Web-based Supplementary Materials for "Fast Approximation of Small p-values in Permutation Tests by Partitioning the Permutations"

Brian D. Segal,* Thomas Braun, Michael Elliott, and Hui Jiang
Department of Biostatistics, University of Michigan
Ann Arbor, Michigan
**email:* bdsegal@umich.edu

Contents

A Proofs

In this appendix, we find the limiting distribution of $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ and $T = |\bar{x} - \bar{y}|$ within each partition, and note the corresponding trend in p-values across the partitions. In the process, we prove the results discussed in Section 3. We structure this appendix around the statistic $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ to help to motivate our discussion, and then extend our results to the statistic $T = |\bar{x} - \bar{y}|$.

As before, we denote the total sample size as N, and we require that $N \geq 2$ to allow for at least one observation in each sample. Let $\{m^N\}_{N=2}^{\infty}$, $\{n_x^N\}_{N=2}^{\infty}$, and $\{n_y^N\}_{N=2}^{\infty}$ be sequences such that $m^N/N \to \tau$ and $n_x^N/N \to \lambda$ as $N \to \infty$, and for all N, $n_y^N = N - n_x^N$. We require that for all N, $0 < m^N \leq n_x^N \leq n_y^N < N$, and similarly, $0 < \tau \leq \lambda \leq 1 - \lambda < 1$. We denote the observed data as $\boldsymbol{x}^N$ and $\boldsymbol{y}^N$, which are $n_x^N \times 1$ and $n_y^N \times 1$ vectors, respectively.

Let $\boldsymbol{\delta}_x^{m^N} = (\delta_{x,1}^{m^N}, \ldots, \delta_{x,n_x^N}^{m^N})'$ and $\boldsymbol{\delta}_y^{m^N} = (\delta_{y,1}^{m^N}, \ldots, \delta_{y,n_y^N}^{m^N})'$ be $n_x^N \times 1$ and $n_y^N \times 1$ indicator vectors, respectively, with 1's corresponding to indices of $\boldsymbol{x}^N$ and $\boldsymbol{y}^N$ that are exchanged for a particular permutation π and zero elsewhere. To be specific, for a permutation $\pi \in \Pi(m^N)$, we define $\delta_{x,i}^{m^N}$ and $\delta_{y,j}^{m^N}$ as

$$\delta_{x,i}^{m^N} = \begin{cases} 1 \text{ if } \pi(i) > n_x^N \\ 0 \text{ if } \pi(i) \leq n_x^N \end{cases} \qquad i = 1, \ldots, n_x^N$$

$$\delta_{y,j}^{m^N} = \begin{cases} 1 \text{ if } \pi(n_x^N + j) \leq n_x^N \\ 0 \text{ if } \pi(n_x^N + j) > n_x^N \end{cases} \qquad j = 1, \ldots, n_y^N.$$

For completeness, we note that for fixed m and $i \neq j$, and dropping dependence on N,

$$
\begin{aligned}
\mathbb{E}[\delta_{x,i}^m] &= m/n_x & \qquad \mathbb{E}[\delta_{y,i}^m] &= m/n_y \\
\operatorname{Var}(\delta_{x,i}^m) &= \frac{m}{n_x}\left(1 - \frac{m}{n_x}\right) & \qquad \operatorname{Var}(\delta_{y,i}^m) &= \frac{m}{n_y}\left(1 - \frac{m}{n_y}\right) \\
\operatorname{Cov}(\delta_{x,i}^m, \delta_{x,j}^m) &= \frac{-m(n_x - m)}{n_x^2(n_x - 1)} & \qquad \operatorname{Cov}(\delta_{y,i}^m, \delta_{y,j}^m) &= \frac{-m(n_y - m)}{n_y^2(n_y - 1)}
\end{aligned}
$$

We denote the ratio of means as $R = \bar{x}/\bar{y}$. With the permutation test, for each permutation π in partition m^N, we calculate the statistic (ignoring for now the max function used earlier)

$$
R(m^N) = \frac{\frac{1}{n_x^N}[(\mathbf{1} - \boldsymbol{\delta}_x^{m^N})'\boldsymbol{x}^N + \boldsymbol{\delta}_y^{m^N\prime}\boldsymbol{y}^N]}{\frac{1}{n_y^N}[\boldsymbol{\delta}_x^{m^N\prime}\boldsymbol{x}^N + (\mathbf{1} - \boldsymbol{\delta}_y^{m^N})'\boldsymbol{y}^N]}.
$$

As for all permutation tests, $R(m^N)$ is conditional on the data. The random quantities are $(\boldsymbol{\delta}_x^{m^N}, \boldsymbol{\delta}_y^{m^N})$, which indexed by N, form a triangular array of identically distributed, dependent random variables. We can rewrite $R(m^N)$ as

$$
\begin{aligned}
R(m^N) &= \frac{n_y^N}{n_x^N}\left(\frac{n_x^N\bar{x} + \left(\sum_{j=1}^{n_y^N}\delta_{y,j}^{m^N}y_j^N - \sum_{i=1}^{n_x^N}\delta_{x,i}^{m^N}x_i^N\right)}{n_y^N\bar{y} - \left(\sum_{j=1}^{n_y^N}\delta_{y,j}^{m^N}y_j^N - \sum_{i=1}^{n_x^N}\delta_{x,i}^{m^N}x_i^N\right)}\right) \\
&= g\Bigg(\underbrace{\sum_{j=1}^{n_y^N}\delta_{y,j}^{m^N}y_j^N - \sum_{i=1}^{n_x^N}\delta_{x,i}^{m^N}x_i^N}_{W(m^N)}\Bigg).
\end{aligned} \tag{1}
$$

Writing $R(m^N)$ as a function of $W(m^N)$ will make it straightforward to generalize our results. We note that conditional on the observed data $\boldsymbol{x}^N$ and $\boldsymbol{y}^N$, all terms in $R(m^N)$ are constant except for $W(m^N)$.

We can further split $W(m^N)$ into

$$
W(m^N) = \underbrace{\sum_{j=1}^{n_y^N}\delta_{y,j}^{m^N}y_j^N}_{W_y(m^N)} - \underbrace{\sum_{i=1}^{n_x^N}\delta_{x,i}^{m^N}x_i^N}_{W_x(m^N)} \tag{2}
$$

Following Theorem 2.8.2 in Lehmann (1999, p. 116), restated in Theorem 1 below, under certain conditions both $W_y(m^N)$ and $W_x(m^N)$ in (2) converge to normal random variables, in which case $W(m^N)$ also converges to a normal random variable.

We make a few observations before stating Theorem 1. The following statements focus on $W_y(m^N)$, but equivalent statements apply to $W_x(m^N)$. First, we note that conditional on $\boldsymbol{y}^N$, $W_y(m^N)$ is the sum of a random sample without replacement of m^N elements from a finite population $\boldsymbol{y}^N = (y_1^N, \ldots, y_{n_y^N}^N)'$. We consider a sequence of populations of increasing size, $\boldsymbol{y}^N, N = 2, 3, \ldots$, and random samples $\boldsymbol{v}^N = (v_1^N, \ldots, v_{m^N}^N)'$ from each $\boldsymbol{y}^N$. To be specific, for fixed $\boldsymbol{\delta}_y^{m^N}$, let $\mathcal{K} = \{j : \delta_{y,j}^{m^N} = 1\}$ be the set of indices corresponding to the selected elements of $\boldsymbol{y}^N$. Then writing $\mathcal{K} = \{k_1, \ldots, k_{m^N}\}$, we have $\boldsymbol{v}^N = (y_{k_1}^N, \ldots, y_{k_{m^N}}^N)'$.

Let $\bar{v}_{m^N} = (1/m^N) \sum_{k=1}^{m^N} v_k^N$, and $\bar{y}_{n_y^N} = (1/n_y^N) \sum_{j=1}^{n_y^N} y_j^N$. Then as shown by Lehmann (1999, p. 116-117),

$$\mathbb{E}[\bar{v}_{m^N} | \boldsymbol{y}^N] = \bar{y}_{n_y^N}$$

$$\mathrm{Var}(\bar{v}_{m^N} | \boldsymbol{y}^N) = \frac{n_y^N - m^N}{m^N(n_y^N - 1)} \frac{1}{n_y^N} \sum_{j=1}^{n_y^N} (y_j^N - \bar{y}_{n_y^N})^2.$$

We can now state Theorem 1.

