Received: 6 April 2022 | Revised: 25 May 2022 | Accepted: 4 June 2022

DOI: 10.1002/sta4.478

ORIGINAL ARTICLE

WILEY

On the effect of rounding on hypothesis testing when sample size is large

Nikolai G. Ushakov[1] | Vladimir G. Ushakov[2]

[1]Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, 7034, Norway

[2]Department of Mathematical Statistics, Moscow State University, Moscow, 119991, Russia

Correspondence
Nikolai G. Ushakov, Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway.
Email: nikolai.ushakov@ntnu.no

It is well known that sample moments are more sensitive and less robust than order statistics for robustness with respect to outliers. In this article, we show that the situation is exactly the opposite for robustness with respect to rounding. For large and very large sample sizes, statistical procedures based on order statistics become non-applicable even for very mild data rounding while procedures based on sample moments work perfectly for this rounding level. The comparison of sample moments and order statistics is made for tests for normality and tests for exponentiality.

KEYWORDS

Monte Carlo simulation, order statistics, robustness, rounded data, sample moments, test for exponentiality, test for normality

1 | INTRODUCTION

In practice, data for a statistical analysis are always rounded. If the discretization step is not too large compared with the measurement error and the sample size is not too large, then rounding usually does not lead to serious troubles. Such situations have been studied by a number of authors; see, in particular, Hall (1982), Tricker (1984a), Tricker (1984b), Tricker (1990a), Tricker (1990b), Tricker (1990c), Härdle and Scott (1992), Hall and Wand (1996), Tricker and Okell (1997), Minnotte (1998), Meintanis and Ushakov (2004), Bai et al. (2009), Schneeweiss et al. (2010), Ushakov and Ushakov (2018), Zhao and Bai (2020) and references therein.

However, if the sample size is large, then the situation changes. Many statistical procedures, tests and estimators become non-robust with respect to the data rounding. In this paper, we study statistical tests and their robustness with respect to rounding when sample sizes are large. It is known that the sample mean is less robust with respect to outliers than the sample median. Here, we show that the situation is exactly the opposite for robustness with respect to rounding: Statistical tests whose test statistics are based on sample moments are much more robust than tests whose test statistics are based on order statistics. We present some preliminary theoretical analysis and then a simulation study of normality and exponentiality tests based on the sample moments and on the order statistics and find out how well these tests control the probability of Type I error, especially for the case when sample size is large.

Let c be a real number. $[c]$ is used to denote the largest integer less than or equal to c (the integer part of c). The fractional part of c is denoted by $\{c\}$: $\{c\} = c - [c]$. For a positive h, the number $h[c/h]$ is called the integer part modulo h and is denoted by $[c]_h$. The fractional part modulo h is denoted by $\{c\}_h$: $\{c\}_h = c - [c]_h$. In this work, we study rounding to the nearest. Other rounding types are studied similarly. There is a rounding lattice $\{x : x = kh, k = 0, \pm 1, \pm 2, \ldots\}$, where $h > 0$ (call it the discretization step). The rounding of a real number c is the nearest to c point of the rounding lattice. Denote it by $c^{(h)}$. Thus, $c^{(h)} = kh$ where k is such that $kh - h/2 \le c < kh + h/2$. The integer part modulo h is called also the rounding down. For a random variable X, its rounding (or discretization) is the discrete random variable $X^{(h)}$ such that $X^{(h)} = kh$ when $kh - h/2 \le X < kh + h/2$. Note that $[X]_h = kh$ when $kh \le X < (k+1)h$.

The closeness of moments of a random variable and moments of its discretization have been studied in a number of works; see, in particular, Tricker (1984b), Janson (2006), Schneeweiss et al. (2010), Ushakov and Ushakov (2018), Samsonov et al. (2019), Ushakov and Ushakov (2020) and references therein. Shortly, the moments are close provided that the discretization step is not too large.

For our purpose, it is important to take into account not only the accuracy of the discretization (the discretization step) but also the spread of a rounded random variable. A natural measure of the level of rounding is the variable $r=h/\sigma$, where h is the step of discretization and σ is the standard deviation of the random variable.

Throughout the paper, we will accept the following convention. To simplify analysis of simulation results, we will choose a certain threshold value of the probability of Type I error and denote probabilities greater than this threshold by bold. Such values can be considered as unacceptable. Let us choose the threshold to be equal to the double significance level, that is, for example, 0.1 if the significance level is 5%. Of course, this agreement is very conditional and is made only for convenience.

2 | THREE ELEMENTARY EXAMPLES

We start with three very elementary examples.

Example 1. Let $X_1,\ldots,X_n$ be a random sample (iid random variables) from a normal distribution with expectation μ and variance σ^2 (both are unknown). The null hypothesis $H_0:\mu=\mu_0$ is tested versus the alternative $H_1:\mu\neq\mu_0$. The significance level is α. Denote the sample mean and variance by $\overline{X}$ and S^2. It is natural to use the Student t test: The null hypothesis is rejected if

$$\left|\sqrt{n}\frac{\overline{X}-\mu_0}{S}\right|>t_{n-1,\alpha/2}.$$

Suppose however that we do not observe $X_1,\ldots,X_n$. Instead we have rounded observations $X_1^{(h)},\ldots,X_n^{(h)}$. The corresponding sample mean and variance are $\overline{X}'$ and S'^2. Respectively, the null hypothesis is rejected if

$$\left|\sqrt{n}\frac{\overline{X}'-\mu_0}{S'}\right|>t_{n-1,\alpha/2}.$$

Using Monte Carlo simulation, we calculate the empirical probability of Type I error for different rounding levels and different sample sizes and see what happens. To this end, we generate 10,000 samples of each size $n=20,50,100,10^3,10^4,10^5,10^6$ from the normal distribution with prescribed expectation μ_0 and variance σ^2. Let for definiteness $\mu_0=0$ and $\sigma=1$. The observations generated are rounded to different rounding levels $r=0.1,0.01,10^{-3},10^{-4},10^{-5},10^{-6}$ (since $\sigma=1$, r simply coincides with h). For convenience, consider also the case $r=0$ (non-rounded data). The significance level $\alpha=0.05$. For each fixed r and fixed n, the proportion of 10,000 samples for which the hypothesis is rejected is the estimated probability of Type I error. The results are presented in Table 1. The empirical significance level coincides with the theoretical for all levels of rounding and all sample sizes. The rounding did not lead to any problems.

