Biometrika (2022), **109**, 3, *pp.* 873–880 doi: 10.1093/biomet/asab053
Advance Access publication 30 December 2021

On the relative efficiency of the intent-to-treat Wilcoxon–Mann–Whitney test in the presence of noncompliance

By LU MAO

Department of Biostatistics and Medical Informatics, School of Medicine and Public Health, University of Wisconsin-Madison, Madison, Wisconsin 53726, U.S.A.
lmao@biostat.wisc.edu

Summary

A general framework is set up to study the asymptotic properties of the intent-to-treat Wilcoxon–Mann–Whitney test in randomized experiments with nonignorable noncompliance. Under location-shift alternatives, the Pitman efficiencies of the intent-to-treat Wilcoxon–Mann–Whitney and t tests are derived. It is shown that the former is superior if the compliers are more likely to be found in high-density regions of the outcome distribution or, equivalently, if the noncompliers tend to reside in the tails. By logical extension, the relative efficiency of the two tests is sharply bounded by least and most favourable scenarios in which the compliers are segregated into regions of lowest and highest density, respectively. Such bounds can be derived analytically as a function of the compliance rate for common location families such as Gaussian, Laplace, logistic and t distributions. These results can help empirical researchers choose the more efficient test for existing data, and calculate sample size for future trials in anticipation of noncompliance. Results for nonadditive alternatives and other tests follow along similar lines.

Some key words: Causal inference; Instrumental variable; Nonignorable missingness; Pitman efficiency; Randomized controlled trial; t test.

1. Motivation

The asymptotic properties of the two-sample Wilcoxon–Mann–Whitney test (Wilcoxon, 1945; Mann & Whitney, 1947) have been thoroughly investigated in the literature, especially in comparison with the t test (Sidak et al., 1999). It is well known that the relative performance of the Wilcoxon–Mann–Whitney and t tests depends on the shape of the underlying outcome distribution, with the former being superior under heavy-tailed distributions. In fact, the asymptotic relative efficiency of the two tests has been explicitly derived for various outcome distributions, such as the Gaussian, logistic and Laplace families, under location-shift alternatives (see, e.g., van der Vaart, 1998, Ch. 14).

The situation may be different in the presence of nonignorable, i.e., informative, noncompliance to group assignment (Angrist et al., 1996). In such settings, an intent-to-treat test that compares the subjects according to their randomization status is commonly employed to circumvent the selection bias in the treatment received (Gupta, 2011). Because of the unknown relationship between the compliance status and potential outcome, however, the power structures of such tests remain opaque. In this paper, we use the instrumental-variable framework (Angrist et al., 1996; Imbens & Rubin, 1997) to study the asymptotic properties of the intent-to-treat Wilcoxon–Mann–Whitney test in the presence of noncompliance, with particular emphasis on comparison with its t-test counterpart in detecting additive treatment effects.

2. ASYMPTOTIC THEORY

2.1. *General set-up for intent-to-treat tests*

Let $Y(z,a)$ denote the continuous potential outcome under randomization status z and received treatment a, where z and a are binary variables, each taking value 1 for the active treatment and value 0 for the control (Rubin, 1978). Similarly, let $A(z)$ denote the potential treatment under randomization status z. This notation defines four compliance classes: always-taker if $A(z)=1$; complier if $A(z)=z$; never-taker if $A(z)=0$; and defier if $A(z)=1-z$ $(z=1,0)$. Under the stable unit treatment value assumption (Imbens & Rubin, 2015), the observed treatment and outcome are $A=A(Z)$ and $Y=Y(Z,A)$, respectively. In addition, we assume that randomization has a nontrivial effect on the treatment received, i.e., $\mathrm{pr}\{A(1)=1\} \neq \mathrm{pr}\{A(0)=1\}$, and that there is no defier in the population, i.e., $A(1) \geqslant A(0)$ almost surely. In the interest of generality, we relax the usual exclusion restriction assumption that $Y(1,a)=Y(0,a)$ $(a=1,0)$ with probability 1 (Angrist et al., 1996) to allow for possible randomization direct effects.

Denote the proportions of the three remaining compliance classes by $p_{\mathrm{a}}=\mathrm{pr}\{A(1)=A(0)=1\}$, $p_{\mathrm{c}}=\mathrm{pr}\{A(1)=1,A(0)=0\}$ and $p_{\mathrm{n}}=\mathrm{pr}\{A(1)=A(0)=0\}$. By Imbens & Rubin (1997), the observed outcomes in each randomized group can be viewed as a mixture of potential outcomes from the three classes. To be specific, let $\omega_z(\cdot \mid \theta)$ denote the cumulative distribution function for $(Y \mid Z=z)$ $(z=1,0)$, where θ is some real-valued parameter. Then

$$\begin{aligned}\omega_1(\cdot \mid \theta) &= p_{\mathrm{c}}\nu_{1\mathrm{c}}(\cdot \mid \theta)+p_{\mathrm{a}}\nu_{1\mathrm{a}}(\cdot \mid \theta)+p_{\mathrm{n}}\nu_{1\mathrm{n}}(\cdot \mid \theta),\\ \omega_0(\cdot \mid \theta) &= p_{\mathrm{c}}\nu_{0\mathrm{c}}(\cdot \mid \theta)+p_{\mathrm{a}}\nu_{0\mathrm{a}}(\cdot \mid \theta)+p_{\mathrm{n}}\nu_{0\mathrm{n}}(\cdot \mid \theta),\end{aligned} \tag{1}$$