Theorem 1 (Theorem 2.8.2, Lehmann (1999))**.**

$$\frac{\bar{v}_{m^N} - \mathbb{E}[\bar{v}_{m^N} | \boldsymbol{y}^N]}{\sqrt{\mathrm{Var}(\bar{v}_{m^N} | \boldsymbol{y}^N)}} \to N(0,1)$$

provided that $m^N \to \infty$ and $n_y^N - m^N \to \infty$ as $N \to \infty$, and either of the following two conditions is satisfied:
i) m^N/n_y^N is bounded away from 0 and 1 as $N \to \infty$, and

$$\frac{\max(y_j^N - \bar{y}_{n_y^N})^2}{\sum_j (y_j^N - \bar{y}_{n_y^N})^2} \to 0$$

or
ii)

$$\frac{\max(y_j^N - \bar{y}_{n_y^N})^2}{\sum_j (y_j^N - \bar{y}_{n_y^N})^2 / n_y^N}$$

remains bounded as $N \to \infty$.

For a proof, please see Lehmann (1999) and references therein, particularly the corollary to Lemma 4.1 in Hájek (1961), and Example 4.1 and Section 5 in Hájek (1961). Our constraints on m^N, n_x^N, and n_y^N imply that $m^N \to \infty$ and $n_y^N - m^N \to \infty$ as $N \to \infty$. The other conditions in Theorem 1 require that the contribution of each deviance to the sum of deviances becomes negligible as the sample size becomes large. This excludes data coming from distributions with a non-finite variance, such as the Cauchy distribution.

Applying Theorem 1 to $W(m^N)$ we get Corollary 1.

Corollary 1. *Conditional on $\boldsymbol{x}^N$ and $\boldsymbol{y}^N$, and assuming the conditions in Theorem 1 hold,*

$$\frac{W(m^N) - \mu(m^N)}{\sqrt{V(m^N)}} \to N(0,1),$$

where $\mu(m^N) = \mu_y(m^N) - \mu_x(m^N)$ and $V(m^N) = V_y(m^N) + V_x(m^N)$, with

$$\begin{aligned}\mu_y(m^N) &= \mathbb{E}[W_y(m^N)|\boldsymbol{y}^N] = m^N \bar{y}_{n_y^N}\\ \mu_x(m^N) &= \mathbb{E}[W_x(m^N)|\boldsymbol{x}^N] = m^N \bar{x}_{n_x^N}\end{aligned}$$

and

$$\begin{aligned}V_y(m^N) &= \mathrm{Var}(W_y(m^N)|\boldsymbol{y}^N) = m^N \frac{n_y^N - m^N}{\left(n_y^N - 1\right) n_y^N} \sum_{j=1}^{n_y^N} (y_j^N - \bar{y}_{n_y^N})^2\\ V_x(m^N) &= \mathrm{Var}(W_x(m^N)|\boldsymbol{x}^N) = m^N \frac{n_x^N - m^N}{\left(n_x^N - 1\right) n_x^N} \sum_{i=1}^{n_x^N} (x_i^N - \bar{x}_{n_x^N})^2.\end{aligned}$$

Before proving Corollary 1, we state Lemma 1.

Lemma 1. *For all m and N,* $\mathrm{Cov}\left(W_x(m^N), W_y(m^N)|\boldsymbol{x},\boldsymbol{y}\right) = 0$.

Proof of Lemma 1. First note that for all m, N, i, and j, $\delta_{x,i}^{m^N} \perp \delta_{y,j}^{m^N}$. This is a direct consequence of the sampling procedure implied by the permutation, in which we condition on the number of elements to exchange (m), and then randomly select m elements of $\boldsymbol{x}$ and m elements of $\boldsymbol{y}$. Therefore, dropping dependence on N,

$$\begin{aligned}\mathbb{E}\left[W_x(m)W_y(m)|\boldsymbol{x},\boldsymbol{y}\right] &= \mathbb{E}\left[\left(\sum_i \delta_{x,i}^m x_i\right)\left(\sum_j \delta_{y,j}^m y_j\right)|\boldsymbol{x},\boldsymbol{y}\right]\\ &= \mathbb{E}\left[\sum_i \sum_j \delta_{x,i}^m x_i \delta_{y,j}^m y_j |\boldsymbol{x},\boldsymbol{y}\right]\\ &= \sum_i \sum_j x_i y_j \mathbb{E}\left[\delta_{x,i}^m \delta_{y,j}^m\right]\\ &= \sum_i x_i \mathbb{E}\left[\delta_{x,i}^m\right] \sum_j y_j \mathbb{E}\left[\delta_{y,j}^m\right] \qquad (\delta_{x,i}^m \perp \delta_{y,j}^m)\\ &= \mathbb{E}\left[W_x(m)|\boldsymbol{x}\right]\mathbb{E}\left[W_y(m)|\boldsymbol{y}\right].\end{aligned}$$

Therefore,

$$\begin{aligned}\operatorname{Cov}\left(W_x(m^N), W_y(m^N)|\boldsymbol{x}, \boldsymbol{y}\right) &= \mathbb{E}\left[W_x(m^N)W_y(m^N)|\boldsymbol{x}, \boldsymbol{y}\right] - \mathbb{E}\left[W_x(m^N)|\boldsymbol{x}\right]\mathbb{E}\left[W_y(m^N)|\boldsymbol{y}\right] \\ &= 0\end{aligned}$$

which proves the lemma. □

Now we prove Corollary 1.

Proof of Corollary 1. Working with the first term in (2), we have

$$W_y(m^N) = \sum_{j=1}^{n_y^N} \delta_{y,j}^{m^N} y_j^N = m^N \bar{v}_{m^N}$$

Therefore, as shown by Lehmann (1999, p. 116-117),

$$\mu_y(m^N) = \mathbb{E}[W_y(m^N)|\boldsymbol{y}^N] = m^N \bar{y}_{n_y^N}$$

and

$$\begin{aligned}V_y(m^N) = \operatorname{Var}(W_y(m^N)|\boldsymbol{y}^N) &= \left(m^N\right)^2 \frac{n_y^N - m^N}{m^N(n_y^N - 1)} \frac{1}{n_y^N} \sum_{j=1}^{n_y^N} (y_j^N - \bar{y}_{n_y^N})^2. \\ &= m^N \frac{n_y^N - m^N}{(n_y^N - 1)} \frac{1}{n_y^N} \sum_{j=1}^{n_y^N} (y_j^N - \bar{y}_{n_y^N})^2.\end{aligned}$$

Similarly, working with the second term in (2),

$$\mu_x(m^N) = \mathbb{E}[W_x(m^N)|\boldsymbol{x}^N] = m^N \bar{x}_{n_x^N}$$

$$V_x(m^N) = m^N \frac{n_x^N - m^N}{(n_x^N - 1)} \frac{1}{n_x^N} \sum_{i=1}^{n_x^N} (x_i^N - \bar{x}_{n_x^N})^2.$$

Applying Theorem 1, we have

$$\frac{W_y(m^N) - \mu_y(m^N)}{\sqrt{V_y(m^N)}} = \frac{\bar{v}_{m^N} - \mathbb{E}[\bar{v}_{m^N}|\boldsymbol{y}^N]}{\sqrt{\operatorname{Var}(\bar{v}_{m^N}|\boldsymbol{y}^N)}} \to N(0,1).$$

Similarly, we have

$$\frac{W_x(m^N) - \mu_x(m^N)}{\sqrt{V_x(m^N)}} \to N(0,1).$$

By Lemma 1, we have

$$\operatorname{Var}\left(W_y(m^N) - W_x(m^N) \,\middle|\, \boldsymbol{x}, \boldsymbol{y}\right) = V_y(m^N) + V_x(m^N).$$

Since uncorrelated normal random variables are independent, for N sufficiently large we also have $W_y(m^N) \perp W_x(m^N)$. Then since the sum of independent normal random variables is also normal, for N sufficiently large we have

$$W(m^N) = W_y(m^N) - W_x(m^N) \sim N\left(\mu_y(m^N) - \mu_x(m^N), V_y(m^N) + V_x(m^N)\right).$$

Equivalently, we have

$$\frac{W(m^N) - \mu(m^N)}{\sqrt{V(m^N)}} \to N(0,1)$$

which proves the corollary. □

In the rest of this appendix, we assume that N is sufficiently large for asymptotic normality to hold for any given partition m, so we drop N from the notation.

