_X^2}{n}$ and $\frac{S_Y^2}{N}$ denote the estimated variances of $\overline{X}$ and $\overline{Y}$, respectively; being the degrees of freedom:

$$d.f. = \frac{\left(\frac{S_X^2}{n} + \frac{S_Y^2}{N}\right)^2}{\frac{S_X^4}{n^2(n-1)} + \frac{S_Y^4}{N^2(N-1)}}.$$

This test will not be affected by the condition $N/n \to \infty$ since, in that case, the variance of $\overline{X} - \overline{Y}$,

$$\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{N}$$

tends to the variance of $\overline{X}$. In fact, when $N >> n$, the degrees of freedom are

approximately $d.f. \simeq n-1$ and the statistic t is approximately equal to

$$\frac{\overline{X}-\mu}{\sqrt{\frac{S_X^2}{n}}},$$

which has a standard normal asymptotic distribution.

2.5 Bias indices

Several indices to measure the amount of bias are defined below. All of them are invariant under location and scale transformations. This means that if we consider any positive constant $a > 0$ and any real number b, the index defined for the two new random variables, $X' = aX + b$ and $Y' = aY + b$, has the same value as for the original random variables, X and Y. This is a very convenient property since the value of the index does not depend on the measure units used. All the indices except i_1 are defined in such a way that they all lie within the interval [0,1]. The value 0 for all those indices corresponds to no bias, while the value 1 is the maximal possible value of them.

The first index considers the absolute value of the difference of the population means of the distributions involved in each case. The average of the standard deviations in the denominator is necessary in order to obtain an scale-invariant index:

$$i_1 = \frac{|\mu_Y - \mu_X|}{\frac{\sigma_X + \sigma_Y}{2}}.$$

The following two indices are based, respectively, on the L_1 and L_2 distances

between the density functions:

$$\begin{aligned} d_{L_1} &= \|f-g\|_1 = \int_a^b |f(x)-g(x)|dx, \\ d_{L_2} &= \|f-g\|_2 = \left[\int_a^b (f(x)-g(x))^2 dx\right]^{1/2}, \end{aligned}$$

where $[a,b]$ is the common support of F and G.

Since the distance d_{L_1} takes values between 0 and 2, the second index is divided by 2 in order to be in the range $[0,1]$:

$$i_2 = \frac{1}{2}\|f-g\|_1.$$

Since the distance d_{L_2} is not an scale-invariant measure, it is transformed to obtain the third relative index in $[0,1]$:

$$i_3 = \frac{\|f-g\|_2}{\|f\|_2 + \|g\|_2}.$$

The fourth and fifth indices consider the Kolmogorov-Smirnov and the Cramer-von Mises distances between the distribution functions, respectively:

$$\begin{aligned} d_{KS} &= \sup_{x\in\mathbb{R}} |F(x)-G(x)|, \\ d_{CvM} &= \int_{-\infty}^{\infty} (F(x)-G(x))^2 \frac{1}{2} d\left(F+G\right)(x). \end{aligned}$$

The Kolmogorov-Smirnov distance is already a location and scale invariant measure that takes values in $[0,1]$, therefore it does not require any modification to obtain the fourth index:

$$i_4 = d_{KS} = \sup_{x\in\mathbb{R}} |F(x)-G(x)|.$$

Since the Cramer-von Mises distance is location and scale invariant but takes values between 0 and 1/3, the fifth index is:

$$i_5 = 3\int_{-\infty}^{\infty}(F(x)-G(x))^2\frac{1}{2}d\left(F+G\right)(x).$$

Finally, an index that measures the proximity of the biasing weight function, w, defined in (1), to its nearest constant function is considered:

$$i_6 = \frac{\|w-c_w\|_2}{\|w\|_2+\|c_w\|_2} = \frac{\left[\int_a^b (w(x)-c_w)^2dx\right]^{1/2}}{\left[\int_a^b w(x)^2dx\right]^{1/2}+(b-a)^{1/2}\cdot c_w},$$

where

$$c_w = \frac{1}{b-a}\int_a^b w(x)dx$$

and the correction in the denominator is introduced to get a location and scale invariant index with values in $[0,1]$.