If we use the Wilcoxon signed rank test instead of the t test, then results are similar except for very large sample sizes and very strong rounding. These results are presented in Table 2.

TABLE 1 Simulated probability of Type I error, Example 1, t test

	$r=0$	$r=10^{-6}$	$r=10^{-5}$	$r=10^{-4}$	$r=10^{-3}$	$r=0.01$	$r=0.1$
$n=20$	0.051	0.0494	0.0499	0.0494	0.0496	0.051	0.0514
$n=50$	0.0496	0.0472	0.0511	0.0513	0.0523	0.0484	0.0507
$n=100$	0.0521	0.0511	0.0493	0.0494	0.0514	0.0505	0.0475
$n=10^3$	0.0488	0.0489	0.052	0.0505	0.0486	0.0495	0.0501
$n=10^4$	0.0532	0.0504	0.0485	0.0479	0.0501	0.0547	0.051
$n=10^5$	0.0519	0.0509	0.0492	0.0506	0.051	0.0501	0.0516
$n=10^6$	0.0467	0.047	0.0496	0.0504	0.045	0.0461	0.0541

In the considered example, we used (for the t test) the optimal (optimal for non-rounded data) test based on a suffisient statistic—the sample mean. In the next example, we again use the optimal test based on a sufficient statistic, but the statistic is an order statistic—the maximum.

Example 2. Let $X_1,\ldots,X_n$ be a random sample from a uniform distribution on the interval $[0,\theta],\theta>0$. The problem is to test the hypothesis $H_0:\theta=\theta_0$ against $H_1:\theta>\theta_0$, where θ_0 is a given positive number. The significance level is α. The likelihood ratio test is based on the maximal observation. The null hypothesis is rejected if

$$\max\{X_1,\ldots,X_n\}>\theta_0(1-\alpha)^{\frac{1}{n}}.$$

Like in Example 1, the sample $X_1,\ldots,X_n$ is not observed. Instead we have rounded observations $X_1^{(h)},\ldots,X_n^{(h)}$ and therefore use the test where H_0 is rejected when

$$\max\{X_1^{(h)},\ldots,X_n^{(h)}\}>\theta_0(1-\alpha)^{\frac{1}{n}}.$$

The probability of Type I error is estimated using Monte Carlo simulation absolutely in the same way as in Example 1; that is, 10,000 samples are generated for each sample size. The significance level is $\alpha=0.05$. The parameter $\theta_0=2\sqrt{3}$ (then $\sigma=1$, and $r=h$). The results are presented in Table 3. It is seen that for any rounding level, the test becomes unsuitable for sufficiently large sample sizes.

Let us try to solve the problem of Example 2 using a test statistic which is neither optimal and nor sufficient but is based on a sample moment—the sample mean.

Example 3. The same problem as in the previous example. The critical region has form $\overline{X}>c(n,\theta_0)$. Since it is difficult to find precisely the critical value c, one can approximate it using the Central Limit Theorem. Due to the theorem, the statistic

TABLE 2 Simulated probability of Type I error, Example 1, Wilcoxon test

	$r=0$	$r=10^{-6}$	$r=10^{-5}$	$r=10^{-4}$	$r=10^{-3}$	$r=0.01$	$r=0.1$
$n=20$	0.049	0.046	0.0494	0.0471	0.0507	0.0488	0.0478
$n=50$	0.0485	0.0463	0.05	0.048	0.0509	0.0488	0.0478
$n=100$	0.051	0.05	0.048	0.0534	0.0498	0.0509	0.0519
$n=10^3$	0.0532	0.0521	0.0483	0.0474	0.0518	0.051	0.049
$n=10^4$	0.0478	0.0469	0.0482	0.0473	0.049	0.0479	0.0475
$n=10^5$	0.0498	0.0501	0.0502	0.0514	0.0469	0.0507	0.0707
$n=10^6$	0.0438	0.0522	0.0466	0.0489	0.05	0.0514	**0.2839**

TABLE 3 Simulated probability of Type I error, Example 2, $\alpha=0.05,\theta_0=2\sqrt{3}$

	$r=0$	$r=10^{-6}$	$r=10^{-5}$	$r=10^{-4}$	$r=10^{-3}$	$r=0.01$	$r=0.1$
$n=20$	0.0545	0.0466	0.0528	0.0516	0.0527	0.0523	0.0773
$n=50$	0.0524	0.0525	0.0493	0.0522	0.0448	0.0597	**0.1828**
$n=100$	0.0493	0.0489	0.0485	0.0509	0.0612	**0.1146**	**0.3357**
$n=10^3$	0.05	0.0508	0.0496	0.0618	**0.1567**	**0.6982**	**0.9826**
$n=10^4$	0.0505	0.0529	0.049	**0.1373**	**0.4443**	**1**	**1**
$n=10^5$	0.0554	0.0622	**0.1818**	**0.4643**	**0.9975**	**1**	**1**
$n=10^6$	0.0524	**0.1591**	**0.6502**	**0.9976**	**1**	**1**	**1**

$$\sqrt{n}\frac{\overline{X}-\theta/2}{\theta/\sqrt{12}}$$

has the standard normal distribution asymptotically; therefore, the approximate critical region is

$$\sqrt{n}\frac{\overline{X}-\theta_0/2}{\theta_0/\sqrt{12}} > z_\alpha.$$

Simulation is fulfilled in the same way as in Example 2. In particular, we take $\theta_0 = 2\sqrt{3}$. Simulation results are presented in Table 4. The test works fine: The probability of Type I error is perfectly controlled for all rounding levels and all sample sizes.