In Corollary 2 below, we apply the delta method to show that for sufficiently large N, the permutation distribution of the statistic $R(m)$ is normal within each partition.

Corollary 2. *Let $R = g(W)$, and suppose that $g'(\mu(m)) > 0$ exists. Also, suppose the conditions in Theorem 1 hold. Then conditional on the observed data $\boldsymbol{x}, \boldsymbol{y}$, and for N sufficiently large, $R(m) \sim N(\nu(m), \sigma^2(m))$, where the mean $\nu(m)$ and variance $\sigma^2(m)$ are functions of the partition m.*

Proof of Corollary 2. By Corollary 1, W is normal for N sufficiently large. Then by the delta method, $g(W)$ also converges to a normal distribution, which proves the corollary. □

The result in Corollary 2 for the one-sided statistic $R(m)$ leads directly to the following result for its two-sided counterpart $T(m)$, given in Corollary 3 below. However, we first define a new function g^{conj}, the conjugate of g.

Definition 1 (Conjugate g^{conj})**.** *Let $g(W)$ be a function of W, in which the only other terms are the constants n_x, n_y, $\bar{x}$ and $\bar{y}$. The conjugate g^{conj} is formed by switching the place of n_x with n_y, and $\bar{x}$ with $\bar{y}$, and reversing the sign on each occurrence of W.*

For example, for $R = \bar{x}/\bar{y}$, we have

$$g = \frac{n_y}{n_x}\left(\frac{n_x\bar{x} + W}{n_y\bar{y} - W}\right) \qquad\qquad g^{\text{conj}} = \frac{n_x}{n_y}\left(\frac{n_y\bar{y} - W}{n_x\bar{x} + W}\right)$$

and for $R = \bar{x} - \bar{y}$, as shown below, we have

$$g = \bar{x} - \bar{y} + \left(\frac{1}{n_x} + \frac{1}{n_y}\right)W \qquad\qquad g^{\text{conj}} = \bar{y} - \bar{x} - \left(\frac{1}{n_y} + \frac{1}{n_x}\right)W.$$

We note that $(g^{\text{conj}})^{\text{conj}} = g$.

Corollary 3. *Let $T(m) = \max\left(g(W(m)), g^{conj}(W(m))\right)$. Under the conditions of Theorem 1, and assuming $g'(\mu(m)) > 0$ and $(g^{conj})'(\mu(m)) > 0$ exist, then for N sufficiently large,*

$$\Pr\left(T(m) \geq t | \boldsymbol{x}, \boldsymbol{y}\right) \approx 2 - \Phi\left[\xi\left(\min\left\{m, 2m_{max} - m\right\}\right)\right] - \Phi\left[\xi^{conj}\left(\min\left\{m, 2m_{max} - m\right\}\right)\right], \tag{3}$$

where Φ is the standard normal CDF, $m_{max} = \arg\max f(m)$, and

$$\xi(m) = \frac{t - g\left(\mu(m)\right)}{g'\left(\mu(m)\right)\sqrt{V(m)}}, \qquad\qquad \xi^{conj}(m) = \frac{t - g^{conj}\left(\mu(m)\right)}{(g^{conj})'\left(\mu(m)\right)\sqrt{V(m)}}.$$

Proof of Corollary 3. For $m = 1, \ldots, m_{\max}$,

$$\begin{aligned}\Pr(T(m) > t|\boldsymbol{x}, \boldsymbol{y}) &= \Pr\left(g(W(m)) > t\right) + \Pr\left(g^{\text{conj}}(W(m)) > t\right)\\ &= \Pr\left(Z > \frac{t - g\left(\mu(m)\right)}{g'\left(\mu(m)\right)\sqrt{V(m)}}\right) + \Pr\left(Z > \frac{t - g^{\text{conj}}\left(\mu(m)\right)}{(g^{\text{conj}})'\left(\mu(m)\right)\sqrt{V(m)}}\right) \qquad (4)\\ &\approx 1 - \Phi\left(\xi(m)\right) + 1 - \Phi\left(\xi^{\text{conj}}(m)\right) \qquad (5)\end{aligned}$$

where Z is a standard normal random variable and $\mu(m)$ and $V(m)$ are given in Corollary 1. Line (4) follows from the delta method, and line (5) follows from Corollary 2 for N sufficiently large.

Furthermore, since the partition-specific p-values are approximately symmetric about $m_{\max}$ (the p-values are exactly symmetric for equal sample sizes, and the symmetry worsens as the sample sizes become more imbalanced), we can get the asymptotic p-value for any partition $m = 1, \dots, \min(n_y, n_x)$ as

$$\Pr\left(T(m) \geq t | \boldsymbol{x}, \boldsymbol{y}\right) \approx 2 - \Phi\left[\xi\left(\min\left\{m, 2m_{\max} - m\right\}\right)\right] - \Phi\left[\xi^{\text{conj}}\left(\min\left\{m, 2m_{\max} - m\right\}\right)\right]. \tag{6}$$

This proves the corollary. □

We also note that when $n_x = n_y$, the approximation in (3) is equally accurate for partitions both smaller and larger than $m_{\max}$. However, for unequal sample size, the approximation is less accurate for partitions larger than $m_{\max}$.

In summary, and to be explicit with all quantities, for the statistic $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$, we have

$$\Pr\left(T(m) \geq t | \boldsymbol{x}, \boldsymbol{y}\right) \approx 2 - \Phi\left[\xi\left(\min\left\{m, 2m_{\max} - m\right\}\right)\right] - \Phi\left[\xi^{\text{conj}}\left(\min\left\{m, 2m_{\max} - m\right\}\right)\right]$$

where Φ is the standard normal CDF, $m_{\max} = \arg\max_m f(m)$, $f(m) = \binom{N}{n_{\min}}^{-1}\binom{n_x}{m}\binom{n_y}{m}$, $n_{\min} = \min(n_x, n_y)$, and [1]

$$\xi(m) = \frac{t - g\left(\mu(m)\right)}{g'\left(\mu(m)\right)\sqrt{V(m)}} \qquad \xi^{\text{conj}}(m) = \frac{t - g^{\text{conj}}\left(\mu(m)\right)}{(g^{\text{conj}})'\left(\mu(m)\right)\sqrt{V(m)}}$$

$$g(\mu(m)) = \frac{n_y}{n_x}\left(\frac{n_x\bar{x} + \mu(m)}{n_y\bar{y} - \mu(m)}\right) \qquad g^{\text{conj}}(\mu(m)) = \frac{n_x}{n_y}\left(\frac{n_y\bar{y} - \mu(m)}{n_x\bar{x} + \mu(m)}\right)$$

$$g'\left(\mu(m)\right) = \frac{n_y}{n_x}\left(\frac{n_y\bar{y} + n_x\bar{x}}{\left(n_y\bar{y} - \mu(m)\right)^2}\right) \qquad (g^{\text{conj}})'\left(\mu(m)\right) = -\frac{n_y}{n_x}\left(\frac{n_x\bar{x} + n_y\bar{y}}{\left(n_x\bar{x} + \mu(m)\right)^2}\right)$$

where

$$\mu(m) = m(\bar{y} - \bar{x})$$

$$V(m) = m\left[\frac{n_y - m}{n_y(n_y - 1)}\sum_{j=1}^{n_y}(y_j - \bar{y})^2 + \frac{n_x - m}{n_x(n_x - 1)}\sum_{i=1}^{n_x}(x_i - \bar{x})^2\right].$$

To get the expected trend shown Figure 1 of Section 3, we set $t = \bar{x}/\bar{y}$ (the observed test statistic), and substituted expected values for the sample quantities. For example, if we generated the elements of $\boldsymbol{x}$ as iid realizations of a random variable X, then we substituted $\mathbb{E}[X]$ for $\bar{x}$, and $\text{Var}(X)$ for $(n_x - 1)^{-1}\sum_{i=1}^{n_x}(x_i - \bar{x})^2$.

