Psicológica (2013), 34, 407-424.

The adequacy of different robust statistical tests in comparing two independent groups

Maribel Peró-Cebollero[*] and Joan Guàrdia-Olmos

Universitat de Barcelona

In the current study, we evaluated various robust statistical methods for comparing two independent groups. Two scenarios for simulation were generated: one of equality and another of population mean differences. In each of the scenarios, 33 experimental conditions were used as a function of sample size, standard deviation and asymmetry. For each condition, 5000 replications per group were generated. The results obtained by this study show an adequate type error I rate but not a high power for the confidence intervals. In general, for the two scenarios studied (mean population differences and not mean population differences) in the different conditions analysed, the Mann-Whitney U-test demonstrated strong performance, and a little worse the t-test of Yuen-Welch.

In social sciences, and particularly in psychology, many of the applied research studies use parametric statistical tests to evaluate their expectations or hypotheses. However, in most cases, the adequacy of the use of those tests is not assessed, and the use of those tests is often of dubious validity because the assumptions of the statistical test are violated. A clear example is the assumption of normal distribution, which is often assumed, although observed distributions do not usually follow a normal distribution. In recent years, there have been increasing numbers of studies that pay attention to the assumptions of statistical tests, and researchers do not use parametric tests indiscriminately, but they use nonparametric tests or even perform

[*] This research was made possible by the PSI2010-21214-C02-01 project and was carried out by members of the Generalitat de Catalunya's SGR 388 Consolidated Research Group. Specials Thanks to the reviewers and editor of Psicologica, for their comments and contributions in order to improve this paper. Correspondence autor: Maribel Peró-Cebollero. Departament de Metodologia de les Ciències del Comportament. Facultat de Psicologia. Universitat de Barcelona. Institut de Recerca en Cervell, Cognició i Conducta (IR3C). Passeig de la Vall d'Hebron, 171. 08035 Barcelona, Spain. E mail: mpero@ub.edu

logarithmic or power transformations on the original variables to obtain new distributions that follow the normal distribution. Some researchers use the Monte Carlo permutation test when they work with small samples and with variables that do not follow a normal distribution, but the resultant performance is not as good as desired (Holmes-Finch and Davenport, 2009). On the other hand, some researchers defend the use of the Bayesian approach in statistical decision-making (De la Fuente, Cañadas, Guàrdia, and Lozano, 2009), although this option is far from being applied to psychological research.

However, it is necessary to mention that in recent years, in the field of psychology, there has been an effort to encourage researchers to provide more information than the *p*-value when they present the results of their research. Therefore, it is becoming advisable to provide, in addition to the statistical test and the *p*-value, other statistical measures, such as the effect size, confidence intervals around the statistical test, and even statistical confidence intervals for effect size (APA, 2010; Bailar and Mosteller, 1988; Belia, Fidler, Williams and Cumming, 2005; Cohen, 1994; Cumming, 2008, 2009; Cumming and Fidler, 2009; Cumming and Finch, 2001, 2005; Wilkinson and the Task Force on Statistical Inference, 1999; Wolfe and Hanley, 2002) or robust estimations of the confidence intervals of the robust effect size (Algina, Keselman and Penfield, 2005; Keselman, Algina, Lix, Wilcox and Deering, 2008). In fact, in 1934, J. Neyman proposed the use of confidence intervals in statistical decision-making (Cowles, 1989). In the same manner, Hagen (1997) or Tryon (2001) indicated that confidence intervals contribute the same information as the hypothesis test, and Coulson, Healey, Fidler and Cumming (2010), Cumming (2008, 2009), Cumming and Fidler (2011) and Cumming and Maillardet (2006) remarked that confidence intervals support the inference without the need to formulate a null hypothesis.

However, few works incorporate the use of the median confidence interval (Bonett and Price, 2002; Dubnicka, 2007; Lin, Newcombe, Lipsitz and Carter, 2009; Strelen, 2001, 2004; Wilcox, 2005; Woodruff, 1952). Bonett and Price (2002) indicate that the decision around the median is adequate when the distribution of the quantitative variable is biased and leptokurtic and the variable distribution does not fit the normal distribution; in those situations the analysis based on the mean is not always robust. Several works (Peró, Delgado and Guàrdia, 2011; Peró, Guàrdia, Freixa and Turbany, 2008) show a good performance for the median confidence intervals comparison to recognize the conditions around the true H_0 when comparing two independent groups, but a bad performance in the conditions around the false H_0.

The main aim of this work is to study the benefit of confidence interval comparison for two independent groups when the distributions are asymmetrical. We compare the confidence intervals around the mean and the confidence intervals around the median to determine which analysis provides the best results in the comparison. For this comparison, we obtain the confidence intervals around the mean, the trimmed mean (Wilcox, 2005) and the median.

The confidence interval for the trimmed mean was obtained applying the following expression:

$$\bar{x}_t \pm t_{(\alpha/2,\nu)} \frac{S_w}{(1-2\gamma)\sqrt{n}}$$

where $\bar{x}_t$ symbolizes the trimmed mean, S_w the winsorized standard deviation and γ the proportion of trimming in each tail.

In the case of the median, we obtain the confidence interval using five different methods: the standard error method (Kendall, 1945; Mothes and Torres-Ibern, 1970), the binomial median confidence interval (Bland, 2003; DeCoster and Burchill, 2000), the McKean & Schraeder method, the Maritz & Jarret method and the adaptive kernel estimation to obtain the median confidence interval (Wilcox, 2005).

According to Kendall (1945) and Mothes and Torrens-Ibern (1970), if the population is normal with mean μ, standard deviation σ and a sufficiently great sample, the median probability distribution tends to be a normal distribution with the following characteristics:

$$E[median] = \mu \qquad VAR[median] = \frac{\pi}{2}\frac{\sigma^2}{n} \qquad SE(median) = 1.253\,\sigma/\sqrt{n}$$

Consequently, the interval is obtained by calculating the following formula:

$$Md \pm t_{(\alpha,\nu)}\ 1.253\ \frac{\sigma}{\sqrt{n}}$$

In the case of small samples (which are common in the social sciences), the sampling distribution of the median is not known, even though authors such as Lane (1999) posit that the sampling distribution of the median is normal if the variable in the original population is distributed normally.

The second method (the calculation of the median confidence interval from the binomial distribution) is based on the positions of the lower limit and the upper limit of the interval from the application of the binomial distribution, given that the number of observations below the percentile *k* follows this distribution with parameters *n* and *k* (Bland, 2003, DeCoster and Burchill, 2000) and the median is the central point of distribution *k* = .5. It is defined by the following parameters:

$$\text{Position that occupies the value of the median: } \frac{n+1}{2}$$

$$\text{Standard error: } \sqrt{n \;\; p \;\; (1-p)}, \quad \textit{where } p = .5$$

Consequently, the interval for the positions is obtained by calculating the following formula:

$$\frac{n+1}{2} \pm t_{(\alpha,\nu)} \;\; \sqrt{n \;\; p \;\; (1-p)}$$

Then, the positions are rounded off to the next whole number, and finally, the values of the observed distribution that occupy these positions are obtained.