Thus, although the test of Example 2 is more powerful than the test of Example 3, the latter is more preferable if data are rounded, and the sample size is large.

3 | SOME ESTIMATES AND PRELIMINARY REMARKS

In this section, we obtain quantitative estimates for the distance between sample moments obtained from rounded and unrounded data and estimates for such a distance between order statistics which shed light on the situation with the examples of the previous section. Although the estimates are obtained under certain restrictions, we believe that they reflect the situation in the general case.

The sample moments are asymptotically normal (as the sample size tends to infinity), and for normally distributed observations, stability with respect to the data rounding is very high provided that the rounding is not too coarse. This follows from the following quantitative estimate.

Theorem 1. *Let $Y_1, Y_2, \ldots$ be a sequence of independent random variables having the same normal distribution with unit variance. Then*

$$P\left(\lim_{n\to\infty}\left|\frac{1}{n}\sum_{i=1}^{n} Y_i^{(h)} - \frac{1}{n}\sum_{i=1}^{n} Y_i\right| < \frac{h}{\pi}\left(1+\frac{h^2}{2\pi^2}\right)e^{-2\pi^2/h^2}\right) = 1. \tag{1}$$

Extension to the case of an arbitrary variance is straightforward. The following theorem gives lower bounds of stability of the minimal and maximal order statistics.

Theorem 2. *Let $Y_1, Y_2, \ldots$ be a sequence of iid random variables with a unimodal bounded density and distribution function $F(x)$. If $F(x) \neq 1$ for all x, then*

$$\liminf_{n\to} P\left(\left|\max\{Y_1^{(h)}, \ldots, Y_n^{(h)}\} - \max\{Y_1, \ldots, Y_n\}\right| \ge \frac{h}{4}\right) \ge \frac{1}{2},$$

and if $F(x) \neq 0$ for all x, then

$$\liminf_{n\to} P\left(\left|\min\{Y_1^{(h)}, \ldots, Y_n^{(h)}\} - \min\{Y_1, \ldots, Y_n\}\right| \ge \frac{h}{4}\right) \ge \frac{1}{2}.$$

TABLE 4 Simulated probability of Type I error, Example 3, $\alpha = 0.05, \theta_0 = 2\sqrt{3}$

	$r = 0$	$r = 10^{-6}$	$r = 10^{-5}$	$r = 10^{-4}$	$r = 10^{-3}$	$r = 0.01$	$r = 0.1$
$n = 20$	0.0442	0.0493	0.0536	0.0498	0.0483	0.0512	0.0539
$n = 50$	0.0485	0.0485	0.0507	0.0526	0.0465	0.0513	0.0492
$n = 100$	0.0497	0.0482	0.0505	0.0574	0.053	0.0507	0.0513
$n = 10^3$	0.0501	0.0527	0.0476	0.0489	0.0484	0.047	0.0528
$n = 10^4$	0.0455	0.0521	0.0517	0.0488	0.0487	0.0491	0.0462
$n = 10^5$	0.0515	0.0488	0.0477	0.0456	0.0491	0.0508	0.0411
$n = 10^6$	0.0491	0.0512	0.0481	0.0536	0.0482	0.0509	0.0327

Similar bounds probably hold for arbitrary order statistics. Proofs of Theorems 1 and 2 are contained in Appendix A1.

The theorems presented in this section as well as the examples given in Section 2 suggest that tests based on the empirical moments must be more stable with respect to the data rounding than tests based on the order statistics. In the next two sections, using Monte Carlo simulation, we find out if this is so.

4 | TESTS FOR NORMALITY

Among existing normality tests, we choose three tests based on the sample moments and four tests based on the order statistics or on the empirical distribution function. Using a simulation study, we check how well these tests control the probability of Type I error for different sample sizes. The tests based on the sample moments are the Jarque–Bera, the kurtosis and the skewness. The tests based on the order statistics and the empirical distribution function are the Kolmogorov–Smirnov, the Cramer–von Mises, the Anderson–Darling and the Pearson. The simulation is performed as follows: 10,000 random samples are generated from a normal distribution for each sample size $n=20, n=50, n=100, n=10^3, n=10^4, n=10^5$. The observations are rounded. Rounding levels are $r=0.1, r=0.01, r=10^{-3}, r=10^{-4}$. For convenience, we provide also simulation results for non-rounded data. All the tests are applied to the same samples. The significance level is 5%. For each sample and each of the seven tests, the hypothesis of normality is rejected if the p value obtained is less than 0.05. For a given test and fixed r and n, the proportion of 10,000 samples for which the hypothesis is rejected is the estimated probability of Type I error. The results are presented in Tables 5–11.

The tables show that the conjecture made in Sections 2 and 3 is fully confirmed. The three tests whose test statistics are based only on the sample moments are much more robust than the other tests if the sample size is large.

5 | TESTS FOR EXPONENTIALITY

The following tests for exponentiality have been selected. The first group: the Epps–Pulley, the Gini, the Atkinon. These tests are based on the sample moments. The second group: the Harris, the Gnedenko, the Epstein. These tests are based on the order statistics. The simulation study is similar to that provided in the previous section; 10,000 random samples are generated from an exponential distribution for each sample size $n=20, n=50, n=100, n=10^3, n=10^4, n=10^5$. The observations are rounded. Rounding levels are $r=0.05, r=0.01, r=10^{-3}, r=10^{-4}$. The results are presented in Tables 12–17. Again tests based on the sample moments are much more robust when the sample size is large.