where $\{Y(z,z) \mid A(1)=1,A(0)=0\} \sim \nu_{z\mathrm{c}}(\cdot \mid \theta)$, $\{Y(z,1) \mid A(1)=A(0)=1\} \sim \nu_{z\mathrm{a}}(\cdot \mid \theta)$ and $\{Y(z,0) \mid A(1)=A(0)=0\} \sim \nu_{z\mathrm{n}}(\cdot \mid \theta)$ $(z=1,0)$. Suppose that $\theta=0$ represents a scenario of the sharp null hypothesis

$$H_0 : Y(z,a)=Y(0,0) \quad (z,a=1,0),$$

so that $\nu_{z\mathrm{c}}(\cdot \mid 0)=\nu_{\mathrm{c}}(\cdot)$, $\nu_{z\mathrm{a}}(\cdot \mid 0)=\nu_{\mathrm{a}}(\cdot)$ and $\nu_{z\mathrm{n}}(\cdot \mid 0)=\nu_{\mathrm{n}}(\cdot)$ $(z=1,0)$ for some ν_{c}, ν_{a} and ν_{n}. Then $\omega_1(\cdot \mid 0)=\omega_0(\cdot \mid 0)=\nu(\cdot)$, where $\nu=p_{\mathrm{c}}\nu_{\mathrm{c}}+p_{\mathrm{a}}\nu_{\mathrm{a}}+p_{\mathrm{n}}\nu_{\mathrm{n}}$.

With a random sample of size n, suppose that a level-α test $(0<\alpha<1)$ rejects H_0 and accepts $H_{\mathrm{A}}:\theta>0$ when $n^{1/2}|T_n-\mu(0)|/\hat{\sigma}_n>z_{1-\alpha/2}$, where T_n is a regular estimator for some function $\mu(\theta)$ of θ, $\hat{\sigma}_n^2$ is a consistent estimator for its asymptotic variance, and $z_{1-\alpha/2}=\Phi^{-1}(1-\alpha/2)$ with $\Phi(\cdot)$ denoting the standard normal cumulative distribution function. The asymptotic power of the test can be evaluated analytically under a sequence of contiguous alternatives approaching H_0 at rate $O(n^{-1/2})$, say $\theta=n^{-1/2}h$ for some $h>0$. Under this θ, the power of the test tends to $\Phi(\zeta h-z_{1-\alpha/2})$ for some $\zeta>0$ as $n\to\infty$ (see van der Vaart, 1998, Ch. 14). The quantity ζ^2, commonly called the Pitman efficiency, is given by $\zeta^2=\dot{\mu}(0)^2/\sigma_0^2$, where σ_0^2 is the asymptotic limit of $\hat{\sigma}_n^2$ under H_0 and $\dot{f}(x)=\mathrm{d}f(x)/\mathrm{d}x$ for any function f. The Pitman efficiency is inversely proportional to the sample size needed to achieve a given power and is used aptly to measure the asymptotic quality of the test.

Nonparametric intent-to-treat tests can often be formulated as contrasts of empirical distributions. Let $\hat{\omega}_z(\cdot)$ denote the empirical analogue of $\omega_z(\cdot \mid \theta)$ $(z=1,0)$. Then the Wilcoxon–Mann–Whitney and t test statistics can be written as $T_{\mathrm{WMW},n}=\int \hat{\omega}_0(y)\hat{\omega}_1(\mathrm{d}y)-2^{-1}$ (Mann & Whitney, 1947) and $T_{t,n}=\int y\{\hat{\omega}_1(\mathrm{d}y)-\hat{\omega}_0(\mathrm{d}y)\}$, respectively. The estimands of $T_{\mathrm{WMW},n}$ and $T_{t,n}$ are thus $\mu_{\mathrm{WMW}}(\theta)=\int \omega_0(y \mid \theta)\omega_1(\mathrm{d}y \mid \theta)-2^{-1}$ and $\mu_t(\theta)=\int y\{\omega_1(\mathrm{d}y \mid \theta)-\omega_0(\mathrm{d}y \mid \theta)\}=-\int\{\omega_1(y \mid \theta)-\omega_0(y \mid \theta)\}\,\mathrm{d}y$, respectively. Write $\nabla\eta(\cdot)=\partial\eta(\cdot \mid \theta)/\partial\theta\big|_{\theta=0}$ for any distribution function η indexed by θ. The following lemma, proved in the Appendix, provides general expressions for the Pitman efficiencies of the two tests.

LEMMA 1. *Under* (1) *with* $\theta=O(n^{-1/2})$, *the Pitman efficiencies for the intent-to-treat Wilcoxon–Mann–Whitney and t tests are*

$$\zeta^2_{\mathrm{WMW}} = 12q(1-q)\left[\int \{\nabla\omega_1(y) - \nabla\omega_0(y)\}\dot{\nu}(y)\,\mathrm{d}y\right]^2,$$

$$\zeta^2_t = q(1-q)V(\nu)^{-1}\left[\int \{\nabla\omega_1(y) - \nabla\omega_0(y)\}\,\mathrm{d}y\right]^2,$$

respectively, where $q = E(Z)$ *and* $V(\nu) = \int\{y - \int y^*\nu(\mathrm{d}y^*)\}^2\dot{\nu}(y)\,\mathrm{d}y$.