The equation proposed by McKean and Schraeder's (1984) for obtaining the median's confidence interval from the estimation of the median standard error is (Wilcox, 2005):

$$SE = \left(\frac{x_{(n-k-1)} - x_k}{2\, z_{.995}}\right), \qquad \text{where } k = \frac{n-1}{2} - z_{.995}\sqrt{\frac{\mathrm{n}}{4}}$$

Maritz and Jarret's (1978) equation for obtaining the median's confidence interval from the estimation of the proposed standard error is (Wilcox, 2005):

$$\lambda\, x_{k+1} + (1-\lambda)\, x_k \;\div\; \lambda\, x_{n-k} + (1-\lambda)\, x_{n-k+1}$$

$$\text{Where:} \quad I = \frac{\gamma_k - 1 - \alpha}{\gamma_k - \gamma_{k+1}}, \quad \lambda = \frac{(n-k)\, I}{k + (n-2k)\, I}$$

Finally, the adaptive kernel estimation consist in compute a confidence interval for the median using an estimate of the standard error based on adaptive kernel density estimator, a good approximation of the true density with small samples (Wilcox, 2005).

METHOD

Procedure. The data analysed in this study were generated with the use of *R* software (R Development Core Team, 2010) under the assumption of a two independent groups design. In particular, two possible scenarios were simulated: one in which the population means were equal (100) and one with different population means (100 and 115, respectively). Both scenarios were simulated with equal population variances in the two groups

(100). For each scenario, 9 possible experimental conditions were studied with varying sample size values of each group (n = 10, n = 30 or n = 50 but with both groups having the same sample size) and asymmetrically generated distributions (g_x = .0, g_x = .8, g_y = .0 or g_y = .8; the sub-index "x" refers to one group, and the sub-index "y" refers to the other group; for equal means the condition of g_x = .8, g_y = .0 was not studied). Also, for both scenarios, another 24 experimental conditions were studied, different sample size (n_x = 10, n_y = 30 and n_x = 30, n_y = 10), equal or different variances (σ_x = σ_y = 10, σ_x = 5, σ_y = 10 and σ_x = 10, σ_y = 5) and asymmetrically generated distributions (g_x = .0, g_x = .8, g_y = .0 or g_y = .8).

In the simulation of each condition, 5000 replications for each group were generated according to the model of the standard normal distribution [R code: `x[i,]<-sort(rnorm(10))` `y[i,]<-sort(rnorm(10)),` we change the value 10 by 30 or 50 to simulate the different sample sizes worked]. In a second step, the asymmetry in the distribution was generated by applying the formula of the distribution *gh* (Field and Genton, 2006 or Hoaglin, 1985):

$$\frac{e^{gz}-1}{g}\,e^{h\,z^2/2}$$

where *g* indicates the asymmetry that can be generated, and *h* indicates the kurtosis that can be generated in the normal distribution. In both cases, a 0 would indicate a perfectly symmetrical and mesokurtic distribution (for all the simulations h was fixed to 0). As the parameters move away from 0, the asymmetry increases and the kurtosis of the curve increases or decreases (Field and Genton, 2006; Hoaglin, 1985 or Wilcox, 2005). Asymmetry was generated in the two comparison groups (the "x" group and the "y" group for both scenarios). Finally, the third step in the simulation consisted of multiplying the distribution by the value of standard deviation and adding the mean to obtain a distribution with a mean of 100 or 115 and an original standard deviation of 10 or 5.

In tables 1 and 2 we show some descriptive indicators of the simulated distributions for all the experimental conditions generated.

For each simulation, we compute Student's t-test of independent groups (t-test), Yuen-Welch's t-test (Wilcox, 2005), the nonparametric Mann-Whitney U-test, the mean confidence intervals, the trimmed mean confidence interval, the medians confidence interval from the standard error (Kendall, 1945), the binomial median confidence intervals, the medians confidence interval based on the standard error from the McKean and Schraeder method (1984), the Marizt and Jarret method (1978) and the adaptive-kernel density estimation (Wilcox, 2005).

Table 1. Mean and standard deviation (in brackets) for the mean, median, trimmed mean, standard deviation and asymmetry of the empirical distribution for the different simulations generated (equal sample size).

Simulation	$n = 10$					$n = 30$					$n = 50$				
	Mean	Md	T mean	SD	As	Mean	Md	T mean	SD	As	Mean	Md	T mean	SD	As
$\mu_x = 100$, $\sigma_x = 10$; $g_x = 0$	99.98	99.98	99.98	9.70	- 0.0002	99.98	99.97	99.99	9.92	- 0.0050	99.99	100.02	100.01	9.93	- 0.0074
	(3.18)	(3.71)	(3.36)	(2.27)	(0.69)	(1.80)	(2.22)	(1.92)	(1.32)	(0.43)	(1.42)	(1.77)	(1.52)	(1.01)	(0.34)
$\mu_x = 100$, $\sigma_x = 10$, $g_x = .8$	104.66	100.66	101.48	14.02	1.2612	104.68	100.18	101.08	15.14	1.9098	104.68	100.16	101.03	15.47	2.1931
	(5.14)	(4.02)	(3.92)	(8.08)	(0.81)	(2.97)	(2.28)	(2.14)	(6.21)	(0.89)	(2.29)	(1.80)	(1.67)	(4.82)	(0.99)
$\mu_y = 100$, $\sigma_y = 10$, $g_y = 0$	99.97	99.97	99.97	9.75	- 0.0061	100.04	100.03	100.03	9.89	0.0086	100.00	100.00	100.00	9.96	0.0007
	(3.11)	(3.64)	(3.31)	(2.35)	(0.69)	(1.80)	(2.21)	(1.92)	(1.31)	(0.42)	(1.42)	(1.74)	(1.52)	(0.99)	(0.34)
$\mu_y = 100$, $\sigma_y = 10$, $g_y = .8$	104.68	100.63	101.46	14.08	1.2653	104.74	100.24	101.12	15.24	1.9319	104.73	100.13	101.01	15.59	2.2062
	(5.11)	(3.99)	(3.91)	(8.16)	(0.80)	(2.92)	(2.27)	(2.14)	(5.69)	(0.89)	(2.31)	(1.77)	(1.68)	(4.86)	(0.96)
$\mu_y = 115$, $\sigma_y = 10$, $g_y = 0$	114.97	114.97	114.97	9.75	- 0.0061	115.04	115.03	115.03	9.89	0.0086	115.00	115.00	115.00	9.96	0.0007
	(3.11)	(3.64)	(3.31)	(2.35)	(0.69)	(1.80)	(2.21)	(1.92)	(1.31)	(0.42)	(1.42)	(1.74)	(1.52)	(0.99)	(0.34)
$\mu_y = 115$, $\sigma_y = 10$, $g_y = .8$	119.68	115.63	116.46	14.08	1.2653	119.74	115.24	116.12	15.24	1.9319	119.73	15.13	116.01	15.59	2.2062
	(5.11)	(3.99)	(3.91)	(8.16)	(0.80)	(2.92)	(2.27)	(2.14)	(5.69)	(0.89)	(2.31)	(1.77)	(1.68)	(4.86)	(0.96)

μ_x : population mean generated in group "x"; σ_x : population standard deviation generated in group "x"; g_x : asymmetry generated in group "x"; μ_y : population mean generated in group "y"; σ_y : population standard deviation generated in group "y"; g_y : asymmetry generated in group "y"; Md: median; T mean: 25% trimmed mean; SD: standard deviation; As: asymmetry.