We note that we get similar results for $T = |\bar{x} - \bar{y}|$. In this case we can write $R(m) = \bar{x} - \bar{y}$

[1] Implementation note: In the fastPerm package, we use the same function to compute ξ and ξ^{conj}, reversing the order of the arguments related to x and y.

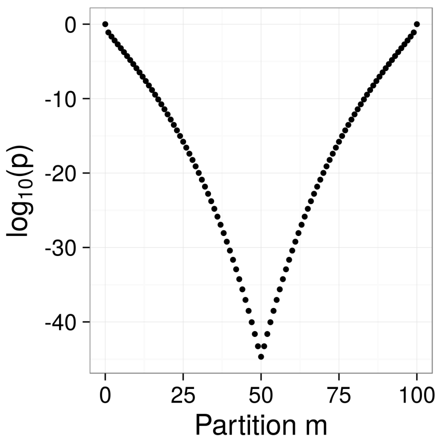


Figure S1: Trend in p-values across the partitions for $T = |\bar{x} - \bar{y}|$ with $n_x = n_y = 100$, $\mu_x = 4, \mu_y = 2$, and $\sigma_x^2 = \sigma_y^2 = 1$.

as

$$\begin{aligned} R(m) &= \frac{1}{n_x}[(\mathbf{1} - \boldsymbol{\delta}_x)'\boldsymbol{x} + \boldsymbol{\delta}_y'\boldsymbol{y}] - \frac{1}{n_y}[\boldsymbol{\delta}_x{}'\boldsymbol{x} + (\mathbf{1} - \boldsymbol{\delta}_y)'\boldsymbol{y}] \\ &= \bar{x} - \bar{y} + \left(\frac{1}{n_x} + \frac{1}{n_y}\right) W(m) \end{aligned}$$

Therefore, (3) still holds, but with $g(\mu(m)) = \bar{x} - \bar{y} + \left(n_x^{-1} + n_y^{-1}\right)\mu(m)$ and $g'(\mu(m)) = \left(n_x^{-1} + n_y^{-1}\right)$, with the corresponding results for g^{conj} and $(g^{\text{conj}})'$. All other formula are the same as those given for the ratio of means. The resulting trend for $T = |\bar{x} - \bar{y}|$ is shown in Figure S1 with $n_x = n_y = 100$, $\mu_x = 4, \mu_y = 2$, and $\sigma_x^2 = \sigma_y^2 = 1$.

While this appendix shows that the nearly log linear trend holds for both $T = |\bar{x} - \bar{y}|$ and $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$, we speculate that the trend might be similar for other statistics that are smooth functions of the means. The results for $R = \bar{x}/\bar{y}$ and $R = \bar{x} - \bar{y}$ above suggest a general formulation of permutation statistics in terms of W, which might help with this effort. This general formulation is presented in Proposition 1, in which R could be any statistic of the sample means, and not necessarily the ratio or difference of means.

Proposition 1. *Let $R(m) = R(\bar{x}^*(m), \bar{y}^*(m)|\boldsymbol{x}, \boldsymbol{y})$ be any statistic of the permuted sample means conditional on observed data $\boldsymbol{x}, \boldsymbol{y}$, where $\bar{x}^*(m)$ and $\bar{y}^*(m)$ are the means of a permuted dataset $(\boldsymbol{x}^{*\prime}, \boldsymbol{y}^{*\prime})'$ corresponding to a permutation $\pi \in \Pi(m)$. Then we can always write $R(m) = g(W(m))$ for some function g that is conditional on the observed data $\boldsymbol{x}, \boldsymbol{y}$.*

Proof of Proposition 1. Noting that $\bar{x}^*(m) = \bar{x} + (1/n_x)W(m)$ and $\bar{y}^*(m) = \bar{y} - (1/n_y)W(m)$, we have

$$\begin{aligned} R\left(\bar{x}^*(m), \bar{y}^*(m) | \boldsymbol{x}, \boldsymbol{y}\right) &= R\left(\bar{x} + (1/n_x)W(m), \bar{y} - (1/n_y)W(m) | \boldsymbol{x}, \boldsymbol{y}\right) \\ &= g\left(W(m)\right) \end{aligned}$$

where the last line follows, because $\bar{x}$, $\bar{y}$, n_x, and n_y are constant conditional on $\boldsymbol{x}$ and $\boldsymbol{y}$, and can be absorbed into the functional form of R. This proves the proposition. □

Then for any one-sided statistic $R = g(W)$, in order for asymptotic normality to hold within each partition for the corresponding two-sided statistic T, we must check the conditions in Theorem 1 and Corollary 3. However, it remains to be shown what additional properties are required to ensure a log concave trend in p-values across the partitions, so we must currently check new statistics on a case-by-base basis.

B Parametric p-values for ratios and differences of gamma random variables

The results in this appendix are used in our simulations of exponential and gamma random variables to obtain parametric approximations to the permutation p-value.

B.1 Ratio of means

Let F be the beta prime CDF (also called a Pearson type VI distribution (Johnson et al., 1995, p. 248)), and let f be the corresponding pdf. Following the form given by Becker and Klößner (2016), for $Z \sim F$,

$$f_Z(z; \alpha_1, \alpha_2, s, q) = \frac{\left(\frac{z-q}{s}\right)^{\alpha_1 - 1} \left(1 + \frac{z-q}{s}\right)^{-\alpha_1 - \alpha_2}}{sB(\alpha_1, \alpha_2)}$$

where B is the beta function. As we show in this section, if $X_i \overset{\text{iid}}{\sim} \text{Exp}(\lambda_x)$ and $Y_j \overset{\text{iid}}{\sim} \text{Exp}(\lambda_y)$, then $\bar{X}/\bar{Y}$ and $\bar{Y}/\bar{X}$ follow scaled beta prime distributions. This allows us to approximate the permutation p-value for the ratio statistic with the p-value from a beta prime. We note that the beta prime p-value is not conditional on the data, so is not the same as the permutation p-value, but simulation results suggest it is a reasonable approximation.

As in Section 5.2, let $x_i, i = 1, \ldots, n_x$, and $y_j, j = 1, \ldots, n_y$, be realizations of the respective random variables $X_i \overset{\text{iid}}{\sim} \text{Exp}(\lambda_x)$ and $Y_j \overset{\text{iid}}{\sim} \text{Exp}(\lambda_y)$. We consider the quantity $T = \max\left(\bar{X}/\bar{Y}, \bar{Y}/\bar{X}\right)$, and denote the observed statistic as $t = \max\left(\bar{x}/\bar{y}, \bar{y}/\bar{x}\right)$. Then under the null hypothesis that $\lambda_x = \lambda_y$, the p-value from the beta prime distribution is

$$\begin{aligned}
p_\beta &= \Pr(T \geq t) \\
&= \Pr\left(\max(\bar{X}/\bar{Y}, \bar{Y}/\bar{X}) \geq t\right) \\
&= \Pr\left(\left\{\bar{X}/\bar{Y} \geq t\right\} \cup \left\{\bar{Y}/\bar{X} \geq t\right\}\right) \\
&= \Pr\left(\bar{X}/\bar{Y} \geq t\right) + \Pr\left(\bar{Y}/\bar{X} \geq t\right) \qquad \text{(disjoint)} \qquad (7) \\
&= \Pr\left(\frac{n_y}{n_x}\frac{\sum_i X_i}{\sum_j Y_j} \geq t\right) + \Pr\left(\frac{n_x}{n_y}\frac{\sum_j Y_j}{\sum_i X_i} \geq t\right) \qquad (8) \\
&= 1 - F\left(t; \alpha_1 = n_x, \alpha_2 = n_y, s = n_y/n_x, q = 0\right) \qquad (9) \\
&+ 1 - F\left(t; \alpha_1 = n_y, \alpha_2 = n_x, s = n_x/n_y, q = 0\right).
\end{aligned}$$

The equality in (7) follows because $\bar{X}/\bar{Y} \geq t$ if and only if $\bar{Y}/\bar{X} < t$ (assuming $t \neq 1$, which occurs with probability one). Line 9 follows from well known properties, which we outline below.