By Lemma 1, the Pitman efficiencies of the two tests are both functionals of

$$\nabla\omega_1 - \nabla\omega_0 = p_{\mathrm{c}}\{\nabla\nu_{1\mathrm{c}} - \nabla\nu_{0\mathrm{c}}\} + p_{\mathrm{a}}\{\nabla\nu_{1\mathrm{a}} - \nabla\nu_{0\mathrm{a}}\} + p_{\mathrm{n}}\{\nabla\nu_{1\mathrm{n}} - \nabla\nu_{0\mathrm{n}}\}, \tag{2}$$

which mixes the treatment and randomization effects across the compliance classes. Under exclusion restriction, for example, the latter two terms on the right-hand side of (2) vanish so that the difference consists only of the treatment effect on the compliers.

2.2. *The Wilcoxon–Mann–Whitney test versus the t test under additive effects*

For ease of comparison with existing theory under perfect compliance (see, e.g., van der Vaart, 1998, § 14.1), we first focus on constant additive effects. Extensions to general patterns of treatment effects are studied in § 3.1. Specifically, consider the location-shift alternatives

$$H_{\mathrm{A},n} : Y(z,a) = Y(0,0) + (rz + a)\theta, \quad \theta = n^{-1/2}h, \tag{3}$$

where $h > 0$ and $r \in \mathbb{R}$. That is, the treatment and randomization add θ and $r\theta$, respectively, to the outcome. The case with exclusion restriction corresponds to $r = 0$. Combining (2) with (3), we obtain that $\nabla\omega_1(y) - \nabla\omega_0(y) = p_{\mathrm{c}}\partial\{\nu_{\mathrm{c}}(y - \theta - r\theta) - \nu_{\mathrm{c}}(y)\}/\partial\theta\big|_{\theta=0} + p_{\mathrm{a}}\partial\{\nu_{\mathrm{a}}(y - \theta - r\theta) - \nu_{\mathrm{a}}(y - \theta)\}/\partial\theta\big|_{\theta=0} + p_{\mathrm{n}}\partial\{\nu_{\mathrm{n}}(y - r\theta) - \nu_{\mathrm{n}}(y)\}/\partial\theta\big|_{\theta=0} = -\{(p_{\mathrm{c}} + r)\dot{\nu}_{\mathrm{c}}(y) + r\dot{\nu}_{\mathrm{a}}(y) + r\dot{\nu}_{\mathrm{n}}(y)\} = -\{p_{\mathrm{c}}\dot{\nu}_{\mathrm{c}}(y) + r\dot{\nu}(y)\}$. Substituting this into Lemma 1, we arrive at the next theorem.

THEOREM 1. *Under the location-shift contiguous alternatives* $H_{\mathrm{A},n}$ *in* (3)*, the Pitman efficiencies for the intent-to-treat Wilcoxon–Mann–Whitney and t tests are*

$$\begin{aligned}\zeta^2_{\mathrm{WMW}}(p_{\mathrm{c}},\nu) &= 12q(1-q)\big(p_{\mathrm{c}}\langle\dot{\nu},\dot{\nu}_{\mathrm{c}}\rangle + r\langle\dot{\nu},\dot{\nu}\rangle\big)^2,\\ \zeta^2_t(p_{\mathrm{c}},\nu) &= q(1-q)V(\nu)^{-1}(p_{\mathrm{c}} + r)^2,\end{aligned} \tag{4}$$

respectively, where $\langle f, g\rangle = \int f(y)g(y)\,\mathrm{d}y$ *for any Lebesgue square-integrable functions* f *and* g*. As a result, the relative efficiency is*

$$\mathcal{R}(p_{\mathrm{c}},\nu) \equiv \frac{\zeta^2_{\mathrm{WMW}}(p_{\mathrm{c}},\nu)}{\zeta_t(p_{\mathrm{c}},\nu)} = 12V(\nu)\left(\frac{p_{\mathrm{c}}\langle\dot{\nu},\dot{\nu}_{\mathrm{c}}\rangle + r\langle\dot{\nu},\dot{\nu}\rangle}{p_{\mathrm{c}} + r}\right)^2. \tag{5}$$

Simulations described in the Supplementary Material show that asymptotic power functions based on (4) approximate the actual power fairly accurately under finite samples.

Either under perfect compliance so that $p_{\mathrm{c}} = 1$, or under uninformative noncompliance so that $\nu_{\mathrm{c}} = \nu$, the far right-hand side of (5) reduces to $12V(\nu)\langle\dot{\nu},\dot{\nu}\rangle^2$, recovering the classical results for the two tests (see, e.g., van der Vaart, 1998, Example 14.13). In that sense, the impact of informative noncompliance shows in the difference between $\mathcal{R}(p_{\mathrm{c}},\nu)$ and $\mathcal{R}(1,\nu) = 12V(\nu)\langle\dot{\nu},\dot{\nu}\rangle^2$, which, in turn because of (5), is determined by the difference between $\langle\dot{\nu},\dot{\nu}_{\mathrm{c}}\rangle$ and $\langle\dot{\nu},\dot{\nu}\rangle$. The following proposition says that this difference depends crucially on how strongly being a complier is correlated with the height of the outcome density.