Table 2. Mean and standard deviation (in brackets) for the mean, median, trimmed mean, standard deviation and asymmetry of the empirical distribution for the different simulations generated (different sample size).

Simulation	$n_x = 10$ $n_y = 30$					$n_x = 30$ $n_y = 10$				
	Mean	Md	T mean	SD	As	Mean	Md	T mean	SD	As
$\mu_x = 100$, $\sigma_x = 5$, $g_x = 0$	99.99	99.98	99.98	4.86	0.0103	99.99	99.98	99.99	4.97	0.0019
	(1.58)	(1.85)	(1.67)	(1.14)	(.68)	(0.91)	(1.12)	(0.97)	(0.65)	(0.42)
$\mu_x = 100$, $\sigma_x = 5$, $g_x = .8$	102.35	100.32	100.74	7.05	1.2621	102.35	100.09	100.54	7.62	1.9197
	(2.60)	(2.00)	(1.96)	(4.13)	(0.80)	(1.48)	(1.14)	(1.07)	(2.84)	(0.89)
$\mu_x = 100$, $\sigma_x = 10$, $g_x = 0$	99.99	99.96	99.97	9.73	0.0103	99.98	99.96	99.98	9.94	0.0019
	(3.16)	(3.69)	(3.33)	(2.27)	(0.68)	(1.82)	(2.24)	(1.94)	(1.30)	(0.42)
$\mu_x = 100$, $\sigma_x = 10$, $g_x = .8$	104.71	100.63	101.48	14.11	1.2621	104.71	100.18	101.08	15.24	1.9197
	(5.20)	(3.99)	(3.91)	(8.27)	(0.80)	(2.97)	(2.29)	(2.14)	(5.69)	(0.89)
$\mu_y = 100$, $\sigma_y = 5$, $g_y = 0$	100.01	100.01	100.01	4.97	-0.0030	100.00	100.01	100.01	4.88	-0.0095
	(0.90)	(1.11)	(0.96)	(0.65)	(0.42)	(1.59)	(1.86)	(1.69)	(1.18)	(0.68)
$\mu_y = 100$, $\sigma_y = 5$, $g_y = .8$	102.38	100.11	100.56	7.64	1.8999	102.37	100.36	100.78	7.03	1.2481
	(1.49)	(1.14)	(1.06)	(3.00)	(0.89)	(2.61)	(2.02)	(1.97)	(4.10)	(0.80)
$\mu_y = 115$, $\sigma_y = 5$, $g_y = 0$	115.01	115.01	115.01	4.97	-0.0030	115.00	115.01	115.01	4.88	-0.0095
	(0.90)	(1.11)	(.096)	(0.65)	(0.42)	(1.59)	(1.86)	(1.69)	(1.18)	(0.68)
$\mu_y = 115$, $\sigma_y = 5$, $g_y = .8$	117.38	115.11	115.56	7.64	1.8999	117.37	115.36	115.78	7.03	1.2481
	(1.49)	(1.14)	(1.06)	(3.00)	(0.89)	(2.61)	(2.02)	(1.97)	(4.10)	(0.80)
$\mu_y = 100$, $\sigma_y = 10$, $g_y = 0$	100.03	100.01	100.03	9.93	-0.0030	100.00	100.02	100.02	9.76	-0.0095
	(1.80)	(2.22)	(1.92)	(1.30)	(0.42)	(3.18)	(3.73)	(3.37)	(2.37)	(0.68)
$\mu_y = 100$, $\sigma_y = 10$, $g_y = .8$	104.76	100.22	101.12	15.28	1.8999	104.75	100.71	101.56	14.06	1.2481
	(2.97)	(2.28)	(2.13)	(5.99)	(0.89)	(5.22)	(4.04)	(3.94)	(8.21)	(0.80)
$\mu_y = 115$, $\sigma_y = 10$, $g_y = 0$	115.03	115.01	115.03	9.93	-0.0030	115.00	115.02	115.02	9.76	-0.0095
	(1.80)	(2.22)	(1.92)	(1.30)	(0.42)	(3.18)	(3.73)	(3.37)	(2.37)	(0.68)
$\mu_y = 115$, $\sigma_y = 10$, $g_y = .8$	119.76	115.22	116.13	15.28	1.8999	119.75	115.71	116.56	14.06	1.2481
	(2.97)	(2.28)	(2.13)	(5.99)	(0.89)	(5.22)	(4.04)	(3.94)	(8.21)	(0.80)

μ_x : population mean generated in group "x"; σ_x : population standard deviation generated in group "x"; g_x : asymmetry generated in group "x"; μ_y : population mean generated in group "y"; σ_y : population standard deviation generated in group "y"; g_y : asymmetry generated in group "y"; Md: median; T mean: 25% trimmed mean; SD: standard deviation; As: asymmetry.

Data analysis. The data analysis consisted of obtaining the percentage of errors in the statistical decision using the t-test of the independent groups (according with the homocedasticity condition) for the 5000 comparisons in the 33 conditions for both scenarios, the nonparametric Mann-Whitney U-test, Yuen-Welch's t-test (Wilcox, 2005) and the comparison of the confidence intervals of means and medians in the seven used procedures. The significance level was set at an alpha of 5% for all hypotheses test. The statistical decisions of the compared confidence intervals were made according to two criteria: one non-strict and one strict. According to the first procedure, the null hypothesis was rejected if the statistics (mean or median) of the first group ("x") were not within the generated confidence interval in the second group ("y"), and the statistics of the second group were not within the generated confidence interval in the first group. As for the second procedure, the null hypothesis was rejected in the case of no overlap between the two confidence intervals generated. Finally, for each condition and each scenario, the percentage of replications with an incorrect decision was obtained.

RESULTS

We show the results obtained from the generated simulations in Tables 3 to 6. In Tables 3 and 4 we show the results for the scenario of no mean population differences, and in Tables 5 and 6 we show the results for the scenario of mean population differences.

We can see in Table 3 that when we work with samples sizes of 30 or 50 subjects in each group and generate asymmetry in "y" group, classical statistical tests, such as the t-test for independent groups and the nonparametric Mann-Whitney U-test, have a high error rate in their decisions. In fact, this error is greater than 20% in the case of the t-test and over 10% in the case of the nonparametric Mann-Whitney U-test. It should be noted that in the case of Yuen-Welch's t-test, the error rate is around the nominal alpha of 5%. If you look at the decision from the comparison of confidence intervals following the non-strict decision criterion (mean or median not inside the confidence intervals computed in each group), we can say that the strategy that presents fewer incorrect decisions is the comparison made for the binomial median confidence intervals. In the case of median confidence intervals computed from McKean and Schraeder's estimation, Marizt and Jarret's estimation or kernel adaptive estimation, it is necessary to note that when the sample size increases, the percentage of incorrect decisions also increases, and this percentage is approximately 10%

when we work with 50 subjects in each group. Finally, we would to comment that if the decision is taken based on the comparison of mean confidence intervals, either from the mean or the trimmed mean, the percentage of incorrect decisions is higher than when comparing median confidence intervals, regardless of the method used to estimate the median confidence interval.