Let $U_1 \sim \text{Gamma}(\alpha_1, \lambda_1)$ and $U_2 \sim \text{Gamma}(\alpha_2, \lambda_2)$, $U_1 \perp U_2$. Also, let $V_1 = h_1(U_1, U_2) = U_1/U_2$ and $V_2 = h_2(U_1, U_2) = U_2$ with respective inverse transformations $U_1 = h^{-1}(V_1, V_2) = V_1V_2$ and $U_2 = h^{-1}(V_1, V_2) = V_2$. Noting that the Jacobian of the transformation is

$$J = \begin{vmatrix} \partial u_1/\partial v_1 & \partial u_1/\partial v_2 \\ \partial u_2/\partial v_1 & \partial u_2/\partial v_2 \end{vmatrix} = \begin{vmatrix} v_2 & v_1 \\ 0 & 1 \end{vmatrix} = v_2,$$

we have

$$\begin{aligned}
f_{V_1,V_2}(v_1, v_2) &= f_{U_1,U_2}\left(h_1^{-1}(v_1, v_2), h_2^{-1}(v_1, v_2)\right)|J| \\
&= \frac{\lambda_1^{\alpha_1}}{\Gamma(\alpha_1)}(v_1v_2)^{\alpha_1 - 1}e^{-\lambda_1 v_1 v_2}\frac{\lambda_2^{\alpha_2}}{\Gamma(\alpha_2)}v_2^{\alpha_2 - 1}e^{-\lambda_2 v_2}v_2 \\
&= \frac{\lambda_1^{\alpha_1}\lambda_2^{\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}v_1^{\alpha_1 - 1}v_2^{\alpha_1 + \alpha_2 - 1}e^{-(\lambda_1 v_1 + \lambda_2)v_2}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
f_{V_1}(v_1) &= \int_0^\infty f_{V_1,V_2}(v_1,v_2)dv_2 \\
&= \frac{\lambda_1^{\alpha_1}\lambda_2^{\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} v_1^{\alpha_1-1} \int_0^\infty v_2^{\alpha_1+\alpha_2-1} e^{-(\lambda_1 v_1+\lambda_2)v_2} dv_2 \\
&= \frac{\lambda_1^{\alpha_1}\lambda_2^{\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} v_1^{\alpha_1-1} \frac{\Gamma(\alpha_1+\alpha_2)}{(\lambda_1 v_1+\lambda_2)^{\alpha_1+\alpha_2}} \\
&= \frac{\left(\frac{v_1}{\lambda_2/\lambda_1}\right)^{\alpha_1-1}\left(1+\frac{v_1}{\lambda_2/\lambda_1}\right)^{-\alpha_1-\alpha_2}}{(\lambda_2/\lambda_1)B(\alpha_1,\alpha_2)},
\end{aligned}$$

which is a generalized beta prime distribution with shape parameters α_1 and α_2, location parameter $q = 0$, and scale parameter $s = \lambda_2/\lambda_1$. In the case where $\lambda_1 = \lambda_2$, this simplifies to the standard beta prime distribution with shape parameters α_1 and α_2. This shows that whenever $U_1 \sim \text{Gamma}(\alpha_1, \lambda)$, $U_2 \sim \text{Gamma}(\alpha_2, \lambda)$, and $U_1 \perp U_2$, we have $U_1/U_2 \sim F(\alpha_1, \alpha_2, 1, 0)$. We note that some sources report that for $U_1 \sim \text{Gamma}(\alpha_1, \lambda_1)$, $U_2 \sim \text{Gamma}(\alpha_2, \lambda_2)$, and $U_1 \perp U_2$, we have $U_1/U_2 \sim F(\alpha_1, \alpha_2, 1, 0)$ if $\lambda_1 = \lambda_2 = 1$ (e.g., Leemis and McQueston, 2008). However, as shown above, this also holds when $\lambda_1 = \lambda_2 \neq 1$.

Now let $Z = \left(\sum_{i=1}^{n_x} X_i\right) / \left(\sum_{j=1}^{n_y} Y_i\right)$. Since $X_i \overset{\text{iid}}{\sim} \text{Exp}(\lambda_x)$ and $Y_j \overset{\text{iid}}{\sim} \text{Exp}(\lambda_y)$, it follows that $\sum_{i=1}^{n_x} X_i \sim \text{Gamma}(n_x, \lambda_x)$ and $\sum_{j=1}^{n_y} Y_j \sim \text{Gamma}(n_y, \lambda_y)$. Then under the null of $\lambda_x = \lambda_y$, the results above give $Z \sim F(n_x, n_y, 1, 0)$ and $1/Z \sim F(n_y, n_x, 1, 0)$.

Now let $W = sZ$. Then by a change of variable, we have

$$f_W(w) = \frac{\left(\frac{w}{s}\right)^{n_x-1}\left(1+\frac{w}{s}\right)^{-n_x-n_y}}{sB(n_x,n_y)}$$

Applying this result to (8), we have

$$\frac{n_y}{n_x}\frac{\sum_{i=1}^{n_x} X_i}{\sum_{j=1}^{n_y} Y_j} \sim F(\cdot; n_x, n_y, n_y/n_x, 0)$$

and similarly,

$$\frac{n_x}{n_y}\frac{\sum_{j=1}^{n_y} Y_j}{\sum_{i=1}^{n_x} X_i} \sim F(\cdot; n_y, n_x, n_x/n_y, 0)$$

Then (9) follows directly from (8).

To compute the CDF values for the scaled beta prime, we used the `PearsonDS` package for R (Becker and Klößner, 2016).

Similarly, for gamma random variables $X_i \overset{\text{iid}}{\sim} \text{Gamma}(\alpha_x, \lambda_x)$ and $Y_j \overset{\text{iid}}{\sim} \text{Gamma}(\alpha_y, \lambda_y)$,

$\sum_{i=1}^{n_x} X_i \sim \text{Gamma}(n_x\alpha_x, \lambda_x)$ and $\sum_{j=1}^{n_y} Y_j \sim \text{Gamma}(n_y\alpha_y, \lambda_y)$. Then letting $Z = \left(\sum_{i=1}^{n_x} X_i\right) / \left(\sum_{j=1}^{n_y} Y_j\right)$, under the null of $H_0 : \lambda_x = \lambda_y, \alpha_x = \alpha_y = \alpha$, we have $Z \sim F(\cdot; n_x\alpha, n_y\alpha, 1, 0)$ and $1/Z \sim F(\cdot; n_y\alpha, n_x\alpha, 1, 0)$, so $(n_y/n_x)Z \sim F(\cdot; n_x\alpha, n_y\alpha, n_y/n_x, 0)$ and $(n_x/n_y)Z \sim F(\cdot; n_y\alpha, n_x\alpha, n_x/n_y, 0)$. Therefore,

$$\begin{aligned} p_\beta = \Pr(T \geq t) &= 1 - F\left(t; n_x\alpha, n_y\alpha, n_y/n_x, 0\right) \\ &\quad + 1 - F\left(t; n_y\alpha, n_x\alpha, n_x/n_y, 0\right). \end{aligned}$$

In our simulations, we generated data under the alternative $H_1 : \lambda_x \neq \lambda_y, \alpha_x = \alpha_y = \alpha$ for various values of α. While we would ideally also simulate under the alternatives $H_1 : \lambda_x \neq \lambda_y, \alpha_x \neq \alpha_y$ and $H_1 : \lambda_x = \lambda_y, \alpha_x \neq \alpha_y$, in these scenarios it is not possible to compute p_β under $H_0 : \alpha_x = \alpha_y, \lambda_x = \lambda_y$, because α does not disappear in the beta prime density. Consequently, we would have to compute p_β under $H_0 : \alpha_x = \alpha_y = c, \lambda_x = \lambda_y$ for a specified constant c. This is more restrictive than the null hypothesis for the permutation test, and consequently, it would not be clear how to compute the parametric p-value to use as an approximation for the true permutation p-value.