PROPOSITION 1. *Let* $Y = Y(0,0)$ *and* $\delta = I\{A(1) = 1, A(0) = 0\}$*. Then* $\langle\dot{\nu},\dot{\nu}_{\mathrm{c}}\rangle - \langle\dot{\nu},\dot{\nu}\rangle = p_{\mathrm{c}}^{-1}\mathrm{cov}\{\delta,\dot{\nu}(Y)\}$ *so that*

$$\mathcal{R}(p_{\mathrm{c}},\nu) = 12V(\nu)\left[\langle\dot{\nu},\dot{\nu}\rangle + (p_{\mathrm{c}} + r)^{-1}\mathrm{cov}\{\delta,\dot{\nu}(Y)\}\right]^2. \tag{6}$$

Proof. With $Y \sim \nu$, since $E(\delta - p_c) = 0$ we have that

$$\operatorname{cov}\{\delta, \dot{\nu}(Y)\} = E\{(\delta - p_c)\dot{\nu}(Y)\} = p_c E\{\dot{\nu}(Y) \mid \delta = 1\} - p_c\langle \dot{\nu}, \dot{\nu}\rangle = p_c(\langle \dot{\nu}, \dot{\nu}_c\rangle - \langle \dot{\nu}, \dot{\nu}\rangle).$$

The identity in (6) then follows from (5). □

Using Proposition 1, we obtain via standard calculations that

$$\mathcal{R}(p_c, \nu) - \mathcal{R}(1, \nu) = 12V(\nu)(p_c + r)^{-2}\operatorname{cov}\{\delta, \dot{\nu}(Y)\}\{p_c\langle \dot{\nu}, \dot{\nu}_c\rangle + (p_c + 2r)\langle \dot{\nu}, \dot{\nu}\rangle\}.$$

This means that the sign of $\mathcal{R}(p_c, \nu) - \mathcal{R}(1, \nu)$ coincides with that of $\operatorname{cov}\{\delta, \dot{\nu}(Y)\}$. Hence, roughly speaking, noncompliance favours the Wilcoxon–Mann–Whitney test over the t test if the compliers are more likely to appear in high-density regions or, conversely, if noncompliers are more likely to appear in the tails. Consider a simple situation where $\dot{\nu}$ is unimodal and symmetric about zero and $\operatorname{pr}(\delta = 1 \mid Y = y)$ is nonincreasing in $|y|$; then one can easily show that $\operatorname{cov}\{\delta, \dot{\nu}(Y)\} \geqslant 0$ so that $\mathcal{R}(p_c, \nu) \geqslant \mathcal{R}(1, \nu)$.

The correspondence between the relative location of compliers/noncompliers and the relative efficiency of the two tests has an intuitive explanation rooted in classical theory, which tells us that the rank-based Wilcoxon–Mann–Whitney test is more robust against noise in the tails. In an intent-to-treat analysis, the compliers are the sole carrier of the treatment signal between the randomized groups, with the noncompliers contributing nothing but random noise. Hence, if the noise contributors tend to reside in the tails, their influence is better contained by the Wilcoxon–Mann–Whitney test than by the t test, allowing the former to pick out the signal more easily.

This qualitative relationship can be carried over to quantitative limits. By (6), the range of $\mathcal{R}(p_c, \nu)$ depends on that of $\operatorname{cov}\{\delta, \dot{\nu}(Y)\}$. Intuitively, $\operatorname{cov}\{\delta, \dot{\nu}(Y)\}$ should be maximized or minimized when the compliers are selected according to the largest or smallest density values, respectively. To make this precise, denote the range of ν_c under fixed ν and p_c by

$$\mathcal{P}(p_c \mid \nu) = \{\nu_c : \nu = p_c\nu_c + (1 - p_c)\eta \ \text{ for some distribution } \eta\}.$$

We can obtain bounds for $\{\langle \dot{\nu}, \dot{\nu}_c\rangle : \nu_c \in \mathcal{P}(p_c \mid \nu)\}$ using the following lemma, whose proof can be found in the Supplementary Material.

LEMMA 2. *Given a ν-integrable function g, let*

$$\begin{aligned} \dot{\nu}_g^{(L)}(y) &= p_c^{-1} I\{g(y) < l\}\,\dot{\nu}(y) + c^{(L)} I\{g(y) = l\}\,\dot{\nu}(y), \\ \dot{\nu}_g^{(U)}(y) &= p_c^{-1} I\{g(y) > u\}\,\dot{\nu}(y) + c^{(U)} I\{g(y) = u\}\,\dot{\nu}(y), \end{aligned} \tag{7}$$

where $l, u \in \mathbb{R}$ and $c^{(L)}, c^{(U)} \in [0, p_c^{-1}]$ are such that $\int \dot{\nu}_g^{(L)}(y)\,\mathrm{d}y = \int \dot{\nu}_g^{(U)}(y)\,\mathrm{d}y = 1$. Then

$$\{\langle g, \dot{\nu}_c\rangle : \nu_c \in \mathcal{P}(p_c \mid \nu)\} = \big[\langle g, \dot{\nu}_g^{(L)}\rangle, \langle g, \dot{\nu}_g^{(U)}\rangle\big].$$

Replacing g in Lemma 2 with $\dot{\nu}$, we obtain the range of $\langle \dot{\nu}, \dot{\nu}_c\rangle$ and, by (5), that of $\mathcal{R}(p_c, \nu)$.