When we use the decision strict criterion for comparing the two confidence intervals (non-overlapping confidence intervals), if the decision is taken based on the comparison of median confidence intervals, regardless the strategy used to compute these confidence intervals, or if the decision is based on the comparison of trimmed mean confidence intervals, the rate of incorrect decisions, in general, is less than 1% for the 9 studied conditions. It is important to note that when we compare the mean confidence intervals, the error rate is approximately 10% for the condition of asymmetry in both groups and sample size of 50 subjects; for the other conditions, the error rate is less than the nominal alpha value of 5%.

We can see in Table 4, that the best performance is for t-test of Yuen-Welch and Mann-Whitney U-test. In the last case, for the group of little sample size (n = 10) and greater standard deviation (σ = 10), the error rate is around 10%. In relation to the confidence intervals, the performance is in general good, except for the mean confidence interval.

In the scenario of mean population differences (Tables 5 and 6), the results are not as easy to comment upon as in the scenario of no mean population differences. In general the best performance is for the Mann-Whitney U-test and the t-test of Yuen-Welch. However it is necessary to comment, that their performance is no good for small sample sizes (n = 10) with standard deviation of 10 and when there is asymmetry in the "x" group. In relation to the comparison of confidence intervals, in general their performance is bad and irregular (in some conditions they have a good performance, for example using the non-strict criterion when n_x = 30 and n_y = 10).

Table 3. Percentage of errors for which the null hypothesis is rejected when, in fact, it is true (the mean of the two populations is 100) according to the different criteria of decisions studied (for the comparison of the confidence intervals, the first value corresponds to the non-strict decision criterion: the means or medians are not included in the interval, and the second value corresponds to the strict decision criterion: the confidence intervals do not overlap).

Condition	t Student	t Yuen-Welch	U Mann-Whitney	Mean CI	Trimmed mean CI	EE Md CI	Binomial Md CI	MkS Md CI	MJ Md CI	k Md CI
x: n = 10, σ = 10, g = .0; y: n = 10, σ = 10, g = .0	4.68	4.88	5.08	9.18/0.50	7.36/0.80	6.52/0.26	1.92/0.12	1.66/0.02	5.64/0.46	10.14/1.24
x: n = 10, σ = 10, g = .0; y: n = 10, σ = 10, g = .8	6.70	4.38	6.66	10.04/0.98	6.70/0.76	4.16/0.20	2.34/0.12	1.00/0.04	4.50/0.42	7.90/1.20
x: n = 10, σ = 10, g = .8; y: n = 10, σ = 10, g = .8	3.78	3.82	5.08	4.94/0.52	5.28/0.74	2.76/0.22	1.90/0.12	0.72/0.04	3.40/0.34	5.74/0.98
x: n = 30, σ = 10, g = .0; y: n = 30, σ = 10, g = .0	4.62	4.84	4.50	11.90/0.52	11.10/0.70	11.46/0.42	6.22/0.28	7.74/0.32	9.10/0.60	12.98/1.06
x: n = 30, σ = 10, g = .0; y: n = 30, σ = 10, g = .8	23.14	5.40	10.50	34.64/4.12	11.26/0.70	5.24/0.18	6.34/0.30	6.16/0.28	8.26/0.54	11.88/1.14
x: n = 30, σ = 10, g = .8; y: n = 30, σ = 10, g = .8	4.10	4.10	4.70	7.70/0.56	8.74/0.58	2.60/0.08	6.20/0.28	4.68/0.20	6.84/0.48	10.12/1.10
x: n = 50, σ = 10, g = .0; y: n = 50, σ = 10, g = .0	5.04	4.84	4.88	13.42/0.50	12.50/0.56	12.88/0.30	5.70/0.30	9.58/0.32	10.46/0.58	13.32/0.80
x: n = 50, σ = 10, g = .0; y: n = 50, σ = 10, g = .8	40.16	6.06	14.78	54.34/11.80	13.38/0.66	5.04/0.08	5.88/0.33	8.22/0.32	9.82/0.58	13.02/0.94
x: n = 50, σ = 10, g = .8; y: n = 50, σ = 10, g = .8	4.88	4.52	4.88	9.54/0.62	10.72/0.48	2.52/0.02	5.76/0.30	6.76/0.26	9.58/0.48	12.30/0.98

n: sample size, σ: standard deviation, *g*: asymmntry generated using distribution *gh*, t Student: Student's t-test, t Yuen-Welch: Yuen-Welch's t-test (trimmed mean and winsorized variance), U Mann-Whitney: Mann-Whitney U test, mean CI: mean confidence intervals, Trimmed mean CI: trimmed mean confidence intervals, EE Md CI: median confidence intervals according to the standard error, Binomial Md CI: median confidence intervals according to the binomial distribution, Mks Md CI: median confidence intervals by McKean and Schraeder's estimation, MJ Md CI: median confidence intervals by Marizt and Jarret's estimation and k Md CI: median confidence intervals by the adaptive-kernel estimation.

Table 4. Percentage of errors for which the null hypothesis is rejected when, in fact, it is true (the mean of the two populations is 100) according to the different criteria of decisions studied (for the comparison of the confidence intervals, the first value corresponds to the non-strict decision criterion: the means or medians are not included in the interval, and the second value corresponds to the strict decision criterion: the confidence intervals do not overlap) in the conditions of equal sample size.