3 Results and Discussion

3.1 Experiments

The performance of the tests proposed in Section 2 is studied via simulation. We generated $1,000$ pairs of datasets, each with sample size $n = 1,000$ in the case of the sample $\mathbf{X}$ and sample size $N = 1,000,000$ for the sample $\mathbf{Y}$.

Let us consider a population with density f,

$$f(x) = \frac{3}{14}(x^2+1)\ \mathbf{1}_{[0,2]}(x),$$

from which the sample $\mathbf{X}$ is generated and the following class of biasing weight

functions,

$$w(x) = \varepsilon^k \, \mathbf{1}_{[0,\varepsilon]}(x) + x^k \, \mathbf{1}_{(\varepsilon,2]}(x),$$

with different choices of $k > 0$ and $\varepsilon > 0$.

The biased density is

$$g(x) = \frac{3}{14c}\varepsilon^k(x^2+1) \, \mathbf{1}_{[0,\varepsilon]}(x) + \frac{3}{14c}(x^{k+2}+x^k) \, \mathbf{1}_{(\varepsilon,2]}(x),$$

being

$$c = \frac{1}{14}\left[\frac{k\cdot\varepsilon^{k+3}+3\cdot 2^{k+3}}{k+3} + \frac{3(k\cdot\varepsilon^{k+1}+2^{k+1})}{k+1}\right],$$

from which we simulate the sample $\mathbf{Y}$.

Different combinations of k and ε are considered in this simulation study, providing very biased situations ($k = 2$, $\varepsilon = 0.1$) and others in which bias is quite significant (see Figure 1), decreasing the degree of bias by decreasing k and increasing ε (see Figure 2), until reaching situations in which bias is almost imperceptible (see Figure 3) or it does not exist ($k = 0$, $\varepsilon = 2$).

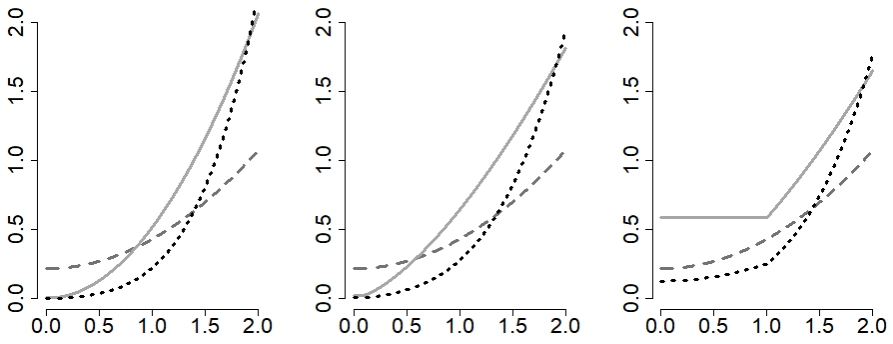

Figure 1: Densities f (dashed gray line) and g (dotted black line) involved in the simulated models for different values of k and ε for the biasing function, w (solid line). Left panel: $k = 2$, $\varepsilon = 0.1$; middle panel: $k = 1.5$, $\varepsilon = 0.1$; right panel: $k = 1.5$, $\varepsilon = 1$.

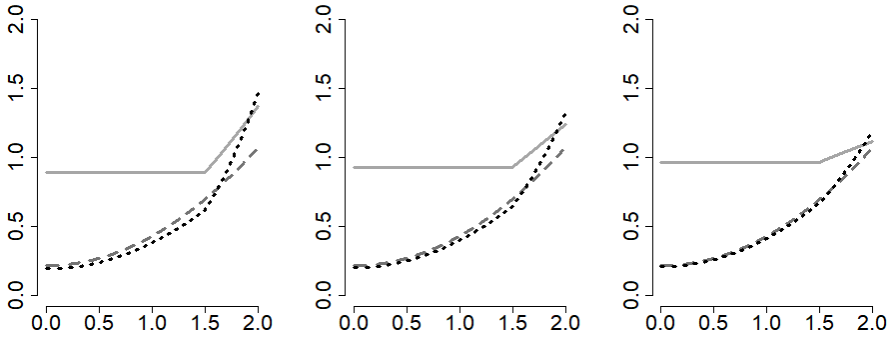

Figure 2: Densities f (dashed gray line) and g (dotted black line) involved in the simulated models for different values of k and ε for the biasing function, w (solid line). Left panel: $k = 1.5$, $\varepsilon = 1.5$; middle panel: $k = 1$, $\varepsilon = 1.5$; right panel: $k = 0.5$, $\varepsilon = 1.5$.

