POINTS OF SIGNIFICANCE

Nonparametric tests

Nonparametric tests robustly compare skewed or ranked data.

We have seen that the *t*-test is robust with respect to assumptions about normality and equivariance[1] and thus is widely applicable. There is another class of methods—nonparametric tests—more suitable for data that come from skewed distributions or have a discrete or ordinal scale. Nonparametric tests such as the sign and Wilcoxon rank-sum tests relax distribution assumptions and are therefore easier to justify, but they come at the cost of lower sensitivity owing to less information inherent in their assumptions. For small samples, the performance of these tests is also constrained because their *P* values are only coarsely sampled and may have a large minimum. Both issues are mitigated by using larger samples.

These tests work analogously to their parametric counterparts: a test statistic and its distribution under the null are used to assign significance to observations. We compare in **Figure 1** the one-sample *t*-test[2] to a nonparametric equivalent, the sign test (though more sensitive and sophisticated variants exist), using a putative sample *X* whose source distribution we cannot readily identify (**Fig. 1a**). The null hypothesis of the sign test is that the sample median m_X is equal to the proposed median, $M = 0.4$. The test uses the number of sample values larger than *M* as its test statistic, *W*—under the null we expect to see as many values below the median as above, with the exact probability given by the binomial distribution (**Fig. 1c**). The median is a more useful descriptor than the mean for asymmetric and otherwise irregular distributions. The sign test makes no assumptions about the distribution—only that sample values be independent. If we propose that the population median is $M = 0.4$ and we observe *X*, we find $W = 5$ (**Fig. 1b**). The chance of observing a value of *W* under the null that is at least as extreme ($W \le 1$ or $W \ge 5$) is $P = 0.22$, using both tails of the binomial distribution (**Fig. 1c**). To limit the test to whether the median of *X* was biased towards values larger than *M*, we would consider only the area for $W \ge 5$ in the right tail to find $P = 0.11$.

The *P* value of 0.22 from the sign test is much higher than that from the *t*-test ($P = 0.04$), reflecting that the sign test is less sensitive. This is because it is not influenced by the actual distance between the sample values and *M*—it measures only 'how many' instead of 'how much'. Consequently, it needs larger sample sizes or more supporting evidence than the *t*-test. For the example of *X*, to obtain $P < 0.05$ we would need to have all values larger than *M* ($W = 6$). Its large *P* values and straightforward application makes the sign test a useful diagnostic. Take, for example, a hypothetical situation slightly different from that in **Figure 1**, where $P > 0.05$ is reported for the case where a treatment has lowered blood pressure in 6 out of 6 subjects. You may think this *P* seems implausibly large, and you'd be right because the equivalent scenario for the sign test ($W = 6$, $n = 6$) gives a two-tailed $P = 0.03$.

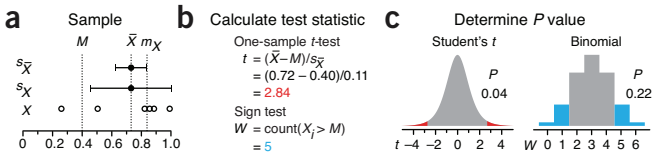

Figure 1 | A sample can be easily tested against a reference value using the sign test without any assumptions about the population distribution. (**a**) Sample *X* ($n = 6$) is tested against a reference $M = 0.4$. Sample mean $\bar{X}$ is shown with s.d. (s_X) and s.e.m. error bars ($s_{\bar{X}}$). m_x is sample median. (**b**) The *t*-statistic compares $\bar{X}$ to *M* in units of s.e.m. The sign test's *W* is the number of sample values larger than *M*. (**c**) Under the null, *t* follows Student's *t*-distribution with five degrees of freedom, whereas *W* is described by the binomial with 6 trials and $P = 0.5$. Two-tailed *P* values are shown.

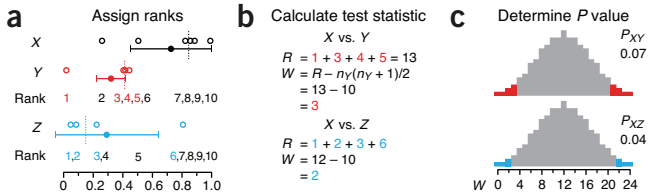

Figure 2 | Many nonparametric tests are based on ranks. (**a**) Sample comparisons of *X* vs. *Y* and *X* vs. *Z* start with ranking pooled values and identifying the ranks in the smaller-sized sample (e.g., 1, 3, 4, 5 for *Y*; 1, 2, 3, 6 for *Z*). Error bars show sample mean and s.d., and sample medians are shown by vertical dotted lines. (**b**) The Wilcoxon rank-sum test statistic *W* is the difference between the sum of ranks and the smallest possible observed sum. (**c**) For small sample sizes the exact distribution of *W* can be calculated. For samples of size (6, 4), there are only 210 different rank combinations corresponding to 25 distinct values of *W*.