Condition	t Student	t Yuen-Welch	U Mann-Whitney	Mean CI	Trimmed mean CI	EE Md CI	Binomial Md CI	MkS Md CI	MJ Md CI	k Md CI
x: n = 10, σ = 5, g = .0; y: n = 30, σ = 10, g = .0	0.76	4.74	2.02	10.10/0.52	8.78/0.78	9.36/0.38	3.62/0.18	5.44/0.22	7.22/0.46	11.38/1.16
x: n = 10, σ = 5, g = .0; y: n = 30, σ = 10, g = .8	1.74	4.96	2.88	33.62/4.28	9.18/0.78	3.42/0.16	3.92/0.18	4.84/0.20	6.26/0.40	10.68/1.06
x: n = 10, σ = 5, g = .8; y: n = 30, σ = 10, g = .0	5.60	4.48	4.16	11.62/0.72	7.86/0.74	5.78/0.24	3.94/0.16	2.92/0.14	6.30/0.48	8.86/0.88
x: n = 10, σ = 5, g = .8; y: n = 30, σ = 10, g = .8	0.94	4.10	2.32	14.42/2.02	6.82/0.72	2.36/0.12	3.74/0.16	2.64/0.14	4.92/0.38	7.52/0.82
x: n = 10, σ = 10, g = .0; y: n = 30, σ = 5, g = .0	14.74	5.86	9.24	5.36/1.74	6.84/2.20	3.88/0.96	2.30/0.40	1.88/0.28	5.58/1.34	9.56/2.90
x: n = 10, σ = 10, g = .0; y: n = 30, σ = 5, g = .8	14.86	5.92	12.28	12.20/2.22	7.30/2.04	3.78/0.48	2.34/0.42	1.96/0.24	5.80/1.34	9.72/3.02
x: n = 10, σ = 10, g = .8; y: n = 30, σ = 5, g = .0	31.86	5.62	9.64	5.96/2.06	6.40/1.92	2.82/0.72	2.30/0.42	1.30/0.26	6.18/1.32	9.38/2.50
x: n = 10, σ = 10, g = .8; y: n = 30, σ = 5, g = .8	14.92	6.10	10.66	6.84/1.54	7.54/1.90	2.68/0.30	2.36/0.42	1.32/0.22	6.30/1.12	9.92/2.64
x: n = 10, σ = 10, g = .0; y: n = 30, σ = 10, g = .0	4.96	5.50	4.48	7.38/0.76	8.18/1.08	5.90/0.30	2.52/0.18	2.90/0.14	7.44/0.52	11.30/1.48
x: n = 10, σ = 10, g = .0; y: n = 30, σ = 10, g = .8	6.58	5.68	8.00	26.40/2.38	8.90/1.08	4.28/0.20	2.72/0.16	2.94/0.16	7.36/0.54	11.42/1.44
x: n = 10, σ = 10, g = .8; y: n = 30, σ = 10, g = .0	18.02	5.04	6.96	7.70/1.08	7.56/1.06	4.40/0.24	2.48/0.18	2.04/0.12	6.24/0.48	9.12/1.32
x: n = 10, σ = 10, g = .8; y: n = 30, σ = 10, g = .8	4.36	5.36	4.48	10.72/1.38	8.70/1.06	2.68/0.18	2.50/0.18	2.04/60.14	5.70/0.38	9.34/1.22
x: n = 30, σ = 5, g = .0; y: n = 10, σ = 10, g = .0	15.66	6.00	9.90	5.98/1.62	6.96/2.24	3.90/0.80	2.64/0.50	1.66/0.22	5.66/1.36	9.24/3.00
x: n = 30, σ = 5, g = .0; y: n = 10, σ = 10, g = .8	32.94	6.08	10.22	6.50/2.16	7.06/2.12	2.98/0.72	2.66/0.58	1.16/0.18	6.10/1.40	9.32/2.62
x: n = 30, σ = 5, g = .8; y: n = 10, σ = 10, g = .0	14.64	6.38	12.54	12.18/2.38	7.36/2.32	3.66/0.38	2.58/0.62	1.72/0.20	5.68/1.36	9.36/3.26
x: n = 30, σ = 5, g = .8; y: n = 10, σ = 10, g = .8	15.10	6.56	10.66	6.70/1.50	7.84/2.20	2.88/0.40	2.60/0.60	1.32/0.16	6.28/1.32	9.64/2.72
x: n = 30, σ = 10, g = .0; y: n = 10, σ = 5, g = .0	0.84	4.90	1.92	10.78/0.60	9.44/0.62	9.22/0.30	4.10/0.16	3.80/0.12	7.76/0.42	11.36/1.06
x: n = 30, σ = 10, g = .0; y: n = 10, σ = 5, g = .8	5.84	4.44	4.76	12.88/0.92	7.96/0.62	5.78/0.20	4.84/0.18	2.20/0.04	6.34/0.38	9.12/0.86
x: n = 30, σ = 10, g = .8; y: n = 10, σ = 5, g = .0	1.54	5.28	2.96	33.20/4.36	9.56/0.58	3.26/0.12	4.16/0.16	2.88/0.10	6.58/0.42	10.48/1.04
x: n = 30, σ = 10, g = .8; y: n = 10, σ = 5, g = .8	1.04	4.14	2.44	13.92/5.00	6.92/0.56	2.22/0.12	4.38/0.14	1.60/0.02	5.00/0.34	7.68/0.86
x: n = 30, σ = 10, g = .0; y: n = 10, σ = 10, g = .0	5.18	5.98	5.12	7.78/0.94	8.40/1.00	5.70/0.30	3.12/0.18	2.20/0.12	6.76/0.58	10.64/1.38
x: n = 30, σ = 10, g = .0; y: n = 10, σ = 10, g = .8	18.50	5.20	7.72	8.28/1.04	7.70/0.84	3.84/0.42	3.16/0.14	1.50/0.08	6.16/0.56	8.78/1.08
x: n = 30, σ = 10, g = .8; y: n = 10, σ = 10, g = .0	6.12	6.20	8.24	25.70/2.36	9.08/1.02	3.98/0.18	3.20/0.18	2.20/0.06	6.74/0.54	11.14/1.40
x: n = 30, σ = 10, g = .8; y: n = 10, σ = 10, g = .8	4.62	5.54	5.12	10.88/1.48	8.72/1.02	2.58/0.20	3.14/0.18	1.32/0.06	5.70/0.48	9.08/1.04

n: sample size, σ: standard deviation, g: asymmntry generated using distribution gh, t Student: Student's t-test, t Yuen-Welch: Yuen-Welch's t-test (trimmed mean and winsorized variance), U Mann-Whitney: Mann-Whitney U test, mean CI: mean confidence intervals, Trimmed mean CI: trimmed mean confidence intervals, EE Md CI: median confidence intervals according to the standard error, Binomial Md CI: median confidence intervals according to the binomial distribution, Mks Md CI: median confidence intervals by McKean and Schraeder's estimation, MJ Md CI: median confidence intervals by Marizt and Jarret's estimation and k Md CI: median confidence intervals by the adaptive-kernel estimation.

Table 5. Percentage of errors for which the null hypothesis is not rejected when, in fact, it is not true (the means of the two populations are 100 and 115, respectively) according to the different criteria of decisions studied (for the comparison of the confidence intervals, the first value corresponds to the non-strict decision criterion: the means or medians are not included in the interval, and the second value corresponds to the strict decision criterion: the confidence intervals do not overlap).