TABLE 5 Simulated probability of Type I error, $\alpha=0.05$, Jarque—Bera test for normality

	$r=0$	$r=10^{-4}$	$r=10^{-3}$	$r=0.01$	$r=0.1$
$n=20$	0.0483	0.0487	0.0504	0.0514	0.0508
$n=50$	0.0532	0.0522	0.051	0.0482	0.0441
$n=100$	0.0445	0.0506	0.0487	0.0527	0.0502
$n=10^3$	0.0487	0.0495	0.0518	0.049	0.0513
$n=10^4$	0.0477	0.0502	0.0504	0.0488	0.0506
$n=10^5$	0.0527	0.0511	0.0496	0.0523	0.0512

TABLE 6 Simulated probability of Type I error, $\alpha=0.05$, kurtosis test for normality

	$r=0$	$r=10^{-4}$	$r=10^{-3}$	$r=0.01$	$r=0.1$
$n=20$	0.05	0.0484	0.0509	0.0456	0.0464
$n=50$	0.0498	0.0521	0.0512	0.0478	0.0524
$n=100$	0.0481	0.0499	0.0499	0.0509	0.0525
$n=10^3$	0.0489	0.0503	0.0478	0.0516	0.0552
$n=10^4$	0.0501	0.0517	0.0513	0.0592	0.0601
$n=10^5$	0.0469	0.0517	0.0529	0.0701	0.0825

TABLE 7 Simulated probability of Type I error, $\alpha = 0.05$, skewness test for normality

	$r=0$	$r=10^{-4}$	$r=10^{-3}$	$r=0.01$	$r=0.1$
$n=20$	0.0488	0.0489	0.05	0.0502	0.0522
$n=50$	0.0509	0.0511	0.0512	0.052	0.0485
$n=100$	0.05	0.0517	0.0499	0.0475	0.0543
$n=10^3$	0.0503	0.0486	0.0507	0.0505	0.0508
$n=10^4$	0.0497	0.0521	0.0495	0.0503	0.0514
$n=10^5$	0.0523	0.0502	0.051	0.0507	0.0492

TABLE 8 Simulated probability of Type I error, $\alpha = 0.05$, Kolmogorov–Smirnov test for normality

	$r=0$	$r=10^{-4}$	$r=10^{-3}$	$r=0.01$	$r=0.1$
$n=20$	0.0482	0.0454	0.0499	0.0483	0.0551
$n=50$	0.0482	0.0484	0.0501	0.0504	0.0737
$n=100$	0.0487	0.0507	0.0509	0.0532	**0.1222**
$n=10^3$	0.0497	0.0478	0.0515	0.0592	**0.9063**
$n=10^4$	0.0411	0.0383	0.0418	**0.1692**	**1**
$n=10^5$	0.0302	0.0364	0.0532	**0.9529**	**1**

TABLE 9 Simulated probability of Type I error, $\alpha = 0.05$, Cramer–von Mises test for normality

	$r=0$	$r=10^{-4}$	$r=10^{-3}$	$r=0.01$	$r=0.1$
$n=20$	0.0456	0.0513	0.0473	0.0479	0.0576
$n=50$	0.0509	0.0508	0.0476	0.0512	0.0577
$n=100$	0.0476	0.0536	0.0512	0.0499	0.0663
$n=10^3$	0.0514	0.0513	0.0489	0.0528	**0.5316**
$n=10^4$	0.0496	0.0526	0.0498	0.0622	**1**
$n=10^5$	0.0489	0.0479	0.0533	**0.5185**	**1**

TABLE 10 Simulated probability of Type I error, $\alpha = 0.05$, Anderson–Darling test for normality

	$r=0$	$r=10^{-4}$	$r=10^{-3}$	$r=0.01$	$r=0.1$
$n=20$	0.0492	0.0495	0.0507	0.047	0.0523
$n=50$	0.0491	0.0525	0.0523	0.0483	0.0527
$n=100$	0.0495	0.0483	0.051	0.0524	0.066
$n=10^3$	0.0476	0.0522	0.0454	0.0522	**0.4766**
$n=10^4$	0.0483	0.0523	0.0471	0.0604	**1**
$n=10^5$	0.0512	0.05	0.0516	**0.4659**	**1**

6 | POWER OF TESTS FOR ROUNDED DATA

In this section, we find and compare the empirical power of the tests in case of exact data ($r=0$) and in case of rounded data (for definiteness $r=0.1$). We show that in contrast to the probability of Type I error, the probability of Type II error practically is not affected by the rounding. We consider only tests for normality because the power of exponentiality tests for rounded data was in detail studied in Ushakov and Ushakov (2021). The power is studied for three alternative distributions (Laplace, Student and Weibull) and two different sample sizes ($n = 100$, $n = 1000$).