B.2 Difference in means

Let $M_X(t)$ be the moment generating function (MGF) for random variable X. Then for $X_i \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \lambda), i = 1, \ldots, n$, $M_{\frac{1}{n}\sum_{i=1}^n X_i}(t) = M_{\sum_{i=1}^n X_i}(t/n) = \prod_{i=1}^n M_{X_i}(t/n) = \left(1 - \frac{1}{n\lambda}t\right)^{-n\alpha}$, which is the MGF for a Gamma distribution with shape parameter $n\alpha$ and rate parameter $n\lambda$. Therefore, $\bar{X} \sim \text{Gamma}(n\alpha, n\lambda)$.

Then for $X_i \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \lambda), i = 1, \ldots, n_x$ and $Y_j \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \lambda), j = 1, \ldots, n_y$, the distribution of $\bar{X} - \bar{Y}$, which we denote as G, is (Klar, 2015)

$$G(z) = \Pr(\bar{X} - \bar{Y} \leq z) = C \underbrace{\int_{\max\{0,-z\}}^{\infty} v^{n_y\alpha - 1} e^{-n_y\lambda v} \gamma\left(n_x\alpha, n_x\lambda(v + z)\right) dv}_{A(z)}, \tag{10}$$

where $\gamma(a, b) = \int_0^b s^{a-1} e^{-s} ds$ is the lower incomplete gamma function, and $C = (n_y\lambda)^{n_y\alpha} / \left(\Gamma(n_x\alpha)\Gamma(n_y\alpha)\right)$ is the normalizing constant. Klar (2015) also gives the density for $\bar{X} - \bar{Y}$, which was derived by Mathai (1993).

However, in our simulations we found that several scenarios led to numerical problems in computing (10) due to large gamma and incomplete gamma function values. These were not solved by computing $G(z) = \exp\{n_y\alpha \log(n_y\lambda) - \log\Gamma(n_x\alpha) - \log\Gamma(n_y\alpha) + \log(A(z))\}$ where $\log\Gamma$ is the log gamma function. As an alternative, we used a saddlepoint approximation for (10). As described below, the saddlepoint approximation is accurate and did not pose computational difficulties.

To compute the saddlepoint approximation, note that under $H_0 : \lambda_x = \lambda_y = \lambda, \alpha_x = \alpha_y = \alpha$, the MGF of $\bar{X} - \bar{Y}$ is

$$M_{\bar{X}-\bar{Y}}(t) = \left(1 - \frac{1}{n_x\lambda}t\right)^{-n_x\alpha} \left(1 + \frac{1}{n_y\lambda}t\right)^{-n_y\alpha} \quad t \in (-n_y\lambda, n_x\lambda),$$

and the cumulant generating function is

$$K(t) = \log\left(M_{\bar{X}-\bar{Y}}(t)\right) = -n_x\alpha \log\left(1 - \frac{t}{n_x\lambda}\right) - n_y\alpha \log\left(1 + \frac{t}{n_y\lambda}\right).$$

After some algebra, we get the derivatives

$$\begin{aligned} K'(t) &= \frac{\alpha(n_x + n_y)t}{(n_x\lambda - t)(n_y\lambda + t)} \\ K''(t) &= \alpha(n_x + n_y)\frac{t^2 + n_x n_y \lambda^2}{[(n_x\lambda - t)(n_y\lambda + t)]^2}. \end{aligned}$$

Let $\hat{t} = \hat{t}(z) \in (-n_y\lambda, n_x\lambda)$ be the solution to $K'(\hat{t}) = z$. Then as Butler (2007) describes, the saddlepoint approximation of the cumulative distribution for $z \neq \mathbb{E}[\bar{X} - \bar{Y}] = 0$ is (Lugannani and Rice, 1980)

$$\hat{G}(z) = \Phi(\hat{w}) + \phi(\hat{w})\left(\frac{1}{\hat{w}} - \frac{1}{\hat{u}}\right), \tag{11}$$

where $\hat{w} = \operatorname{sgn}(\hat{t})\sqrt{2\left[\hat{t}z - K(\hat{t})\right]}$, $\hat{u} = \hat{t}\sqrt{K''(\hat{t})}$, and Φ and ϕ are the standard normal distribution and density, respectively. The two-sided p-value is then $p_{\text{saddle}} = \Pr(T \geq t) = 1 - \hat{G}(t; n_x, n_y, \lambda, \alpha) + \hat{G}(-t; n_x, n_y, \lambda, \alpha)$.

Figure S2 compares the true distribution (10) and saddlepoint approximation (11) for $n_x = n_y = 100$, $\alpha = 1$, and $\lambda = 4$. Figure S2 shows agreement between the true distribution and saddlepoint approximation far into the tail. The trend is similar for other parameter values (not shown), and appears to be reliable up to quantile values of around 10^{-200}. We also note that through simulations, we found that both the true distribution and the saddlepoint approximation agreed with the empirical distribution for a variety of parameter values.

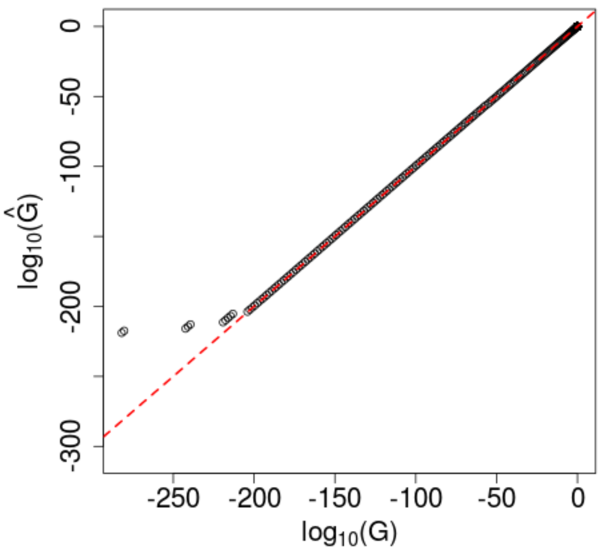


Figure S2: Comparison of true (G) and saddlepoint approximation ($\hat{G}$) distributions of the difference of gamma random variables. The diagonal dashed line has slope of 1 and an intercept of 0, and indicates agreement. This figure appears in color in the electronic version of this article.

Both the true distribution (10) and saddlepoint approximation (11) are functions of α and λ. Neither parameter disappears under the null of $H_0 : \alpha_x = \alpha_y = \alpha, \lambda_x = \lambda_y = \lambda$, so we must set α and λ to fixed values to compute p-values. To do this in the simulations, we pooled the generated data, computed the maximum likelihood estimates (MLEs), and plugged the MLEs into (11). In the simulations, we found that allowing both α and λ to vary led to less reliable p-values from the saddlepoint approximation than allowing just one parameter to vary. To be consistent with our simulations for the ratio of gamma means, we fixed α and used the MLE estimate for λ in the simulations.

We note that this procedure for obtaining a parametric approximation to the permutation p-value involves three approximations: 1) approximating the permutation p-value (conditional on the data) with a parametric distribution (not conditional on the data), 2) approximating the parametric distribution with a saddlepoint approximation, and 3) approximating the general null $H_0 : \lambda_x = \lambda_y$ with the more restrictive null $H_0 : \lambda_x = \lambda_y = \hat{\lambda}$, where $\hat{\lambda}$ is the MLE from the pooled data.