Condition	t Student	t Yuen-Welch	U Mann-Whitney	Mean CI	Trimmed mean CI	EE Md CI	Binomial Md CI	MkS Md CI	MJ Md CI	k Md CI
x: $n = 10$, $\sigma = 10$, $g = .0$; y: $n = 10$, $\sigma = 10$, $g = .0$	5.08	10.58	6.26	4.20/41.34	9.90/54.76	12.66/65.62	19.86/77.40	22.16/89.04	11.38/65.94	6.18/50.60
x: $n = 10$, $\sigma = 10$, $g = .0$; y: $n = 10$, $\sigma = 10$, $g = .8$	0.80	5.94	1.12	0.62/27.38	6.56/50.58	10.64/76.14	17.24/60.72	19.80/92.72	9.56/64.70	4.72/48.70
x: $n = 10$, $\sigma = 10$, $g = .8$; y: $n = 10$, $\sigma = 10$, $g = .0$	40.30	23.88	24.72	41.64/71.90	24.66/61.06	37.32/75.64	50.64/85.80	46.56/91.88	18.30/66.40	13.58/53.98
x: $n = 10$, $\sigma = 10$, $g = .8$; y: $n = 10$, $\sigma = 10$, $g = .8$	27.86	19.14	11.96	27.88/64.20	20.54/57.92	35.04/82.54	47.92/76.82	44.36/94.10	16.24/64.16	12.00/52.50
x: $n = 30$, $\sigma = 10$, $g = .0$; y: $n = 30$, $\sigma = 10$, $g = .0$	0.00	0.02	0.02	0.00/0.18	0.02/0.74	0.06/3.70	0.06/4.44	0.10/8.42	0.10/7.04	0.02/3.00
x: $n = 30$, $\sigma = 10$, $g = .0$; y: $n = 30$, $\sigma = 10$, $g = .8$	0.00	0.00	0.00	0.00/0.02	0.00/0.14	0.06/16.26	0.00/0.86	0.06/7.54	0.06/5.26	0.00/1.08
x: $n = 30$, $\sigma = 10$, $g = .8$; y: $n = 30$, $\sigma = 10$, $g = .0$	13.30	0.68	0.78	15.66/37.44	0.80/6.84	4.50/21.42	2.38/15.24	1.16/13.90	0.80/10.15	0.38/4.82
x: $n = 30$, $\sigma = 10$, $g = .8$; y: $n = 30$, $\sigma = 10$, $g = .8$	4.48	0.38	0.06	5.34/20.34	0.46/4.00	4.14/40.30	2.16/9.14	1.00/14.06	0.70/8.80	2.92/0.06
x: $n = 50$, $\sigma = 10$, $g = .0$; y: $n = 50$, $\sigma = 10$, $g = .0$	0.00	0.00	0.00	0.00/0.00	0.00/0.00	0.00/0.02	0.00/0.26	0.00/0.38	0.00/0.42	0.00/0.02
x: $n = 50$, $\sigma = 10$, $g = .0$; y: $n = 50$, $\sigma = 10$, $g = .8$	0.00	0.00	0.00	0.00/0.00	0.00/0.00	0.00/2.34	0.00/0.00	0.00/0.18	0.00/0.16	0.00/0.00
x: $n = 50$, $\sigma = 10$, $g = .8$; y: $n = 50$, $\sigma = 10$, $g = .0$	5.22	0.02	0.00	6.64/17.98	0.04/0.48	0.58/3.90	0.18/2.40	0.02/1.28	0.02/1.34	0.00/0.32
x: $n = 50$, $\sigma = 10$, $g = .8$; y: $n = 50$, $\sigma = 10$, $g = .8$	0.76	0.00	0.00	1.26/6.22	0.02/0.16	0.50/12.80	0.14/1.06	0.00/0.98	0.02/0.82	0.00/0.18

n: sample size, *σ*: standard deviation, *g*: asymmntry generated using distribution *gh*, t Student: Student's t-test, t Yuen-Welch: Yuen-Welch's t-test (trimmed mean and winsorized variance), U Mann-Whitney: Mann-Whitney U test, mean CI: mean confidence intervals, Trimmed mean CI: trimmed mean confidence intervals, EE Md CI: median confidence intervals according to the standard error, Binomial Md CI: median confidence intervals according to the binomial distribution, Mks Md CI: median confidence intervals by McKean and Schraeder's estimation, MJ Md CI: median confidence intervals by Marizt and Jarret's estimation and k Md CI: median confidence intervals by the adaptive-kernel estimation.

Table 6. Percentage of errors for which the null hypothesis is not rejected when, in fact, it is not true (the means of the two populations are 100 and 115, respectively) according to the different criteria of decisions studied (for the comparison of the confidence intervals, the first value corresponds to the non-strict decision criterion: the means or medians are not included in the interval, and the second value corresponds to the strict decision criterion: the confidence intervals do not overlap).

Condition	t Student	t Yuen-Welch	U Mann-Whitney	Mean CI	Trimmed mean CI	EE Md CI	Binomial Md CI	MkS Md CI	MJ Md CI	k Md CI
x: $n = 10$, $\sigma = 5$, $g = .0$; y: $n = 30$, $\sigma = 10$, $g = .0$	0.02	0.02	0.00	0.00/0.22	0.02/1.22	0.06/3.18	0.16/5.72	0.22/11.56	0.10/5.82	0.00/1.70
x: $n = 10$, $\sigma = 5$, $g = .0$; y: $n = 30$, $\sigma = 10$, $g = .8$	0.12	0.00	0.00	0.00/0.06	0.02/0.24	0.00/15.34	0.10/1.56	0.16/10.72	0.02/3.46	0.00/0.26
x: $n = 10$, $\sigma = 5$, $g = .8$; y: $n = 30$, $\sigma = 10$, $g = .0$	4.26	1.60	1.58	8.28/18.82	2.70/9.60	5.74/19.46	14.40/32.92	11.46/33.36	1.22/10.78	0.60/6.38
x: $n = 10$, $\sigma = 5$, $g = .8$; y: $n = 30$, $\sigma = 10$, $g = .8$	2.12	0.86	0.14	3.48/10.28	1.84/7.28	5.34/34.14	13.42/27.50	10.72/33.52	1.00/8.72	0.52/4.00
x: $n = 10$, $\sigma = 10$, $g = .0$; y: $n = 30$, $\sigma = 5$, $g = .0$	0.06	2.36	0.34	1.62/4.96	6.48/12.82	8.00/17.90	14.96/28.50	17.46/34.72	7.60/17.32	2.86/8.48
x: $n = 10$, $\sigma = 10$, $g = .0$; y: $n = 30$, $\sigma = 5$, $g = .8$	0.06	1.74	0.08	0.30/2.40	5.48/11.22	7.54/24.86	14.40/25.16	17.02/35.16	7.36/16.54	2.68/7.40
x: $n = 10$, $\sigma = 10$, $g = .8$; y: $n = 30$, $\sigma = 5$, $g = .0$	17.98	16.34	8.66	40.08/47.67	22.40/29.86	34.42/44.80	49.50/59.36	44.64/56.76	16.08/24.44	11.72/18.58
x: $n = 10$, $\sigma = 10$, $g = .8$; y: $n = 30$, $\sigma = 5$, $g = .8$	14.06	14.92	6.44	31.46/41.78	20.72/28.14	34.10/50.46	49.18/57.48	44.12/57.00	15.80/23.90	11.44/17.68
x: $n = 10$, $\sigma = 10$, $g = .0$; y: $n = 30$, $\sigma = 10$, $g = .0$	0.94	3.72	1.36	2.50/14.16	7.24/25.10	9.58/35.36	16.32/46.44	19.22/55.44	9.18/33.86	4.06/21.36
x: $n = 10$, $\sigma = 10$, $g = .0$; y: $n = 30$, $\sigma = 10$, $g = .8$	0.46	2.04	0.18	0.30/2.40	5.12/20.54	8.86/52.70	15.50/38.74	18.14/57.22	8.56/32.16	3.56/17.58
x: $n = 10$, $\sigma = 10$, $g = .8$; y: $n = 30$, $\sigma = 10$, $g = .0$	24.88	17.44	13.48	40.20/56.76	23.00/39.42	35.02/55.94	49.64/69.22	44.76/69.36	16.78/38.06	12.74/29.14
x: $n = 10$, $\sigma = 10$, $g = .8$; y: $n = 30$, $\sigma = 10$, $g = .8$	16.5	14.64	6.02	25.10/44.20	19.66/36.32	34.16/67.96	48.90/65.66	44.12/70.34	16.16/36.78	12.18/26.18
x: $n = 30$, $\sigma = 5$, $g = .0$; y: $n = 10$, $\sigma = 10$, $g = .0$	0.06	2.50	0.36	0.00/5.16	0.00/13.32	0.04/18.36	0.08/28.34	0.02/49.68	0.06/29.14	0.02/19.70
x: $n = 30$, $\sigma = 5$, $g = .0$; y: $n = 10$, $\sigma = 10$, $g = .8$	0.00	0.58	0.00	0.00/1.96	0.00/5.98	0.00/36.28	0.00/0.08	0.00/64.74	0.00/24.04	0.00/13.94
x: $n = 30$, $\sigma = 5$, $g = .8$; y: $n = 10$, $\sigma = 10$, $g = .0$	4.18	3.40	1.64	1.22/25.52	0.08/18.00	0.34/28.00	0.28/34.16	0.10/52.10	0.08/30.28	0.04/20.26
x: $n = 30$, $\sigma = 5$, $g = .8$; y: $n = 10$, $\sigma = 10$, $g = .8$	0.90	0.76	0.00	0.10/11.96	0.00/9.50	0.10/45.12	0.02/2.18	0.02/65.92	0.02/25.16	0.00/14.62
x: $n = 30$, $\sigma = 10$, $g = .0$; y: $n = 10$, $\sigma = 5$, $g = .0$	0.00	0.02	0.00	0.00/0.00	0.00/1.24	0.04/2.94	0.02/5.90	0.04/13.90	0.14/6.30	0.02/2.78
x: $n = 30$, $\sigma = 10$, $g = .0$; y: $n = 10$, $\sigma = 5$, $g = .8$	0.00	0.00	0.00	0.00/0.00	0.00/0.76	0.04/12.24	0.00/0.26	0.02/29.76	0.06/5.48	0.00/2.04
x: $n = 30$, $\sigma = 10$, $g = .8$; y: $n = 10$, $\sigma = 5$, $g = .0$	31.64	0.56	1.58	16.02/25.48	0.88/7.62	4.6/20.20	2.04/17.12	0.86/19.98	0.08/9.50	0.38/4.38
x: $n = 30$, $\sigma = 10$, $g = .8$; y: $n = 10$, $\sigma = 5$, $g = .8$	21.38	0.24	0.56	9.84/17.02	0.56/6.16	4.04/31.52	1.78/7.26	0.82/35.36	0.64/8.86	0.32/3.84
x: $n = 30$, $\sigma = 10$, $g = .0$; y: $n = 10$, $\sigma = 10$, $g = .0$	0.92	3.66	1.16	0.04/14.82	0.16/26.74	0.86/35.32	0.92/47.42	0.84/67.42	0.90/45.86	0.54/33.82
x: $n = 30$, $\sigma = 10$, $g = .0$; y: $n = 10$, $\sigma = 10$, $g = .8$	0.02	0.92	0.02	0.00/4.90	0.00/17.28	0.18/52.36	0.28/11.42	0.22/77.68	0.38/41.80	0.10/29.06
x: $n = 30$, $\sigma = 10$, $g = .8$; y: $n = 10$, $\sigma = 10$, $g = .0$	36.10	7.72	8.26	19.70/63.32	2.28/37.48	7.60/55.46	4.58/56.62	2.52/69.20	1.92/47.68	1.18/33.82
x: $n = 30$, $\sigma = 10$, $g = .8$; y: $n = 10$, $\sigma = 10$, $g = .8$	19.26	3.00	0.94	8.56/51.08	1.04/30.03	5.68/64.48	2.96/26.42	1.48/78.92	1.22/43.68	0.70/29.76