TABLE 11 Simulated probability of Type I error, $\alpha = 0.05$, Pearson test for normality

	$r = 0$	$r = 10^{-4}$	$r = 10^{-3}$	$r = 0.01$	$r = 0.1$
$n = 20$	0.0487	0.0535	0.0514	0.0526	0.059
$n = 50$	0.0524	0.0523	0.0476	0.0556	0.0807
$n = 100$	0.0519	0.0499	0.0507	0.0517	**0.1255**
$n = 10^3$	0.0516	0.049	0.0511	0.0776	**1**
$n = 10^4$	0.0416	0.0473	0.0633	**1**	**1**
$n = 10^5$	0.053	0.0537	**0.9034**	**1**	**1**

TABLE 12 Simulated probability of Type I error, $\alpha = 0.05$, Epps-Pulley test for exponentiality

	$r = 0$	$r = 10^{-4}$	$r = 10^{-3}$	$r = 0.01$	$r = 0.05$
$n = 20$	0.0402	0.0384	0.0399	0.0395	0.0407
$n = 50$	0.0478	0.0434	0.0466	0.0511	0.0457
$n = 100$	0.046	0.0455	0.0468	0.0478	0.049
$n = 10^3$	0.0508	0.0499	0.0518	0.048	0.0509
$n = 10^4$	0.0498	0.0502	0.05	0.0487	0.0494
$n = 10^5$	0.048	0.0501	0.0448	0.0526	0.0595

TABLE 13 Simulated probability of Type I error, $\alpha = 0.05$, Gini test for exponentiality

	$r = 0$	$r = 10^{-4}$	$r = 10^{-3}$	$r = 0.01$	$r = 0.05$
$n = 20$	0.0449	0.0499	0.0482	0.0482	0.0509
$n = 50$	0.0538	0.0494	0.0483	0.047	0.0494
$n = 100$	0.0482	0.0508	0.0463	0.0502	0.0498
$n = 10^3$	0.0481	0.05	0.0536	0.0461	0.049
$n = 10^4$	0.0512	0.0533	0.0513	0.0527	0.0558
$n = 10^5$	0.0528	0.0484	0.0537	0.0516	0.0544

TABLE 14 Simulated probability of Type I error, $\alpha = 0.05$ Atkinson test for exponentiality

	$r = 0$	$r = 10^{-4}$	$r = 10^{-3}$	$r = 0.01$	$r = 0.05$
$n = 20$	0.0348	0.0332	0.0349	0.036	0.0371
$n = 50$	0.0392	0.0409	0.0441	0.0404	0.0442
$n = 100$	0.0483	0.0466	0.043	0.0464	0.0514
$n = 10^3$	0.0489	0.0464	0.0464	0.0482	0.0553
$n = 10^4$	0.0457	0.0511	0.0545	0.0501	0.0558
$n = 10^5$	0.0496	0.0523	0.0475	0.0522	**0.1198**

The alternative distributions have the following parameters. Laplace: zero expectation, the mean absolute deviation $1/\sqrt{2r}$; Student: the variable $X/\sqrt{2r}$, where X has the Student distribution with 4 degrees of freedom; Weibull: the shape parameter 2, the scale parameter $1/\sqrt{1-\pi/4r}$. The choice of parameters is dictated by the level of rounding r when the discretization step is one, that is, when the generated data are rounded to the nearest integer.

From each of the three distributions above, 10,000 random samples of size n are generated. For each sample and each of the 7 tests, the hypothesis of normality is not rejected if the p value obtained is greater than 0.05, and we calculate the proportion of times when the hypothesis is rejected. In other words, we calculate the proportion of times when the p value is less than 0.05. This is the empirical power of the test for the

TABLE 15 Simulated probability of Type I error, $\alpha = 0.05$, Harris test for exponentiality

	$r = 0$	$r = 10^{-4}$	$r = 10^{-3}$	$r = 0.01$	$r = 0.05$
$n = 20$	0.0506	0.0503	0.0506	0.0501	0.0516
$n = 50$	0.0502	0.0515	0.0486	0.049	0.0514
$n = 100$	0.0489	0.0493	0.0501	0.0492	0.0533
$n = 10^3$	0.0506	0.0506	0.0483	0.0525	**0.115**
$n = 10^4$	0.0515	0.0516	0.0484	0.0741	**0.3241**
$n = 10^5$	0.0493	0.0536	0.0516	**0.2502**	**0.999**

TABLE 16 Simulated probability of Type I error, $\alpha = 0.05$, Gnedenko test for exponentiality

	$r = 0$	$r = 10^{-4}$	$r = 10^{-3}$	$r = 0.01$	$r = 0.05$
$n = 20$	0.0477	0.046	0.0464	0.0508	0.0539
$n = 50$	0.0505	0.0492	0.0477	0.0509	0.0504
$n = 100$	0.0471	0.0514	0.0492	0.0585	0.0519
$n = 10^3$	0.0439	0.0473	0.0472	0.0484	0.0771
$n = 10^4$	0.0519	0.0491	0.0495	0.0561	0.0793
$n = 10^5$	0.0496	0.0477	0.0511	**0.1704**	**0.5999**

TABLE 17 Simulated probability of Type I error, $\alpha = 0.05$, Epstein test for exponentiality

	$r = 0$	$r = 10^{-4}$	$r = 10^{-3}$	$r = 0.01$	$r = 0.05$
$n = 20$	0.0448	0.0506	**0.1279**	**0.662**	**0.996**
$n = 50$	0.0451	0.0908	**0.4748**	**0.999**	**1**
$n = 100$	0.0388	**0.249**	**0.9169**	**1**	**1**
$n = 10^3$	0.0308	**1**	**1**	**1**	**1**
$n = 10^4$	0.0163	**1**	**1**	**1**	**1**
$n = 10^5$	**0.5773**	**1**	**1**	**1**	**1**

TABLE 18 Empirical power for non-rounded and rounded data for three distributions ($n = 100$)

	Laplace		Student		Weibull	
	$r = 0$	$r = 0.1$	$r = 0$	$r = 0.1$	$r = 0$	$r = 0.1$
Pearson	0.4743	0.5931	0.3052	0.4511	0.2714	0.3676
Kolmogorov–Smirnov	0.6922	0.8298	0.5117	0.6553	0.3784	0.5611
Cramer–von Mises	0.8152	0.8463	0.6084	0.6757	0.5017	0.5623
Anderson–Darling	0.8136	0.8423	0.6614	0.7231	0.6081	0.6524
Jarque–Bera	0.7823	0.7847	0.7772	0.7906	0.5483	0.5442
Kurtosis	0.8071	0.8112	0.7914	0.8325	0.1781	0.1532
Skewness	0.4167	0.4125	0.5312	0.5146	0.6781	0.6871

given the alternative distribution and given the sample size. The results are presented in Tables 18 and 19. It is seen that either the power for rounded data is (practically) the same as the power for exact data or the power for rounded data is greater than the power for exact data (the Pearson and Kolmogorov–Smirnov tests).