To obtain the MLE estimates, let $\boldsymbol{z} = (\boldsymbol{x}', \boldsymbol{y}')'$ be the pooled data, $N = n_x + n_y$ be the total sample size, and $\bar{z} = N^{-1} \sum_{i=1}^{N} z_i$, $s^2 = (N-1)^{-1} \sum_i (z_i - \bar{z})^2$ be the sample mean and variance, respectively. Then assuming iid observations, the joint log likelihood is

$$\ell = N\alpha \log(\lambda) - N \log\left(\Gamma(\alpha)\right) + (\alpha - 1)\sum_i \log(z_i) - N\lambda\bar{z}.$$

Taking the derivative with respect to λ and setting to zero, we get $\lambda = \alpha/\bar{z}$. Then taking $\partial\ell/\partial\alpha$ and substituting in $\lambda = \alpha/\bar{z}$, we get

$$\begin{aligned}\ell'(\alpha) &= N\log\left(\frac{\alpha}{\bar{z}}\right) - N\Psi(\alpha) + \sum_i \log(x_i)\\ \ell''(\alpha) &= \frac{N}{\alpha} - N\Psi'(\alpha),\end{aligned}$$

where $\Psi(\alpha) = d\log(\Gamma(\alpha))/d\alpha$ is the digamma function, and $\Psi'(\alpha) = d\Psi(\alpha)/d\alpha$ is the trigamma function. We used Newton-Raphson until convergence of $\ell(\alpha)$ to get the MLE $\hat{\alpha}$, where each update is given by $\alpha^{k+1} = \alpha^k - \ell'\left(\alpha^k\right)/\ell''\left(\alpha^k\right)$, and then set $\hat{\lambda} = \hat{\alpha}\bar{z}$. To get initial values for α, we used the method of moments and set $\alpha^0 = \bar{z}^2/s^2$.

C Additional simulations

In this section, we present simulation results under additional scenarios.

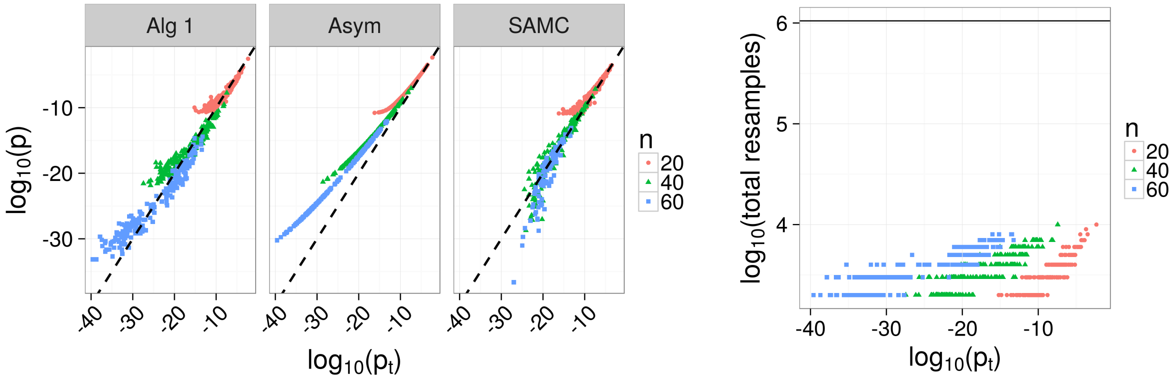
C.1 Difference in means with normal data

In this subsection, we use the statistic $T = |\bar{x} - \bar{y}|$ with data generated as normal random variables.

C.1.1 Small sample sizes

We generated data $x_i, i = 1, \ldots, n_x$ and $y_j, j = 1, \ldots, n_y$ as realizations of the respective random variables $X_i \overset{\text{iid}}{\sim} N(\mu_x, 1)$ and $Y_j \overset{\text{iid}}{\sim} N(\mu_y, 1)$. For equal sample sizes, we set $n = n_x = n_y = 20, 40, 60$, and for unequal sample sizes we set $n_x = 20, 40, 60$ and $n_y = 100$. For both equal and unequal sample sizes and for each each n or n_x, we set $\mu_x = 2$ or 3, and $\mu_y = 0$, and simulated 100 datasets for each combination of parameters. We used the p-value from a t-test with equal variance, denoted as p_t, as an approximation for the true permutation p-value.

Results for equal and unequal sample size are shown in Figures S3 and S4, respectively. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *SAMC* is the SAMC algorithm, and p_t is a two-sided t-test with



(a) p-values

(b) Number of resamples in Alg 1

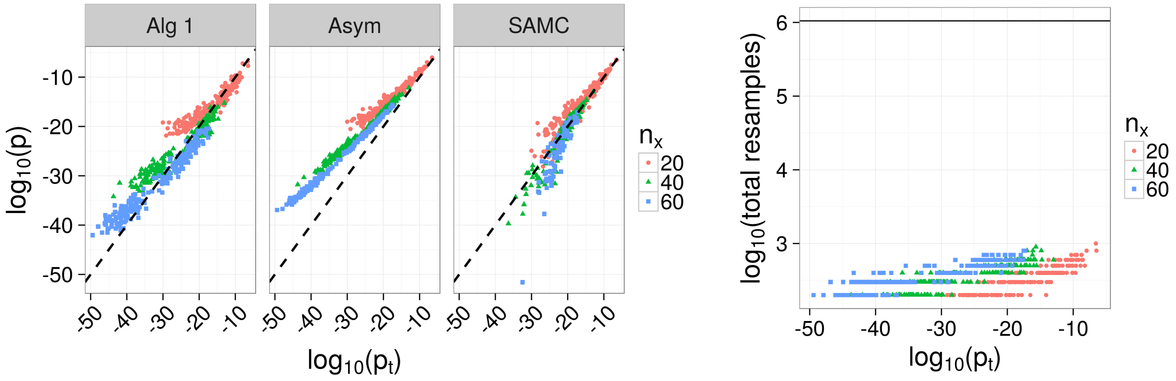
Figure S3: Simulation results using the statistic $T = |\bar{x} - \bar{y}|$ with normal data, $\mu_x = 2$ or 3, $\mu_y = 0$, and equal sample sizes of $n = n_x = n_y = 20$, 40, 60. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *SAMC* is the SAMC algorithm, and p_t is a two-sided t-test with equal variance. The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods. The horizontal line in S3b shows the number of iterations used in the SAMC algorithm (set in advance and independent of p-value). This figure appears in color in the electronic version of this article.

equal variance. The number of resamples used by our algorithm is shown in Figures S3b and S4b. We note that the bias shown in Figures S3a and S4a are similar to that obtained with moment-corrected correlation (MCC) (Zhou and Wright, 2015), shown in Figure S20 of Web Appendix D.

C.1.2 Under the null hypothesis $P_x = P_y$

We generated data $x_i, i = 1, \ldots, n_x$ and $y_j, j = 1, \ldots, n_y$ as realizations of the respective random variables $X_i \overset{\text{iid}}{\sim} N(0, 1)$ and $Y_j \overset{\text{iid}}{\sim} N(0, 1)$. For equal sample sizes, we set $n = n_x = n_y = 20, 40, 60$, and for unequal sample sizes we set $n_x = 20, 40, 60$ and $n_y = 100$. For both equal and unequal sample sizes and for each each n or n_x, we simulated 1,000 datasets (we used 1,000 datasets instead of 100 to better investigate the type I error rate). We used the p-value from simple Monte Carlo resampling with 10^5 resamples, denoted as $\tilde{p}$, as an approximation for the true permutation p-value.

Results for equal and unequal sample size are shown in Figures S5 and S6, respectively. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *t-test* shows the p-value from a two-sided t-test with equal variance, and $\tilde{p}$ is from simple Monte Carlo resampling with 10^5 resamples. We compare



(a) p-values

(b) Number of resamples in Alg 1

Figure S4: Simulation results using the statistic $T = |\bar{x} - \bar{y}|$ with normal data, $\mu_x = 2$ or 3, $\mu_y = 0$, and unequal sample sizes, where $n_y = 100$ and $n_x = 20, 40, 60$. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *SAMC* is the SAMC algorithm, and p_t is a two-sided t-test with equal variance. The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods. The horizontal line in S4b shows the number of iterations used in the SAMC algorithm (set in advance and independent of p-value). This figure appears in color in the electronic version of this article.

p-values from the t-test against $\tilde{p}$, which shows close agreement. We do not show results from the SAMC algorithm, because the EXPERT package (Yu et al., 2011) does not provide results for p-values $> 10^{-3}$.