n: sample size, *σ*: standard deviation, *g*: asymmntry generated using distribution *gh*, t Student: Student's t-test, t Yuen-Welch: Yuen-Welch's t-test (trimmed mean and winsorized variance), U Mann-Whitney: Mann-Whitney U test, mean CI: mean confidence intervals, Trimmed mean CI: trimmed mean confidence intervals, EE Md CI: median confidence intervals according to the standard error, Binomial Md CI: median confidence intervals according to the binomial distribution, Mks Md CI: median confidence intervals by McKean and Schraeder's estimation, MJ Md CI: median confidence intervals by Marizt and Jarret's estimation and k Md CI: median confidence intervals by the adaptive-kernel estimation.

DISCUSSION

The first conclusion that we obtain from this work is that the comparison of confidence intervals based on the median has a low type I error rate, regardless of the method used to compute the confidence interval, but this method's power is not as good as expected. In fact, this result has also been obtained by Peró et al. (2011) and Peró et al. (2008). Also, these results are in line with the results obtained by Holmes-Finch and Davenport (2009) in the use of a Monte Carlo permutation for small samples when the dependent variable does not follow a normal distribution in a MANOVA design.

Although we noted previously that the decision from the comparison median confidence intervals does not exhibit the expected results, it is important to note that when there is no mean population differences, their performance is good, and in the scenario of mean population differences when working with large samples sizes (more than 30 subjects in each group), the decision based on the comparison of median confidence intervals from a binomial distribution, McKean and Schraeder estimation, Marizt and Jarret estimation and adaptive-kernel estimation have good power, if the decision criterion is non-strict.

If we had to recommend a statistical decision strategy for comparing two independent groups, based on our results, the best strategies are Mann-Whitney U-test and the t-test of Yuen Welch. The computation of this last test is easily implemented in R [R code: `yuen(x,y,tr=0.2,alpha=0.05)`]. This result is congruent because the Mann-Whitney U-test is a non-parametric test adequate for asymmetrical distributions and the computation of t-test of Yuen-Welch is based on robust and resistant measures, namely, the trimmed mean and the winsorized variance (Wilcox, 2005).

Complementarily, it is important to point out the bias effect that some coefficients of symmetry could generate in the data simulated when we generate complex distributions. Our results show that in very small samples this effect could generate some relevant distortion.

Our results lead us to reflect about the statistical decision. In this work, we present empirical evidence concerning the incorrect performance of the t-test. However, this test is not the only statistical tests that could have a wrong performance when the conditions of application are violated (for example, ANOVA and the chi-square test), and there are a few works that present empirical evidence about their performance (e.g., Holmes-Finch and Davenport, 2009). We should rethink the use of traditional statistical tests in applied research; these tests are adequate when all of the assumptions are satisfied, but the fulfilment of these assumptions is not the norm in applied research; for example, the random sampling of the units studied is often not satisfied.

It is possible that in these decisions, we can act as *Homo Heuristicus* and use only the useful information to make a decision (Gigerenzer, 1991; Gigerenzer and Brighton, 2009; and Goldstein and Gigerenzer, 2002). Moreover, it is necessary to comment that in this work, we consider the use of confidence intervals based on the mean or the median to be similar to null hypothesis significance testing (NHST), and this approach may be complemented, as Coulson et al. (2010) and Cumming and Fidler (2011) propose. As those authors note, the correct option may be not to interpret the confidence intervals subsumed to the 0 difference and to consider the precision of the intervals and their graphical representation.

RESUMEN

Estudio de la adecuación de diferentes pruebas estadísticas robustas para la comparación de dos grupos independientes. En el presente estudio, se evalúan diferentes pruebas estadísticas robustas para la comparación de dos grupos independientes. Se han generado dos escenarios de simulación: uno para igualdad de medias poblacionales y otro para desigualdad. Para cada escenario se han utilizado 33 condiciones experimentales manipulando los valores de tamaño de muestra, desviación estándar y asimetría. Para cada condición, se han generado 5000 replicaciones por grupo. Los resultados obtenidos muestran una tasa adecuada de error tipo I pero la potencia asociada a los intervalos de confianza no es adecuada. En general, para los dos escenarios estudiados (diferencias y no diferencias de medias poblacionales) en las diferentes condiciones analizadas, la prueba U de Mann-Whitney es la que presenta el mejor rendimiento, y un poco peor la prueba t de Yuen-Welch.