TABLE 19 Empirical power for non-rounded and rounded data for four distributions ($n = 1000$)

	Laplace		Student		Weibull	
	$r = 0$	$r = 0.1$	$r = 0$	$r = 0.1$	$r = 0$	$r = 0.1$
Pearson	1	1	0.9901	1	1	1
Kolmogorov–Smirnov	1	1	1	1	1	1
Cramer–von Mises	1	1	1	1	1	1
Anderson–Darling	1	1	1	1	1	1
Jarque–Bera	1	1	1	1	1	1
Kurtosis	1	1	1	1	0.3812	0.3677
Skewness	0.5013	0.5031	0.7644	0.7582	1	1

7 | CONCLUSION

Statistical tests based on the empirical moments are more robust with respect to data rounding than tests based on the order statistics and the empirical distribution function. Anyway, this is true for tests for normality and exponentiality. With a large sample size and rounded data tests based on the order statistics should not be used.

ACKNOWLEDGEMENT

The authors thank the two anonymous referees for useful comments.

CONFLICT OF INTEREST

No potential conflict of interest was reported by the authors.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in No at https://developer.mozilla.org.

ORCID

Nikolai G. Ushakov https://orcid.org/0000-0002-6521-1664

REFERENCES

Bai, Z., Zheng, S., Zhang, B., & Hu, G. (2009). Statistical analysis for rounded data. *Journal of Statistical Planning and Inference*, *139*, 2526–2542.

Feller, W. (1971). *An Introduction to Probability Theory and its Applications volume 2* (2nd ed.). New York - London - Sydney - Toronto: John Wiley and Sons.

Hall, P. (1982). The influence of rounding errors on some nonparametric estimators of a density and its derivatives. *SIAM Journal of Applied Mathematics*, *42*, 390–399.

Hall, P., & Wand, M. P. (1996). On the accuracy of binned kernel density estimators. *Journal of Multivariate Analysis*, *56*, 165–184.

Härdle, W., & Scott, D. W. (1992). Smoothing by weighted averaging of rounded points. *Computational Statistics*, *7*, 97–128.

Janson, S. (2006). Rounding of continuous random variables and oscillatory asymptotics. *The Annals of Probability*, *34*, 1807–1826.

Meintanis, S., & Ushakov, N. G. (2004). Binned goodness-of-fit tests based on the empirical characteristic function. *Statistics and Probability Letters*, *69*, 305–314.

Minnotte, M. C. (1998). Achieving high-order convergence rates for density estimation with binned data. *Journal of American Statistical Association*, *93*, 663–672.

Samsonov, S. V., Ushakov, N. G., & Ushakov, V. G. (2019). Estimation of the second moment based on rounded data. *Journal of Mathematical Sciences*, *237*, 819–825.

Schneeweiss, H., Komlos, J., & Ahmad, A. S. (2010). Symmetric and asymmetric rounding: a review and some new results. *AStA Advances in Statistical Analysis*, *94*, 247–271.

Tricker, A. R. (1984a). Effects of rounding data sampled from the exponential distribution. *Journal of Applied Statistics*, *11*, 54–87.

Tricker, A. R. (1984b). Effects of rounding on the moments of a probability distribution. *The Statistician*, *33*, 381–390.

Tricker, A. R. (1990a). The effect of rounding on the significance level of certain normal test statistics. *Journal of Applied Statistics*, *17*, 31–38.

Tricker, A. R. (1990b). The effect of rounding on the power level of certain normal test statistics. *Journal of Applied Statistics*, *17*, 219–228.

Tricker, A. R. (1990c). The effect of rounding on the significance level and power of certain test statistics for non-normal data. *Journal of Applied Statistics*, *17*, 329–340.

Tricker, A. R., & Okell, E. (1997). The effect of rounding on sequential and fixed size sample hypothesis tests. *Communications in Statistics. Simulation and Computation*, *26*, 1413–1429.

Ushakov, N. G., & Ushakov, V. G. (2018). Statistical analysis of rounded data: Measurement errors vs rounding errors. *Journal of Mathematical Sciences*, *234*, 770–773.

Ushakov, N. G., & Ushakov, V. G. (2020). Accuracy of estimating the mean from rounded data. *Journal of Mathematical Sciences*, *246*, 565–568.

Ushakov, N. G., & Ushakov, V. G. (2021). On sensitivity of exponentiality tests to data rounding: A Monte Carlo simulation study. Communications in Statistics - Simulation and Computation. Ahead-of-print, 1-10. https://doi.org/10.1080/03610918.2021.2009868

Zhao, N., & Bai, Z. (2020). Bayesian statistical inference based on rounded data. *Communications in Statistics. Simulation and Computation*, *49*, 135–146.

How to cite this article: Ushakov, N. G., & Ushakov, V. G. (2022). On the effect of rounding on hypothesis testing when sample size is large. *Stat*, *11*(1), e478. https://doi.org/10.1002/sta4.478

APPENDIX

In this appendix, we give proofs of Theorems 1 and 2.