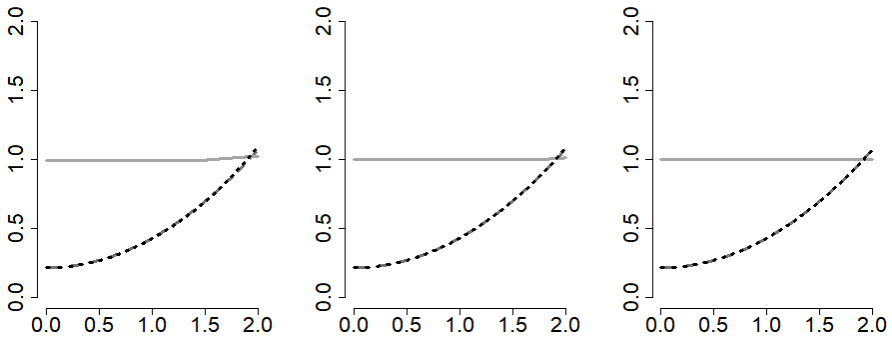

Figure 3: Densities f (dashed gray line) and g (dotted black line) involved in the simulated models for different values of k and ε for the biasing function, w (solid line). Left panel: $k = 0.1$, $\varepsilon = 1.5$; middle panel: $k = 0.1$, $\varepsilon = 1.8$; right panel: $k = 0$, $\varepsilon = 2$.

Table 1 shows the values of the different bias indices considered. It is clearly observed that in the most biased situation the value of all indices is greater, decreasing as the bias decreases.

For the implementation of the two sample KS test in R the *pkolmim* function

Table 1: Values of the six relative bias indices for the nine scenarios considered.

k	ε	i_1	i_2	i_3	i_4	i_5	i_6
2	0.1	0.661	0.276	0.263	0.276	0.121	0.382
1.5	0.1	0.573	0.231	0.223	0.231	0.085	0.333
1.5	1	0.344	0.165	0.173	0.165	0.041	0.192
1.5	1.5	0.114	0.067	0.086	0.067	0.006	0.065
1	1.5	0.074	0.044	0.057	0.044	0.002	0.042
0.5	1.5	0.037	0.022	0.028	0.022	0.001	0.020
0.1	1.5	0.007	0.004	0.006	0.004	10^{-5}	0.004
0.1	1.8	0.001	0.001	0.002	0.001	10^{-6}	0.001
0	2	0	0	0	0	0	0

of the *kolmim* package[19] is used. This is an improved version of the function *ks.test*. The reason for using this package is that the *ks.test* function returns approximated values in case of ties, being the *pkolmim* function more efficient since it returns the exact values. For the implementation of the two-sample Cramer-von Mises criterion we use the *cvm_test* function of the *twosamples* package with 1000 bootstrap iterations and for the Mann-Whitney test the *wilcox.test* function of the *stats* package. As for the *ks.test* function, *wilcox.test* returns approximated values due to the presence of ties. For equal means testing, the *t.test* with unequal variances is used.

Table 2 shows the rejection proportions obtained for different test proposed for bias testing. Except for the Cramer-von Mises criterion, whose bad results apparently come from a malfunction of the *twosamples* package, the rest of the methods considered to test $F = G$ offer similar conclusions. As the indices considered in Table 1 showed, for the first combinations of k and ε the absence of bias is totally rejected with probability 1, while in the last cases, H_0 is rejected only 5% of the times. Regarding the two-sample test for equal means, the conclusions are similar.