THEOREM 2. *Under $H_{A,n}$ in (3) with fixed (p_c, ν), the range of the relative efficiency between the intent-to-treat Wilcoxon–Mann–Whitney and t tests is*

$$12V(\nu)\left(\frac{p_c\langle \dot{\nu}, \dot{\nu}_c^{(L)}\rangle + r\langle \dot{\nu}, \dot{\nu}\rangle}{p_c + r}\right)^2 \leqslant \mathcal{R}(p_c, \nu) \leqslant 12V(\nu)\left(\frac{p_c\langle \dot{\nu}, \dot{\nu}_c^{(U)}\rangle + r\langle \dot{\nu}, \dot{\nu}\rangle}{p_c + r}\right)^2, \tag{8}$$

where $\{\dot{\nu}_c^{(L)}, \dot{\nu}_c^{(U)}\}$ is $\{\dot{\nu}_g^{(L)}, \dot{\nu}_g^{(U)}\}$ defined in (7) with g replaced by $\dot{\nu}$.

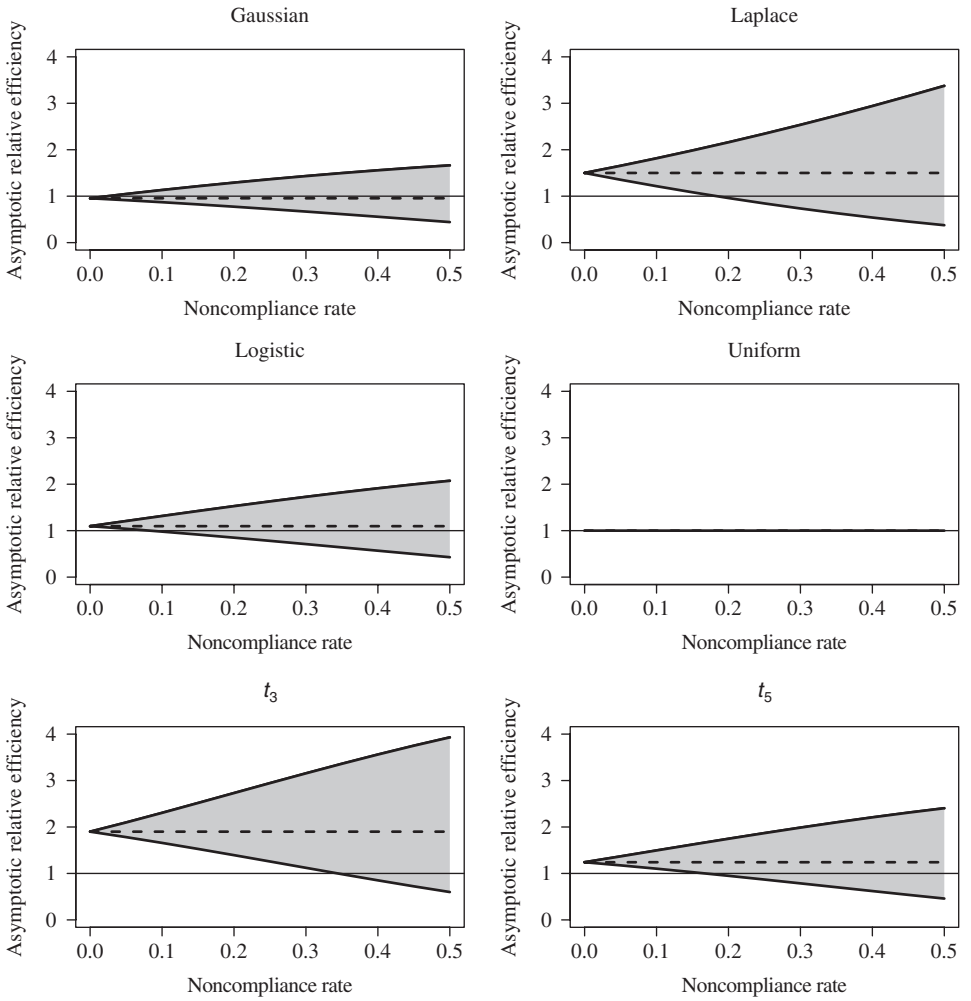

Fig. 1. The range of asymptotic relative efficiency of the intent-to-treat Wilcoxon–Mann–Whitney test versus the t test against $H_{\mathrm{A},n}$ under exclusion restriction for different outcome distributions. The dashed line in each panel represents the asymptotic relative efficiency under perfect compliance; t_d refers to the t distribution with d degrees of freedom.