Lemma 1. Let Y be an absolutely continuous random variable with the probability density function $f(y)$ and the characteristic function $\varphi(t)$. If $\varphi(t)$ is absolutely integrable, then

$$\mathrm{E}\{Y\}_h = h\left(\frac{1}{2} - \sum_{n=1}^{\infty} \frac{\Im\varphi(2\pi n/h)}{\pi n}\right).$$

Proof. We have

$$\mathrm{E}\{Y\}_h = \int_0^h y \sum_{n=-\infty}^{\infty} f(y+nh)dy.$$

Due to the Poisson summation formula (see Feller, (1971), p. 632, formula 5.9 with $\zeta = 0, \lambda = \pi$),

$$\sum_{n=-\infty}^{\infty} f(y+nh) = \frac{1}{h} \sum_{k=-\infty}^{\infty} \varphi(2\pi k/h) e^{-iy2\pi k/h};$$

therefore,

$$\begin{aligned}\mathrm{E}\{Y\}_h = & \ h \sum_{k=-\infty}^{\infty} \varphi(2\pi k/h) \int_0^1 y e^{-iy2\pi k} dy = h\left(\frac{1}{2} + \sum_{k \neq 0} \varphi(2\pi k/h) \frac{i}{2\pi k}\right) = \\ = & \ h\left(\frac{1}{2} - \sum_{k=1}^{\infty} \frac{\Im\varphi(2\pi k/h)}{\pi k}\right)\end{aligned}$$

(we have used the fact that the real part of any characteristic function is even and the imaginary part is odd).

Proof of Theorem 1. Denote the expectation of Y_i by μ. Since

$$\begin{aligned}& \frac{1}{n} \sum_{i=1}^{n} Y_i^{(h)} \overset{a.s.}{\longrightarrow} \mathrm{E}\left[Y_i + \frac{h}{2}\right]_h = \\ = & \ \mathrm{E}\left(Y_i + \frac{h}{2}\right) - \mathrm{E}\left\{Y_i + \frac{h}{2}\right\}_h = \mu + \frac{h}{2} - \mathrm{E}\left\{Y_i + \frac{h}{2}\right\}_h,\end{aligned}$$

the limit in Theorem 1 is equal to $|h/2 - \mathrm{E}\{Y_i + h/2\}_h|$. Denote $t_n = 2\pi n/h$. Then, using Lemma 1, we obtain

$$\left|\frac{h}{2} - \mathrm{E}\left\{Y_i + \frac{h}{2}\right\}_h\right| = \left|h \sum_{n=1}^{\infty} \frac{e^{-t_n^2/2} \sin((\mu + h/2)t_n)}{\pi n}\right| \leq \frac{h}{\pi} \sum_{n=1}^{\infty} \frac{e^{-t_n^2/2}}{n}.$$

Estimate the sum in the right hand side:

$$\sum_{n=1}^{\infty}\frac{e^{-t_n^2/2}}{n}=e^{-2\pi^2/h^2}+\sum_{n=2}^{\infty}\frac{e^{-2\pi^2n^2/h^2}}{n}<e^{-2\pi^2/h^2}+\sum_{n=2}^{\infty}e^{-2\pi^2n/h^2}<$$
$$<e^{-2\pi^2/h^2}+\int_1^{\infty}e^{-2\pi^2x/h^2}dx=e^{-2\pi^2/h^2}+\frac{h^2}{2\pi^2}e^{-2\pi^2/h^2}.$$

Thus,

$$\left|\frac{h}{2}-\mathrm{E}\left\{Y_i+\frac{h}{2}\right\}_h\right|<\frac{h}{\pi}\left(1+\frac{h^2}{2\pi^2}\right)e^{-2\pi^2/h^2}$$

that implies the theorem.

The concentration function of a random variable X is defined as

$$Q(X\,;\,l)=\sup_x P(x\ \leq\ X\ \leq\ x+l),\ l\geq 0.$$

Lemma 2. Let X have an absolutely continuous unimodal distribution with the probability density function $f(x)$. If $f(x)\leq A$, then

$$P\left(|X^{(h)}-X|\geq\frac{h}{4}\right)\geq\frac{1}{2}(1-Ah).$$

Proof. Let k be an arbitrary integer. Then

$$|X^{(h)}-X|<\frac{h}{4}$$

if

$$X\in\left(hk-\frac{h}{4},\,hk+\frac{h}{4}\right)$$

and

$$|X^{(h)}-X|\geq\frac{h}{4}$$

if

$$X\in\left[hk-\frac{h}{2},\,hk-\frac{h}{4}\right]\cup\left[hk+\frac{h}{4},\,hk+\frac{h}{2}\right],$$

that is,

$$|X^{(h)}-X|\geq\frac{h}{4}$$

if

$$X\in\left[hk+\frac{h}{4},\,hk+\frac{3h}{4}\right],\ k=0,\pm1,\pm2,...$$

Let $f(x)$ attain its maximum at x_0. Denote by k_0 the integer such that

$$x_0 \in \left[hk_0 - \frac{h}{2}, hk_0 + \frac{h}{2}\right).$$

Then (to the right of the mode)

$$P\left(X \in \left[hk_0 + \frac{h}{4},\ h(k_0+1) - \frac{h}{4}\right]\right) \geq P\left(X \in \left[h(k_0+1) - \frac{h}{4},\ h(k_0+1)\right]\right),$$
$$P\left(X \in \left[hk + \frac{h}{4},\ hk + \frac{3h}{4}\right]\right) \geq P\left(X \in \left[h(k+1) - \frac{h}{4},\ h(k+1) + \frac{h}{4}\right]\right)$$

for $k = k_0+1, k_0+2, \ldots$, and (to the left of the mode)

$$P\left(X \in \left[hk_0 - \frac{3h}{4},\ hk_0 - \frac{h}{4}\right]\right) \geq P\left(X \in \left[h(k_0-1),\ h(k_0-1) + \frac{h}{4}\right]\right),$$
$$P\left(X \in \left[hk - \frac{3h}{4},\ hk - \frac{h}{4}\right]\right) \geq P\left(X \in \left[h(k-1) - \frac{h}{4},\ h(k-1) + \frac{h}{4}\right]\right)$$

for $k = k_0 - 1, k_0 - 2, \ldots$.