Table 2: Rejection proportions when testing the equality of the distributions, F and G, using the-two sample KS test, through the *ks.test* and the *pkolmim* functions, the two-sample Cramer-von Mises criterion and the Mann-Whitney-Wilcoxon U-test and rejection proportions for the two sample means test using the Welch's t-test for different values of k and ε ($n = 1,000$, $N = 1,000,000$, trials=$1,000$, $\alpha = 0.05$)

k	ε	*ks.test*	*pkolmim*	*cvm_test*	*wilcox.test*	*t.test*
2	0.1	1	1	1	1	1
1.5	0.1	1	1	1	1	1
1.5	1	1	1	1	1	1
1.5	1.5	0.990	0.991	1	0.990	0.956
1	1.5	0.781	0.786	0.915	0.774	0.651
0.5	1.5	0.237	0.247	0.468	0.273	0.209
0.1	1.5	0.053	0.055	0.161	0.059	0.050
0.1	1.8	0.048	0.050	0.150	0.052	0.052
0	2	0.049	0.050	0.154	0.051	0.050

3.2 Real data application

The airline on-time performance (AOTP) data set consists of nearly 180 million records about flight arrival and departure details for all commercial flights within the US, from October 1987 to December 2021. It is available at Bureau of Transportation Statistics.[20]

The main interest is the mean arrival delay time (in minutes) of US flights for the whole year 2017. It is assumed that all the 2017 are not available, but only data until January 11th, 2017 have been collected. Presence of bias in other big data sets is studied. The sample $\mathbf{Y}$ is considered as the whole data set for the year 2016 ($N = 5,617,658$). Within 2017, only the arrival delay time for the flights of January 11th, 2017, $\mathbf{X}$, ($n = 14,568$) is assumed to be available. Since the first days of January are atypical due to the holiday period and since weekends and Mondays do not always accurately reflect the behavior of a normal labor day, the first available Wednesday (January 11th, 2017) has been consider in order to obtain something close to a SRS of the true 2017 population.

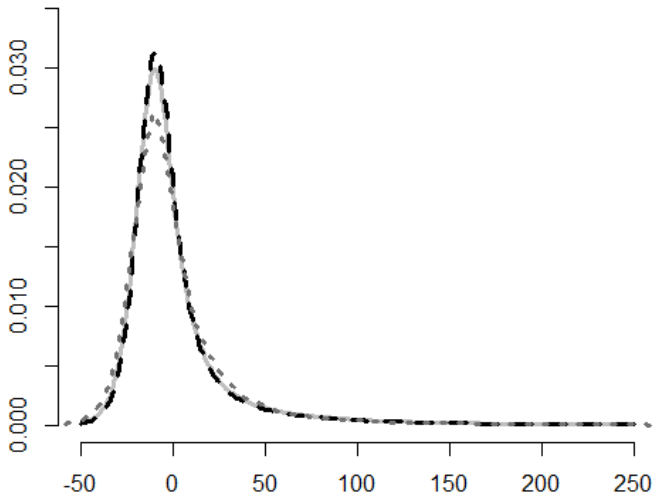

Figure 4: Estimated densities involved in the case study with AOTP data. Densities of arrival delays in 2016 (dashed black line), 2017 (solid gray line) and January 11th, 2017 (dotted gray line).

To illustrate the difference between the density functions of the arrival delay of US flights in 2016 and 2017, two kernel density estimations have been plotted in Figure 4. These density estimates are based on nearly 6 million data each. Although the two estimated annual densities are very similar, they exhibit some subtle differences, for instance the level of the density at the mode. Figure 4 also contains the kernel density estimation based on the arrival delays of January 11th, 2017.

To test for sampling bias, we use some of the methods proposed in Section 2. Looking at the values of the estimated bias indices shown in Table 3, it seems that bias is practically imperceptible. However, the value of the sixth index in Table 3 and the results obtained in Table 4 offer different conclusions.

Table 3: Comparison of the estimated relative bias indices.