For a unimodal symmetric $\dot{\nu}$, it is clear from the above results that the least and most favourable scenarios for the Wilcoxon–Mann–Whitney test are those in which the compliers exclusively occupy the tails or modal regions, respectively. Standard derivation along this line gives us

$$\langle \dot{\nu}, \dot{\nu}_{\mathrm{c}}^{(\mathrm{L})} \rangle = 2p_c^{-1} \int_{-\infty}^{\nu^{-1}(p_{\mathrm{c}}/2)} \dot{\nu}(y)^2 \, \mathrm{d}y, \quad \langle \dot{\nu}, \dot{\nu}_{\mathrm{c}}^{(\mathrm{U})} \rangle = 2p_c^{-1} \int_{0}^{\nu^{-1}\{(1+p_{\mathrm{c}})/2\}} \dot{\nu}(y)^2 \, \mathrm{d}y. \tag{9}$$

Using (9) in conjunction with (8), we derive the range of $\mathcal{R}(p_{\mathrm{c}}, \nu)$ as a function of the noncompliance rate $1 - p_{\mathrm{c}}$ for Gaussian, Laplace, logistic, uniform and t distributions under exclusion restriction, i.e., $r = 0$. The results are plotted in Fig. 1, with the underlying analytical formulae given in the Supplementary Material. The dashed lines represent the asymptotic relative efficiency values under perfect compliance, which agree with the numbers tabulated in Table 14.12 of van der Vaart (1998). Not surprisingly, $\mathcal{R}(p_{\mathrm{c}}, \nu) \equiv 1$ for the uniform distribution, whose constant density makes the upper and lower bounds the same. Otherwise, we find by numerical calculation that the lower bounds of $\mathcal{R}(p_{\mathrm{c}}, \nu)$ under the Laplace, logistic, t_3 and t_5 distributions intersect with the unit horizontal line at $1 - p_{\mathrm{c}} = 18.4\%, 8.3\%, 34.4\%$ and 16.7%, respectively. This means that the Wilcoxon–Mann–Whitney test is always more efficient

than the t test for these distributions, as long as the noncompliance rate is below the corresponding threshold.

2.3. *Practical considerations*

The theoretical results in § 2.2 are practically instructive in two ways. First, we can use Theorem 1 to pick the more efficient test for analysing a current trial in the presence of noncompliance. In (5) with $r = 0$, for example, p_c can be estimated by the sample analogue of $E(A = 1 \mid Z = 1) - E(A = 1 \mid Z = 0)$ (Angrist et al., 1996). Moreover, $\dot{\nu}(\cdot)$ and $\dot{\nu}_c(\cdot)$ can be estimated by a kernel density estimator based on the randomized control group and one using the methods of Imbens & Rubin (1997), respectively. Plugging in these estimates, we obtain an estimate for $\mathcal{R}(p_c, \nu)$, telling us which of the Wilcoxon–Mann–Whitney and t tests is likely to be more powerful. Intuitively, this data-dependent selection of test should not affect the Type I error rate as it does not use the part of the data that is informative about the treatment effect.

As a demonstration, the above procedure was run on data collected from 11 204 participants in the well-known U.S. National Job Training Partnership Act Study, a randomized trial conducted between 1987 and 1989 to assess the effect of a job-training programme on the trainee's subsequent earnings (Abadie et al., 2002). The estimated compliance rate is $p_c = 62.7\%$. As shown in the Supplementary Material, both $\dot{\nu}$ and $\dot{\nu}_c$ appear heavily bimodal. Using these density estimates, we find that $\mathcal{R}(p_c, \nu) \approx 4.9$, suggesting that the intent-to-treat Wilcoxon–Mann–Whitney test is likely much more powerful than the t test. This is substantiated by the actual test results, with the former producing a p-value of 0.001 and the latter 0.023.

Second, one can calculate the sample size for a future trial by inverting the power function with the derived Pitman efficiencies. By Theorem 1 with $r = 0$, the Pitman efficiency for the t test is p_c^2 times that under perfect compliance. Provided that a good estimate for p_c can be obtained from historical data, the investigator need only inflate the standard sample size by p_c^{-2}. The same is true for the Wilcoxon–Mann–Whitney test under uninformative noncompliance. With informative noncompliance, and without further knowledge about the compliance structure, the best one can do is use the lower bound on $\langle \dot{\nu}, \dot{\nu}_c \rangle$ constructed by Lemma 2 to obtain a conservative estimate for the sample size. For example, under exclusion restriction with additive treatment effect θ in units of standard deviation, it can be shown that the sample sizes needed to achieve a power of $1 - \beta$ for the Gaussian and Laplace distributions are bounded by

$$\frac{\pi (z_{1-\alpha/2} + z_{1-\beta})^2}{12q(1-q)\Phi\{2^{1/2}\Phi^{-1}(p_c/2)\}^2\theta^2}, \quad \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2}{3q(1-q)p_c^4\theta^2},$$

respectively. As an example, the data from the above job-training study are used as pilot data to calculate the sample size for a new trial. The details are provided in the Supplementary Material.