To compare two samples, the Wilcoxon rank-sum test is widely used and is sometimes referred to as the Mann-Whitney or Mann-Whitney-Wilcoxon test. It tests whether the samples come from distributions with the same median. It doesn't assume normality, but as a test of equality of medians, it requires both samples to come from distributions with the same shape. The Wilcoxon test is one of many methods that reduce the dynamic range of values by converting them to their ranks in the list of ordered values pooled from both samples (**Fig. 2a**). The test statistic, *W*, is the degree to which the sum of ranks is larger than the lowest possible in the sample with the lower ranks (**Fig. 2b**). We expect that a sample from a population with a smaller median will be converted to a set of smaller ranks.

Because there is a finite number (210) of combinations of rank-ordering for *X* ($n_x = 6$) and *Y* ($n_Y = 4$), we can enumerate all outcomes of the test and explicitly construct the distribution of *W* (**Fig. 2c**) to assign a *P* value to *W*. The smallest value of $W = 0$ occurs when all values in one sample are smaller than those in the other. When they are all larger, the statistic reaches a maximum, $W = n_X n_Y = 24$. For *X* versus *Y*, $W = 3$, and there are 14 of 210 test outcomes with $W \le 3$ or $W \ge 21$. Thus, $P_{XY} = 14/210 = 0.067$. For *X* versus *Z*, $W = 2$, and $P_{XZ} = 8/210 = 0.038$. For cases in which both samples are larger than 10, *W* is approximately normal, and we can obtain the *P* value from a *z*-test of $(W - \mu_W)/\sigma_W$, where $\mu_W = n_1(n_1 + n_2 + 1)/2$ and $\sigma_W = \sqrt{(\mu_W n_2/6)}$.

The ability to enumerate all outcomes of the test statistic makes calculating the *P* value straightforward (**Figs. 1c** and **2c**), but there is an important consequence: there will be a minimum *P* value, P_{min}. Depending on the size of samples, P_{min} can be relatively large. For comparisons of samples of size $n_X = 6$ and $n_Y = 4$ (**Fig. 2a**), $P_{min} = 1/210 = 0.005$ for a one-tailed test, or 0.01 for a two-tailed test, corresponding to $W = 0$. Moreover, because there are only 25 distinct values of *W* (**Fig. 2c**), only two other two-tailed *P* values are <0.05: $P = 0.02$ ($W = 1$) and $P = 0.038$ ($W = 2$). The next-largest *P* value ($W = 3$) is $P = 0.07$. Because there is no *P* with value 0.05, the test cannot be set to reject the null at a type I rate of 5%. Even if we test at $\alpha = 0.05$, we will be rejecting the null at the

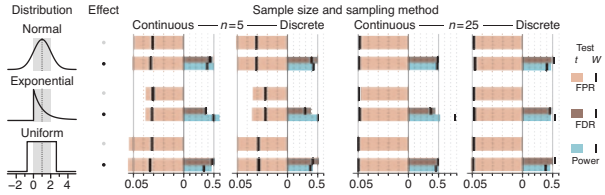

Figure 3 | The Wilcoxon rank-sum test can outperform the *t*-test in the presence of discrete sampling or skew. Data were sampled from three common analytical distributions with $\mu = 1$ (dotted lines) and $\sigma = 1$ (gray bars, $\mu \pm \sigma$). Discrete sampling was simulated by rounding values to the nearest integer. The FPR, FDR and power of Wilcoxon tests (black lines) and *t*-tests (colored bars) for 100,000 sample pairs for each combination of sample size ($n = 5$ and 25), effect chance (0 and 10%) and sampling method. In the absence of an effect, both sample values were drawn from a given distribution type with $\mu = 1$. With effect, the distribution for the second sample was shifted by d ($d = 1.4$ for $n = 5$; $d = 0.57$ for $n = 25$). The effect size was chosen to yield 50% power for the *t*-test in the normal noise scenario. Two-tailed P at $\alpha = 0.05$.

next lower P—for an effective type I error of 3.8%. We will see how this affects test performance for small samples further on. In fact, it may even be impossible to reach significance at $\alpha = 0.05$ because there is a limited number of ways in which small samples can vary in the context of ranks, and no outcome of the test happens less than 5% of the time. For example, samples of size 4 and 3 offer only 35 arrangements of ranks and a two-tailed $P_{min} = 2/35 = 0.057$. Contrast this to the *t*-test, which can produce any P value because the test statistic can take on an infinite number of values.

This has serious implications in multiple-testing scenarios discussed in the previous column[3]. Recall that when N tests are performed, multiple-testing corrections will scale the smallest P value to NP. In the same way as a test may never yield a significant result ($P_{min} > \alpha$), applying multiple-testing correction may also preclude it ($NP_{min} > \alpha$). For example, making $N = 6$ comparisons on samples such as X and Y shown in **Figure 2a** ($n_X = 6$, $n_Y = 4$) will never yield an adjusted P value lower than $\alpha = 0.05$ because $P_{min} = 0.01 > \alpha/N$. To achieve two-tailed significance at $\alpha = 0.05$ across $N = 10$, 100 or 1,000 tests, we require sample sizes that produce at least 400, 4,000 or 40,000 distinct rank combinations. This is achieved for sample pairs of size of (5, 6), (7, 8) and (9, 9), respectively.