Variable	i_1	i_2	i_3	i_4	i_5	i_6
Arrival delay	0.0258	0.0149	0.0624	0.0337	0.0025	0.3194

To test the equality of distributions we use the two-sample Kolmogorov–Smirnov test through the *pkolmim* function since it is the only one that returns the exact p-value. We also test the equality of means using the Student's t-test. The p-values obtained in Table 4 allow to reject the null hypothesis of no bias with both methods. So there exists statistical evidence of the presence of bias and difference of means. Therefore, although looking at Figure 4 and the values of the indices obtained in Table 3 bias seems to be negligible, we can conclude that the difference is big enough to take it into account. In fact, the sample mean of January 11th, 2017 is $\overline{X} = 4.742243$, quite different from $\overline{Y} = 3.519290$, the sample mean of the whole 2016. Using the correction method proposed,[2] the bias-corrected estimate is $\hat{\mu} = 4.326778$, relatively close to the sample mean of the whole 2017, 4.326357.

Table 4: p-values obtained using the two-sample Kolmogorov–Smirnov test for equality of distributions through the *pkolmim* function and using the two-sample Student's t-test for equality of means.

Variable	*pkolmim*	*t.test*
Arrival delay	4.3×10^{-14}	0.005194

4 Conclusions

In the era of big data, sampling bias is more present than before in statistical data analysis. Testing for sampling bias is an extremely important issue in a big data context. Several existing methods have been used to test for no sampling bias (i.e. $F = G$). Of course, the fact that $N/n \to \infty$ makes a difference with the

classical asymptotic theory. So the test procedures have been adapted. Several relative bias indices have been proposed in order to quantify the amount of existing bias. The performance of the proposed tests has been studied through a simulation study, showing their good behavior when detecting the presence of bias. These techniques have been also applied to a real data set. The results show how, even in situations with little bias, it is likely that this bias would be considerable enough to be taken into account.

Author Disclosure Statement

No competing financial interests exist.

Funding Information

This research has been supported by the MICINN Grant PID2020-113578RB-I00 and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2020-14 and Centro de Investigación del Sistema Universitario de Galicia ED431G 2019/ 01), all of them through the European Regional Development Fund (ERDF).

References

[1] Borrajo L, and Cao R. Big-But-Biased Data Analytics for Air Quality. Electronics. 2020;9:1551. https://doi.org/10.3390/electronics9091551.

[2] Borrajo L, and Cao R. Nonparametric Estimation for Big-But-Biased Data. TEST. 2021;30(4),861-883. https://doi.org/10.1007/s11749-020-00749-5.

[3] Crawford K. The hidden biases in big data. Harvard Business Review. 2013. https://hbr.org/2013/04/the-hidden-biases-in-big-data.

[4] Kolmogorov A. Sulla determinazione empirica di una legge di distribuzione. Giornale dell'Istituto Italiano degli Attuari. 1933;4:83-91.

[5] Smirnov N. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. Bulletin Moscow University. 1939;2:3-14.

[6] Feller W. On the Kolmogorov-Smirnov limit theorems for empirical distributions. The Annals of Mathematical Statistics. 1948;19:177-189.

[7] Doob J. Heuristic approach to the Kolmogorov-Smirnov theorems. The Annals of Mathematical Statistics. 1949;20:393-403.

[8] Smirnov N. Table for estimating the goodness of fit of empirical distributions. The Annals of Mathematical Statistics. 1948;19:279-281.

[9] Massey F. The Kolmogorov-Smirnov test for goodness of fit. Journal of the American Statistical Association. 1951;46:68-78.

[10] Cramer H. On the composition of elementary errors: First paper: Mathematical deductions. Scandinavian Actuarial Journal. 1928;1928:13-74.

[11] von Mises R. Statistik und wahrheit. Julius Springer. 1928;20.

[12] Anderson T. On the distribution of the two-sample Cramer-von Mises criterion. The Annals of Mathematical Statistics. 1962;33:1148-1159.

[13] Rosenblatt M. Limit theorems associated with variants of the von Mises statistic. The Annals of Mathematical Statistics. 1952;23:617-623.