3. EXTENSIONS AND DISCUSSIONS

3.1. *Inhomogeneous effects*

The framework of § 2.1 accommodates all patterns of treatment and randomization effects, to the extent that $\nabla\omega_1(y) - \nabla\omega_0(y)$ can be derived under the specific contiguous alternatives. Indeed, consider a general alternative hypothesis that transforms the null outcome by a function $\mathcal{T}$ indexed by treatment and randomization effect parameters θ and $r\theta$, respectively; that is,

$$H_{A,n} : Y(z, a) = \mathcal{T}\{Y(0, 0); a\theta, zr\theta\},$$

where $\mathcal{T}(y; 0, 0) = y$ for all $y \in \mathbb{R}$. For example, the location-shift alternative in (3) corresponds to $\mathcal{T}(y; a\theta, zr\theta) = y + a\theta + zr\theta$. Multiplicative effects can be formulated as $\mathcal{T}(y; a\theta, zr\theta) = y \exp(a\theta + zr\theta)$. Sometimes subjects with lower baseline outcomes stand to gain more from the treatment than do those with already high values. In such cases, it would be appropriate to consider $\mathcal{T}(y; a\theta, zr\theta) = y + (rz + a)\theta\kappa(y)$,

where $\theta > 0$ and $\kappa(y)$ is a decreasing function, e.g., $\kappa(y) = \{1+\exp(cy)\}^{-1}$ for some $c > 0$. The following proposition, proved in the Supplementary Material, expresses $\nabla\omega_1(y) - \nabla\omega_0(y)$ in terms of a tangent function of $\mathcal{T}(y; a\theta, zr\theta)$ as $\theta \to 0$.

LEMMA 3. *Under regularity conditions specified in the Supplementary Material, suppose that* $\partial\mathcal{T}(y; a\theta, zr\theta)/\partial\theta\big|_{\theta=0} = \dot{\mathcal{T}}_{a,z}(y)$ *for some function* $\dot{\mathcal{T}}_{a,z}(y)$. *Write* $\dot{\mathcal{T}}_{\mathrm{c}}(y) = \dot{\mathcal{T}}_{1,1}(y) - \dot{\mathcal{T}}_{0,0}(y)$, $\dot{\mathcal{T}}_{\mathrm{a}}(y) = \dot{\mathcal{T}}_{1,1}(y) - \dot{\mathcal{T}}_{1,0}(y)$ *and* $\dot{\mathcal{T}}_{\mathrm{n}}(y) = \dot{\mathcal{T}}_{0,1}(y) - \dot{\mathcal{T}}_{0,0}(y)$. *Then*

$$\nabla\omega_1(y) - \nabla\omega_0(y) = -\big\{p_{\mathrm{c}}\dot{\mathcal{T}}_{\mathrm{c}}(y)\dot{\nu}_{\mathrm{c}}(y) + p_{\mathrm{a}}\dot{\mathcal{T}}_{\mathrm{a}}(y)\dot{\nu}_{\mathrm{a}}(y) + p_{\mathrm{n}}\dot{\mathcal{T}}_{\mathrm{n}}(y)\dot{\nu}_{\mathrm{n}}(y)\big\}.$$

Under the inhomogeneous additive model $\mathcal{T}(y; a\theta, zr\theta) = y + (rz+a)\theta\kappa(y)$, for example, it is immediate that $\dot{\mathcal{T}}_{a,z}(y) = (rz + a)\kappa(y)$ so that $\dot{\mathcal{T}}_{\mathrm{c}}(y) = (1 + r)\kappa(y)$, $\dot{\mathcal{T}}_{\mathrm{a}}(y) = \dot{\mathcal{T}}_{\mathrm{n}}(y) = r\kappa(y)$ and, by Lemma 3, $\nabla\omega_1(y) - \nabla\omega_0(y) = -\kappa(y)\{p_{\mathrm{c}}\dot{\nu}_{\mathrm{c}}(y) + r\dot{\nu}(y)\}$. Using the last expression in Lemma 1, we obtain that

$$\begin{aligned}\zeta^2_{\mathrm{WMW}}(p_{\mathrm{c}}, \nu) &= 12q(1-q)\left(p_{\mathrm{c}}\langle\kappa\dot{\nu}, \dot{\nu}_{\mathrm{c}}\rangle + r\langle\kappa\dot{\nu}, \dot{\nu}\rangle\right)^2,\\ \zeta^2_t(p_{\mathrm{c}}, \nu) &= q(1-q)V(\nu)^{-1}\left(p_{\mathrm{c}}\langle\kappa, \dot{\nu}_{\mathrm{c}}\rangle + r\langle\kappa, \dot{\nu}\rangle\right)^2.\end{aligned} \tag{10}$$

Taking $\kappa(y) \equiv 1$ reduces (10) to (4) of Theorem 1. Although the Pitman efficiencies under nonconstant $\kappa(\cdot)$ appear a bit more complex, one can nonetheless use Lemma 2 to derive bounds for them. The case with multiplicative effects is considered in the Supplementary Material.

3.2. *General intent-to-treat tests*

The intent-to-treat Wilcoxon–Mann–Whitney and t tests can both be formulated as contrasts between the empirical distributions in the randomized groups. As seen from the main steps in § 2, what determines their respective asymptotic properties is the functional derivative, or more precisely the Hadamard derivative (van der Vaart, 1998, Ch. 20), of the contrast in question. This helps us generalize the previous results to a broad class of nonparametric intent-to-treat tests. The following proposition is proved in the Supplementary Material.