The P values from the Wilcoxon test ($P_{XY} = 0.07$, $P_{XZ} = 0.04$) in **Figure 2a** appear to be in conflict with those obtained from the *t*-test ($P_{XY} = 0.04$, $P_{XZ} = 0.06$). The two methods tell us contradictory information—or do they? As mentioned, the Wilcoxon test concerns the median, whereas the *t*-test concerns the mean. For asymmetric distributions, these values can be quite different, and it is conceivable that the medians are the same but the means are different. The *t*-test does not identify the difference in means of X and Z as significant because the standard deviation, s_Z, is relatively large owing to the influence of the sample's largest value (0.81). Because the *t*-test reacts to any change in any sample value, the presence of outliers can easily influence its outcome when samples are small. For example, simply increasing the largest value in X (1.00) by 0.3 will increase s_X from 0.28 to 0.35 and result in a P_{XY} value that is no longer significant at $\alpha = 0.05$. This change does not alter the Wilcoxon P value because the rank scheme remains unaltered. This insensitivity to changes in the data—outliers and typical effects alike—reduces the sensitivity of rank methods.

The fact that the output of a rank test is driven by the probability that a value drawn from distribution A will be smaller (or larger) than one drawn from B without regard to their absolute difference has an interesting consequence: we cannot use this probability (pairwise preferences, in general) to impose an order on distributions. Consider a case of three equally prevalent diseases for which treatment A has cure times of 2, 2 and 5 days for the three diseases, and treatment B has 1, 4 and 4. Without treatment, each disease requires 3 days to cure—let's call this control C. Treatment A is better than C for the first two diseases but not the third, and treatment B is better only for the first. Can we determine which of the three options (A, B, C) is better? If we try to answer this using the probability of observing a shorter time to cure, we find $P(A < C) = 67\%$ and $P(C < B) = 67\%$ but also that $P(B < A) = 56\%$—a rock-paper-scissors scenario.

The question about which test to use does not have an unqualified answer—both have limitations. To illustrate how the *t*- and Wilcoxon tests might perform in a practical setting, we compared their false positive rate (FPR), false discovery rate (FDR) and power at $\alpha = 0.05$ for different sampling distributions and sample sizes ($n = 5$ and 25) in the presence and absence of an effect (**Fig. 3**). At $n = 5$, Wilcoxon FPR = $0.032 < \alpha$ because this is the largest P value it can produce smaller than α, not because the test inherently performs better. We can always reach this FPR with the *t*-test by setting $\alpha = 0.032$, where we'll find that it will still have slightly higher power than a Wilcoxon test that rejects at this rate. At $n = 5$, Wilcoxon performs better for discrete sampling—the power (0.43) is essentially the same as the *t*-test's (0.46), but the FDR is lower. When both tests are applied at $\alpha = 0.032$, Wilcoxon power (0.43) is slightly higher than *t*-test power (0.39). The differences between the tests for $n = 25$ diminishes because the number of arrangements of ranks is extremely large and the normal approximation to sample means is more accurate. However, one case stands out: in the presence of skew (e.g., exponential distribution), Wilcoxon power is much higher than that of the *t*-test, particularly for continuous sampling. This is because the majority of values are tightly spaced and ranks are more sensitive to small shifts. Skew affects *t*-test FPR and power in a complex way, depending on whether one- or two-tailed tests are performed and the direction of the skew relative to the direction of the population shift that is being studied[4].

Nonparametric methods represent a more cautious approach and remove the burden of assumptions about the distribution. They apply naturally to data that are already in the form of ranks or degree of preference, for which numerical differences cannot be interpreted. Their power is generally lower, especially in multiple-testing scenarios. However, when data are very skewed, rank methods reach higher power and are a better choice than the *t*-test.

Corrected after print 23 May 2014.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Martin Krzywinski & Naomi Altman

1. Krzywinski, M. & Altman, N. *Nat. Methods* **11**, 215–216 (2014).
2. Krzywinski, M. & Altman, N. *Nat. Methods* **10**, 1041–1042 (2013).
3. Krzywinski, M. & Altman, N. *Nat. Methods* **11**, 355–356 (2014).
4. Reineke, D.M, Baggett, J. & Elfessi, A. *J. Stat. Educ.* **11** (2003).

Martin Krzywinski is a staff scientist at Canada's Michael Smith Genome Sciences Centre. Naomi Altman is a Professor of Statistics at The Pennsylvania State University.

Corrigendum: Nonparametric tests

Martin Krzywinski & Naomi Altman
Nat. Methods **11, 467–468 (2014); published online 29 April 2014; corrected after print 23 May 2014**

In the version of this article initially published, the expression X ($n_X = 6$) was incorrectly written as X ($n_Y = 6$). The error has been corrected in the HTML and PDF versions of the article.