[14] Wilcoxon F. Individual comparisons by ranking methods. Biometrics Bulletin. 1945;1:80-83.

[15] Mann H, and Whitney D. On a test of whether one of two random variables is stochastically larger than the other. The Annals of Mathematical Statistics. 1947;18:50-60.

[16] Kasuya E. Mann-Whitney U test when variances are unequal. Animal Behaviour. 2001;61:1247-1249.

[17] Welch B. The generalization of student's' problem when several different population variances are involved. Biometrika. 1947;34:28-35.

[18] Welch B. On the comparison of several mean values: an alternative approach. Biometrika. 1951;38:330-336.

[19] Carvalho L. An improved evaluation of Kolmogorov's distribution. Journal of Statistical Software. 2015;65:1-7.

[20] Bureau of Transportation Statistics. Reporting Carrier On-Time Performance (1987-present). Available online at `https://www.transtats.bts.gov/Tables.asp?QO_VQ=EFD&QO_anzr=Nv4yv0r%FDb0-gvzr%FDcr4s14zn0pr%FDQn6n&QO_fu146_anzr=b0-gvzr`

Supplementary material

Proposition 1. *The two-sample Kolmogorov-Smirnov test is a distribution-free test under H_0 if $F = G$ is continuous.*

Proof. Let us define the generalized inverse of F (or quantile function) by

$$F^{-1}(t) = \min\{x : F(x) \geq t\}.$$

Taking into account the change of variable $t = F(x)$ or $x = F^{-1}(t)$, we can write the statistic as

$$D_{N,n} = \sup_{x \in \mathbb{R}} |G_N(x) - F_n(x)| = \sup_{0<t<1} |G_N(F^{-1}(t)) - F_n(F^{-1}(t))|.$$

Using the definitions of the ecdfs (2) and (3), under the null hypothesis H_0, we obtain:

$$\begin{aligned} F_n(F^{-1}(t)) &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq F^{-1}(t)\}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{F(X_i) \leq t\}}, \\ G_N(F^{-1}(t)) &= \frac{1}{N}\sum_{j=1}^{N} \mathbf{1}_{\{Y_j \leq F^{-1}(t)\}} = \frac{1}{N}\sum_{j=1}^{N} \mathbf{1}_{\{F(Y_j) \leq t\}}, \end{aligned}$$

and therefore,

$$\sup_{0<t<1} |G_N(F^{-1}(t)) - F_n(F^{-1}(t))| = \sup_{0<t<1} \left| \frac{1}{N}\sum_{j=1}^{N} \mathbf{1}_{\{F(Y_j) \leq t\}} - \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{F(X_i) \leq t\}} \right|.$$

The distributions of $F(X_i)$ and $F(Y_j)$ are uniform on the interval $[0, 1]$ since

$$P(F(X_i) \leq t) = P(X_i \leq F^{-1}(t)) = F(F^{-1}(t)) = t$$

and

$$P(F(Y_j) \leq t) = P(Y_j \leq F^{-1}(t)) = F(F^{-1}(t)) = t.$$

Therefore, the random variables $U_i = F(X_i), i = 1, \ldots, n$ and $V_j = F(Y_j), j = 1, \ldots, N$ are independent and have uniform distribution on $[0, 1]$, so:

$$D_{N,n} = \sup_{x \in \mathbb{R}} |G_N(x) - F_n(x)| = \sup_{0<t<1} \left| \frac{1}{N} \sum_{j=1}^{N} \mathbf{1}_{\{V_j \leq t\}} - \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{U_i \leq t\}} \right|,$$

which clearly does not depend on F. □

Proposition 2. *Assuming $F = G$ and $N/n \to \infty$, the statistic $\sqrt{\frac{N \cdot n}{N+n}} D_{N,n}$ has the same asymptotic distribution as the statistics $\sqrt{\frac{Nn}{N+n}} D_n^F$ and $\sqrt{n} D_n^F$ when $F = F_0$.*