Thus,

$$\begin{aligned}
&\sum_{k=-\infty}^{\infty} P\left(X \in \left[hk + \frac{h}{4},\ hk + \frac{3h}{4}\right]\right) = \\
&= \sum_{k=-\infty}^{\infty} P\left(X \in \left[hk - \frac{h}{2},\ hk - \frac{h}{4}\right]\right) + \sum_{k=-\infty}^{\infty} P\left(X \in \left[hk + \frac{h}{4},\ hk + \frac{h}{2}\right]\right) \geq \\
&\geq P\left(X \in \left[h(k_0+1) - \frac{h}{4},\ h(k_0+1)\right]\right) + P\left(X \in \left[h(k_0-1),\ h(k_0-1) + \frac{h}{4}\right]\right) + \\
&+ \sum_{k=-\infty}^{k_0-1} P\left(X \in \left[h(k-1) - \frac{h}{4},\ h(k-1) + \frac{h}{4}\right]\right) + \\
&+ \sum_{k=k_0+1}^{\infty} P\left(X \in \left[h(k+1) - \frac{h}{4},\ h(k+1) + \frac{h}{4}\right]\right).
\end{aligned}$$

Denote the expression in the left hand side by S and the expression in the right hand side by R. We have

$$S \geq R, \tag{A1}$$

and

$$\begin{aligned}
S + R + P\left(X \in \left[hk_0 - \frac{h}{4},\ hk_0 + \frac{h}{4}\right]\right) + P\left(X \in \left[h(k_0+1),\ h(k_0+1) + \frac{h}{4}\right]\right) + \\
+ P\left(X \in \left[h(k_0-1) - \frac{h}{4},\ h(k_0-1)\right]\right) = 1.
\end{aligned}$$

The sum of the three last probabilities in the left hand side does not exceed the concentration function $Q(X;h)$; therefore,

$$S + R + Q(X;h) \geq 1. \tag{A2}$$

From (A1) and (A2), we obtain

$$S \geq \frac{1}{2}(1 - Q(X;h)) \geq \frac{1}{2}(1 - Ah).$$

Thus, finally,

$$P\left(|X^{(h)} - X| \geq \frac{h}{4}\right) \geq S \geq \frac{1}{2}(1 - Ah).$$

Lemma 3. Let $X_1,\ldots,X_n$ be iid random variables with the unimodal bounded density $f(x)$ and the distribution function $F(x)$. Denote the probability density function of $X_{(n)} = \max\{X_1,\ldots,X_n\}$ by $f_{\max}(x)$ and the density of $X_{(1)} = \min\{X_1,\ldots,X_n\}$ by $f_{\min}(x)$.

If $F(x) \neq 1$ for all x, then there exist two constants C_1 and Δ_1 ($0 < \Delta_1 < 1$), depending on f but not depending on n, such that

$$f_{\max}(x) \leq C_1 \Delta_1^n. \tag{A3}$$

If $F(x) \neq 0$ for all x, then there exist two constants C_2 and Δ_2 ($0 < \Delta_2 < 1$), depending on f but not depending on n, such that

$$f_{\min}(x) \leq C_2 \Delta_2^n. \tag{A4}$$

Proof. We have

$$f_{\max}(x) = nF^{n-1}(x) f(x).$$

Without loss of generality, suppose that $f(x)$ is differentiable and that its mode is equal to zero. Denote the mode of the density $f_{\max}(x)$ by x_n^*. x_n^* is a solution of the equation

$$(n-1)f^2(x) + F(x)f'(x) = 0. \tag{A5}$$

The first summand in the left hand side of the equation is positive. The second summand is positive for $x < 0$ and negative for $x > 0$. Since $F(0)f'(0) = 0$ and $\lim_{x\rightarrow} F(x)f'(x) = 0$, there exists the point x_0 where $F(x)f'(x)$ attains its minimum.

The first summand in the left hand side of (A5) increases as n increases while the second summand does not change; therefore, $F(x_n^*)f'(x_n^*)$ decreases in n, and the sequence $\{x_n^*\}$ increases while remaining less than or equal to x_0. Hence,

$$f_{\max}(x) \leq nF^{n-1}(x_n^*) f(x_n^*) \leq nF^{n-1}(x_0) f(0).$$

This inequality implies (A3). Inequality (A4) is obtained similarly.

Proof of Theorem 2. Denote the probability density function of $\max\{Y_1,\ldots,Y_n\}$ by $f_{\max}(x)$. Then, due to Lemma 2,

$$\begin{aligned} &P\left(\left|\max\{Y_1^{(h)}, \ldots, Y_n^{(h)}\} - \max\{Y_1, \ldots, Y_n\}\right| \geq \frac{h}{4}\right) = \\ &= P\left(\left|(\max\{X_1, \ldots, X_n)^{(h)}\} - \max\{X_1, \ldots, X_n\}\right| \geq \frac{h}{4}\right) \geq \frac{1}{2}(1 - A_n h), \end{aligned} \tag{A6}$$

where $A_n = \max_x f_{\max}(x)$. But, due to Lemma 2,

$$\lim_{n\rightarrow\infty} A_n = 0. \tag{A7}$$

(A6) and (A7) imply the theorem.