REFERENCES

Algina, J., Keselman, H.J., & Penfield, R.D. (2005). An alternative to Cohen's standardized mean difference effect size: a robust parameter and confidence interval in the two independent groups case. *Psychological Methods*, *10*, 317-328.

American Psychological Association (APA) (2010). *Publication manual of the American Psychological Association. Sixth edition*. Washington: American Psychological Association.

Bailar, J., & Mosteller, F. (1988). Guidelines for statistical reporting in articles for medical journals. Amplifications and explanations. *Annals of Internal Medicine*, *108*, 266-273.

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*, *10*, 389-396.

Bland, M. (2003). *Confidence interval for a median and other quantiles*. Document retrieved on 5th May 2005 from Word Wide Web: http://www-users.york.ac.uk/~mb55/intro/cicent.htm.

Bonett, D.G., & Price, R.M. (2002). Statistical inference for a linear function of medians: confidence intervals, hypothesis testing, and sample size requirements. *Psychological Methods*, *7*, 370-383.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist*, *49*, 997-1003.

Coulson, M., Healey, M., Fidler, F., & Cumming, G. (2010). Confidence intervals permit, but do not guarantee, better inference than statistical significance testing. *Frontiers in Psychology*, *1*, 1-8. doi: 10.3389/fpsyg.2010.00026.

Cowles, M. (1989). *Statistics in psychology: an historical perspective*. Hillsdale, New Jersey: Lawrence Erlbaum Associates (LEA), Inc, Publishers.

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, *3*, 286-300.

Cumming, G. (2009). Inference by eye: Reading the overlap of independent confidence intervals. *Statistics in Medicine*, *28*, 205-220.

Cumming, G., & Fidler, F. (2009). Confidence intervals: Better answers to better questions. *Zeitschrift für Psychologie / Journal of Psychology*, *217*, 15-26.

Cumming, G., & Fidler, F. (2011). From hypothesis testing to parameter estimation: an example of evidence-based practice in statistics. In A.T. Panter, & S.K. Sterba (Eds.), *Handbook of ethics in quantitative methodology* (pp. 293-312). New York: Routledge Taylor & Francis Group.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, *61*, 532-574.

Cumming, G., & Finch, S. (2005). Inference by eye. Confidence intervals and how to read pictures of data. *American Psychologist*, *60*, 170-180.

Cumming, G., & Maillardet, R. (2006). Confidence intervals and replications: where will the next mean fall? *Psychological Methods*, *11*, 217-227.

DeCoster, C., & Burchill, C. (2000). *Confidence interval of the median*. Document retrieved on 5th May 2005 from Word Wide Web: http://www.umanitoba.ca/centres/MCHP/concept/dict/ci_median.

De la Fuente, E.I., Cañadas, G.R., Guàrdia, J., & Lozano, L.M. (2009). Hypothesis probability or statistical significance? *Methodology*, *5* (1), 35-39.

Dubnicka, S.R. (2007). A confidence interval for the median of a finite population under unequal probability sampling: a model-assisted approach. *Journal of Statistical Planning and Inference*, *137*, 2429-2438.

Field, C., & Genton, M.G. (2006). The multivariate g-and-h distribution. *Technometrics*, *48*, 104-111.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: beyond “heuristics and biases”. *European Review of Social Psychology*, *2*, 83-115.

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: why biased minds make better inference. *Topics in Cognitive Science*, *1*, 107-143.

Goldstein, D.G., & Gigerenzer, G. (2002). Models of ecological rationality: the recognition heuristic. *Psychological Review*, *109*, 75-90.

Hagen, R.L. (1997). In praise of the null hypothesis statistical test. *American Psychologist*, *52*, 15-24.

Hoaglin, D.C. (1985). Summarizing shape numerically: the g-and-h distribution. In D. Hoaglin, F. Mosteller, & J.W. Tukey (Eds), *Exploring data tables, trends and shapes* (pp. 79-98). New York: John Willey and Sons.

Holmes-Finch, W., & Davenport, T. (2009). Performance of Monte Carlo permutation and approximate tests for multivariate means comparisons with small sample sizes when parametric assumptions are violated. *Methodology*, *5* (2), 60-70.

Kendall, M.G. (1945). *The advanced theory of statistics. Volume I.* London: Charles Griffin & Company Limited.

Keselman, H.J., Algina, J., Lix, L.M., Wilcox, R.R., & Deering, K.N. (2008). A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes. *Psychological Methods*, *13*, 110-129.

Lane, D. (1999). Sampling distribution of median. Document retrieved on 19th May 2005 from World Wide Web: http://davidmlane.com/hyperstat/A106993.html.

Lin, Y., Newcombe, R.G., Lipsitz, S., & Carter, R.E. (2009). Fully specified bootstrap confidence intervals for the difference of two independent binomial proportions based on the median unbiased estimator. *Statistics in Medicine*, *28*, 2879-2890.

Maritz, J.S., & Jarrett, R.L. (1978). A note on estimating the variance of the sample median. *Journal of the American Statistical Association*, *73*, 194-196.

McKean, J.W., & Schrader, R.M. (1984). A comparison of methods for studentizing the sample median. *Communications in Statistics--Simulation and Computation*, *13*, 751-773.

Mothes, J., & Torrens-Ibern, J. (1970). *Estadística aplicada a la ingeniería*. Barcelona. Ariel.

Peró, M., Delgado, J., & Guàrdia, J. (2011). Comparison of confidence intervals based on the median as a robust alternative to the classic hypotheses test. *Advances and Applications in Statistics*, *24* (1), 67-82.

Peró, M., Guàrdia, J., Freixa, M., & Turbany, J. (2008). Técnicas basadas en la mediana como alternativa a las pruebas clásicas de decisión. *Psicothema*, *20*, 857-862.

R Development Core Team. (2010). *R: A Languaje and Enviroment for Statistical Computing (Version 2.11.1) {Computer software}.*Viena: R Foundation for Statistical Computing.

Strelen, J.C. (2001). *Median confidence intervals*. Document retrieved on 11th April 2007 from World Wide Web: http://web.informatik.uni-bonn.de/IV/strelen/Forschung/Publikationen/ESM2001.pdf.

Strelen, J.C. (2004). *The accuracy of a new confidence interval method*. Document retrieved on 11th April 2007 from World Wide Web: http://ieeexplore.ieee.org/iel5/9441/29988/01371373.pdf.

Tryon, W.W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: an integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods*, *6* (4), 371-386.

Wilcox, R.R. (2005). *Introduction to robust estimation and hypothesis testing*. Massachusetts: Elsevier Academic Press.

Wilkinson, L., & the Task Force on Statistical Inference (1999). Statistical methods in psychology journals. Guidelines and explanations. *American Psychologist*, *54*, 594-604.

Wolfe, R., & Hanley, J. (2002). If we're so different, why do we keep overlapping? When 1 plus 1 doesn't make 2. *Canadian Medical Association Journal*, *166*, 65-66.

Woodruff, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, *47*, 635-646.

(Manuscript received: 14 May 2012; accepted: 7 February 2013)