Proof. Assuming $F = G$ and defining

$$D_N^G = \sup_{x \in \mathbb{R}} |G_N(x) - G(x)|$$

and

$$D_n^F = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|,$$

give

$$\begin{aligned} \sqrt{\frac{N \cdot n}{N+n}} D_{N,n} &= \sqrt{\frac{N \cdot n}{N+n}} \sup_{x \in \mathbb{R}} |G_N(x) - G(x) + F(x) - F_n(x)| \\ &\leq \sqrt{\frac{N \cdot n}{N+n}} \sup_{x \in \mathbb{R}} |G_N(x) - G(x)| + \sqrt{\frac{N \cdot n}{N+n}} \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \\ &= \sqrt{\frac{n}{N+n}} \sqrt{N} D_N^G + \sqrt{\frac{N}{N+n}} \sqrt{n} D_n^F \xrightarrow{d} K, \end{aligned} \tag{4}$$

since, when $N/n \to \infty$,

$$\sqrt{\frac{n}{N+n}} \simeq \sqrt{\frac{n}{N}} = o(1),$$

$$\sqrt{N} D_N^G \xrightarrow{d} K,$$

$$\sqrt{\frac{N}{N+n}} \simeq \sqrt{\frac{N}{N}} = 1$$

and

$$\sqrt{n} D_n^F \xrightarrow{d} K.$$

On the other hand, under $F_0 = F = G$, the one-sample test statistic, D_n^F, satisfies:

$$\begin{aligned} \sqrt{n} D_n^F &= \sqrt{n} \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| = \sqrt{n} \sup_{x \in \mathbb{R}} |F_n(x) - G_N(x) + G_N(x) - G(x)| \\ &\leq \sqrt{n} D_{N,n} + \sqrt{n} D_N^G = \sqrt{n} D_{N,n} + \sqrt{\frac{n}{N}} \sqrt{N} D_N^G, \end{aligned}$$

which implies that

$$\sqrt{\frac{N}{N+n}} \sqrt{n} D_n^F \leq \sqrt{\frac{Nn}{N+n}} D_{N,n} + \sqrt{\frac{n}{N+n}} \sqrt{N} D_N^G$$

and therefore

$$\sqrt{\frac{N}{N+n}} \sqrt{n} D_n^F - \sqrt{\frac{n}{N+n}} \sqrt{N} D_N^G \leq \sqrt{\frac{Nn}{N+n}} D_{N,n}. \tag{5}$$

Considering (4) and (5), we obtain:

$$\begin{aligned} \sqrt{\frac{N}{N+n}} \sqrt{n} D_n^F - \sqrt{\frac{n}{N+n}} \sqrt{N} D_N^G &\leq \sqrt{\frac{N \cdot n}{N+n}} D_{N,n} \\ &\leq \sqrt{\frac{N}{N+n}} \sqrt{n} D_n^F + \sqrt{\frac{n}{N+n}} \sqrt{N} D_N^G \end{aligned}$$

and since $\sqrt{\frac{n}{N+n}}\sqrt{N}D_N^G \simeq o_p(1)$, it is concluded that the asymptotic distribution of $\sqrt{\frac{N\cdot n}{N+n}}D_{N,n}$ is the same as that of $\sqrt{\frac{N}{N+n}}\sqrt{n}D_n^F$, which, since $N/n \to \infty$, is the same as the asymptotic distribution of $\sqrt{n}D_n^F$. □

Proposition 3. *Assuming $F = G$ and $N/n \to \infty$, the statistic $T_{N,n}$ has the same asymptotic distribution that the statistic T_n under $F = F_0$.*

Proof. Defining

$$T_N^G = N\int_{-\infty}^{\infty}[G_N(x) - G(x)]^2 dG(x),$$

$$T_n^F = n\int_{-\infty}^{\infty}[F_n(x) - F(x)]^2 dF(x)$$

and

$$T_{N,n}^F = \frac{Nn}{N+n}\int_{-\infty}^{\infty}[F_n(x) - G_N(x)]^2 dF(x),$$

it can be proven that, under $F = G$, $T_{N,n}$ has the same asymptotic distribution that $T_{N,n}^F$.