PROPOSITION 2. *Consider an intent-to-treat test that rejects* H_0 *if* $n^{1/2}|\mathcal{H}(\hat{\omega}_1, \hat{\omega}_0)|/\hat{\sigma}_n > z_{1-\alpha/2}$, *where* $\mathcal{H}(\cdot, \cdot)$ *is a functional satisfying* $\mathcal{H}(\omega_1, \omega_0) = -\mathcal{H}(\omega_0, \omega_1)$ *for all* ω_1 *and* ω_0, *and* $\hat{\sigma}^2_n$ *is a consistent variance estimator for* $n^{1/2}|\mathcal{H}(\hat{\omega}_1, \hat{\omega}_0)|$. *Let* $\dot{\mathcal{H}}_\nu$ *denote the Hadamard derivative of* $\mathcal{H}(\omega_1, \omega_0)$ *with respect to* ω_1 *at* $\omega_1 = \omega_0 = \nu$. *Then the Pitman efficiency of the test is*

$$\zeta^2_{\mathcal{H}} = q(1-q)\left[\int \dot{\mathcal{H}}^2_\nu\{I(y \leqslant \cdot) - \nu(\cdot)\}\dot{\nu}(y)\,\mathrm{d}y\right]^{-1}\dot{\mathcal{H}}^2_\nu(\nabla\omega_1 - \nabla\omega_0). \tag{11}$$

In (11), the expressions for $\nabla\omega_1 - \nabla\omega_0$ and $\dot{\mathcal{H}}_\nu$ are determined by the contiguous alternatives and the specific test, respectively. To recover the results of Lemma 1 for the Wilcoxon–Mann–Whitney and t tests under location-shift alternatives, simply plug in $\dot{\mathcal{H}}_\nu\{\psi(\cdot)\} = -\int\psi(y)\dot{\nu}(y)\,\mathrm{d}y$ and $-\int\psi(y)\,\mathrm{d}y$, respectively, along with $\nabla\omega_1 - \nabla\omega_0 = -\{p_{\mathrm{c}}\dot{\nu}_{\mathrm{c}}(y) + r\dot{\nu}(y)\}$. For a new example, consider the τ-quantile test based on $\mathcal{H}(\hat{\omega}_1, \hat{\omega}_0) = \hat{\omega}_1^{-1}(\tau) - \hat{\omega}_0^{-1}(\tau)$ $(0 < \tau < 1)$. The Hadamard derivative of this $\mathcal{H}$ is $\dot{\mathcal{H}}_\nu\{\psi(\cdot)\} = -\dot{\nu}(z_{\nu,\tau})^{-1}\psi(z_{\nu,\tau})$, where $z_{\nu,\tau} = \nu^{-1}(\tau)$ (e.g., van der Vaart, 1998, Lemma 21.3). A straightforward derivation in the Supplementary Material shows that the Pitman efficiency under the location-shift alternatives is $\zeta^2_\tau = q(1-q)\tau^{-1}(1-\tau)^{-1}\left\{p_{\mathrm{c}}\dot{\nu}_{\mathrm{c}}(z_{\nu,\tau}) + r\dot{\nu}(z_{\nu,\tau})\right\}^2$. As one might have guessed, the τ-quantile test is powerful when the neighbourhood of the corresponding quantile is densely populated by compliers, driving up the value of $\dot{\nu}_{\mathrm{c}}(z_{\nu,\tau})$. Formal analysis and comparison with the Wilcoxon–Mann–Whitney and t tests is left to the interested reader.

ACKNOWLEDGEMENT

This research was supported by the U.S. National Science Foundation and National Institutes of Health. The author very much appreciates the helpful comments from the editor, associate editor and two referees.

SUPPLEMENTARY MATERIAL

Supplementary Material includes technical and numerical results.

APPENDIX

Proof of Lemma 1. The null asymptotic variance of $T_{\mathrm{WMW},n}$ is the distribution-free constant $12^{-1}q^{-1}(1-q)^{-1}$ by Example 14.13 in van der Vaart (1998). On the other hand,

$$\dot{\mu}_{\mathrm{WMW}}(0) = \frac{\partial}{\partial\theta}\bigg|_{\theta=0} \int \omega_0(y \mid \theta)\omega_1(\mathrm{d}y \mid \theta) = -\int \{\nabla\omega_1(y) - \nabla\omega_0(y)\}\dot{\nu}(y)\,\mathrm{d}y,$$

where the second equality follows from integration by parts. The case for ζ_t^2 can be similarly derived. □

REFERENCES

ABADIE, A., ANGRIST, J. & IMBENS, G. (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica* **70**, 91–117.

ANGRIST, J. D., IMBENS, G. W. & RUBIN, D. B. (1996). Identification of causal effects using instrumental variables. *J. Am. Statist. Assoc.* **91**, 444–55.

GUPTA, S. K. (2011). Intention-to-treat concept: A review. *Perspect. Clin. Res.* **2**, 109–12.

IMBENS, G. W. & RUBIN, D. B. (1997). Estimating outcome distributions for compliers in instrumental variables models. *Rev. Econ. Stud.* **64**, 555–74.

IMBENS, G. W. & RUBIN, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge: Cambridge University Press.

MANN, H. B. & WHITNEY, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statist.* **18**, 50–60.

RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6**, 34–58.

SIDAK, Z., SEN, P. K. & HAJEK, J. (1999). *Theory of Rank Tests*. London: Academic Press.

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

WILCOXON, F. (1945). Individual comparisons by ranking methods. *Biomet. Bull.* **1**, 80–3.

[*Received on* 4 *October* 2020. *Editorial decision on* 13 *October* 2021]