On the one hand, using the triangular inequality:

$$\begin{aligned}&\left(\int_{-\infty}^{\infty}[F_n(x) - F(x) + G(x) - G_N(x)]^2 dF(x)\right)^{1/2}\\ &\quad\leq \left(\int_{-\infty}^{\infty}[F_n(x) - F(x)]^2 dF(x)\right)^{1/2} + \left(\int_{-\infty}^{\infty}[G_N(x) - G(x)]^2 dG(x)\right)^{1/2},\end{aligned}$$

which, using $F = G$, implies that

$$\begin{aligned}T_{N,n}^F &= \frac{Nn}{N+n}\int_{-\infty}^{\infty}[F_n(x) - F(x) + G(x) - G_N(x)]^2 dF(x) \qquad (6)\\ &\leq \frac{Nn}{N+n}\int_{-\infty}^{\infty}[F_n(x) - F(x)]^2 dF(x) + \frac{Nn}{N+n}\int_{-\infty}^{\infty}[G_N(x) - G(x)]^2 dG(x)\\ &+ \frac{2Nn}{N+n}\left(\int_{-\infty}^{\infty}[F_n(x) - F(x)]^2 dF(x)\right)^{1/2}\left(\int_{-\infty}^{\infty}[G_N(x) - G(x)]^2 dG(x)\right)^{1/2}.\end{aligned}$$

On the other hand, using again the triangular inequality:

$$\left(\int_{-\infty}^{\infty}[F_n(x)-G_N(x)+G_N(x)-G(x)]^2dF(x)\right)^{1/2}$$
$$\leq \left(\int_{-\infty}^{\infty}[F_n(x)-G_N(x)]^2dF(x)\right)^{1/2}+\left(\int_{-\infty}^{\infty}[G_N(x)-G(x)]^2dG(x)\right)^{1/2},$$

then, using that $F_0 = F = G$, the one-sample test statistic, T_n^F, satisfies:

$$\begin{aligned}T_n^F &= n\int_{-\infty}^{\infty}[F_n(x)-F(x)]^2dF(x)\\ &= n\int_{-\infty}^{\infty}[F_n(x)-G_N(x)+G_N(x)-G(x)]^2dF(x)\\ &\leq n\int_{-\infty}^{\infty}[F_n(x)-G_N(x)]^2dF(x)+n\int_{-\infty}^{\infty}[G_N(x)-G(x)]^2dG(x)\\ &+ 2n\left(\int_{-\infty}^{\infty}[F_n(x)-G_N(x)]^2dF(x)\right)^{1/2}\left(\int_{-\infty}^{\infty}[G_N(x)-G(x)]^2dG(x)\right)^{1/2},\end{aligned}$$

which implies that

$$\frac{N}{N+n}T_n^F-\frac{n}{N+n}T_N^G-\frac{2\sqrt{Nn}}{N+n}\left(T_n^F\right)^{1/2}\left(T_N^G\right)^{1/2}\leq T_{N,n}^F. \tag{7}$$

Considering (6) and (7), we obtain:

$$\begin{aligned}&\frac{N}{N+n}T_n^F-\frac{n}{N+n}T_N^G-\frac{2\sqrt{Nn}}{N+n}\left(T_n^F\right)^{1/2}\left(T_N^G\right)^{1/2}\leq T_{N,n}^F\\ &\leq \frac{N}{N+n}T_n^F+\frac{n}{N+n}T_N^G+\frac{2\sqrt{Nn}}{N+n}\left(T_n^F\right)^{1/2}\left(T_N^G\right)^{1/2}\end{aligned}$$

and since, when $N/n\to\infty$,

$$\frac{n}{N+n}\simeq\frac{n}{N}=o(1),$$

$$\frac{\sqrt{Nn}}{N+n}\simeq\frac{\sqrt{n}}{\sqrt{N+n}}=o(1),$$

$$T_n^F = O_P(1)$$

and

$$T_N^G = O_P(1),$$

it is concluded that the asymptotic distribution of $T_{N,n}^F$ is the same as that of $\frac{N}{N+n}T_n^F$, which, since $N/n \to \infty$, is the same as the asymptotic distribution of T_n^F. □