Tables S1 and S2 show the Type I error rates under the null $H_0 : P_x = P_y$ for the equal and unequal sample size simulations, respectively. *MC* is the unadjusted p-value from simple Monte Carlo resampling with 10^5 resamples, *t-test* is a two-sided t-test with equal variance, *Alg 1* is our resampling algorithm, and *Asymptotic* is our asymptotic approximation.

C.2 Ratio of means with exponential data

In this subsection, we use the statistic $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ with data generated as exponential random variables.

C.2.1 Small sample sizes

We generated data $x_i, i = 1, \ldots, n_x$ and $y_j, j = 1, \ldots, n_y$ as realizations of the respective random variables $X_i \overset{\text{iid}}{\sim} \text{Exp}(\lambda_x)$ and $Y_j \overset{\text{iid}}{\sim} \text{Exp}(\lambda_y)$. For equal sample sizes, we set $n = n_x = n_y = 20, 40, 60$, and for unequal sample sizes, we set $n_x = 20, 40, 60$ and $n_y = 100$. For each n or n_x, we set $\lambda_y = 5$ or 10, and $\lambda_x = 1$. For both equal and unequal sample sizes, we

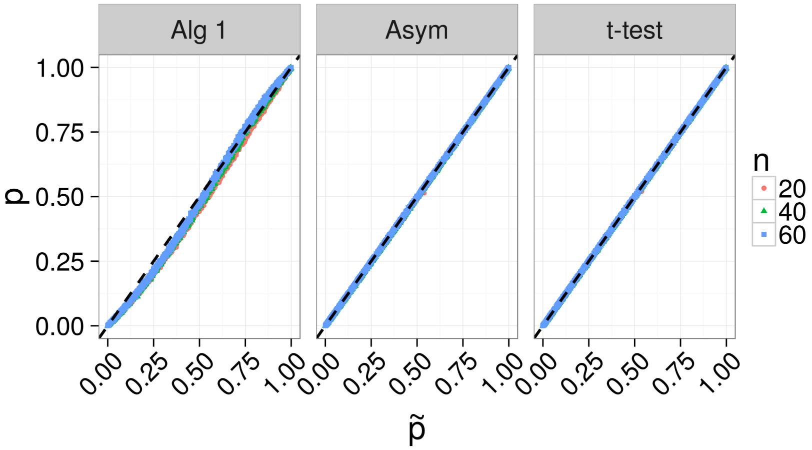


Figure S5: Simulation results using the statistic $T = |\bar{x} - \bar{y}|$ with normal data under the null $P_x = P_y$ with equal sample sizes of $n = n_x = n_y = 20, 40, 60$. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *t-test* shows the p-value from a two-sided t-test with equal variance, and $\tilde{p}$ is from simple Monte Carlo resampling with 10^5 resamples. The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods. This figure appears in color in the electronic version of this article.

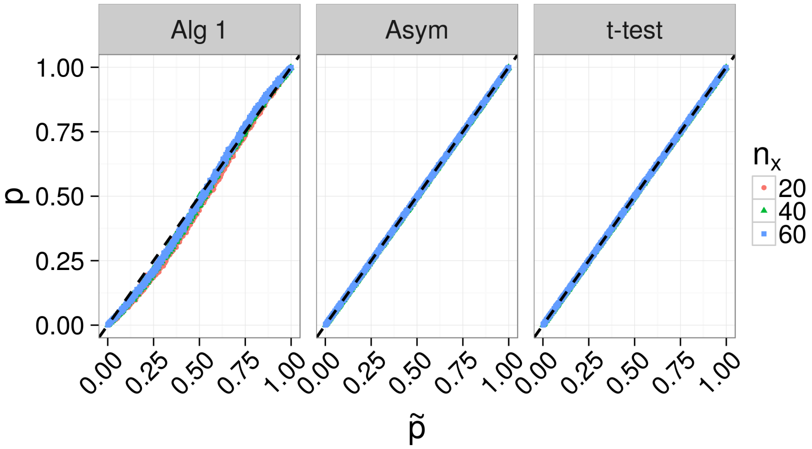


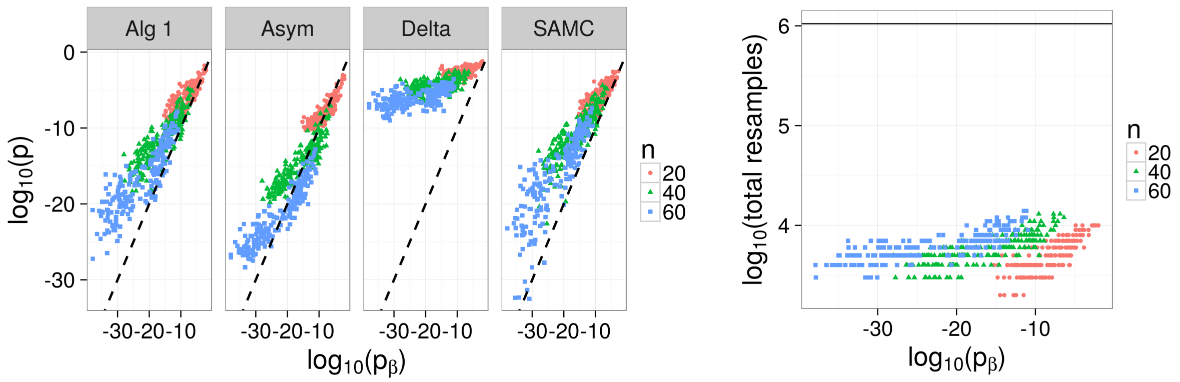
Figure S6: Simulation results using the statistic $T = |\bar{x} - \bar{y}|$ with normal data under the null $P_x = P_y$ with unequal sample sizes of $n_x = 20, 40, 60$ and $n_y = 100$. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *t-test* shows the p-value from a two-sided t-test with equal variance, and $\tilde{p}$ is from simple Monte Carlo resampling with 10^5 resamples. The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods. This figure appears in color in the electronic version of this article.

Table S1: Type I error rates Pr(p-value $\leq$ signif level$|H_0$) for $T = |\bar{x} - \bar{y}|$ with normal data and equal sample sizes $n = n_x = n_y$. *MC* is the unadjusted p-value from simple Monte Carlo resampling with 10^5 resamples, *t-test* is a two-sided t-test with equal variance, *Alg 1* is our resampling algorithm, and *Asymptotic* is our asymptotic approximation.

signif level	n	MC	t-test	Alg 1	Asymptotic
0.01	20	0.010	0.010	0.015	0.010
	40	0.013	0.013	0.015	0.013
	60	0.010	0.010	0.011	0.010
0.05	20	0.048	0.050	0.064	0.050
	40	0.055	0.055	0.075	0.056
	60	0.049	0.050	0.061	0.050
0.1	20	0.098	0.098	0.14	0.11
	40	0.11	0.11	0.14	0.11
	60	0.10	0.10	0.12	0.10

Table S2: Type I error rates Pr(p-value $\leq$ signif level$|H_0$) for $T = |\bar{x} - \bar{y}|$ with normal data and unequal sample sizes $n_x \neq n_y$ (n_x shown and $n_y = 100$). *MC* is the unadjusted p-value from simple Monte Carlo resampling with 10^5 resamples, *t-test* is a two-sided t-test with equal variance, *Alg 1* is our resampling algorithm, and *Asymptotic* is our asymptotic approximation.

signif level	n_x	MC	t-test	Alg 1	Asymptotic
0.01	20	0.013	0.013	0.018	0.013
	40	0.016	0.016	0.018	0.016
	60	0.010	0.010	0.013	0.010
0.05	20	0.049	0.049	0.075	0.049
	40	0.047	0.047	0.066	0.047
	60	0.044	0.044	0.057	0.044
0.1	20	0.090	0.090	0.14	0.092
	40	0.10	0.10	0.14	0.11
	60	0.090	0.090	0.13	0.090



(a) p-values

(b) Number of resamples in Alg 1

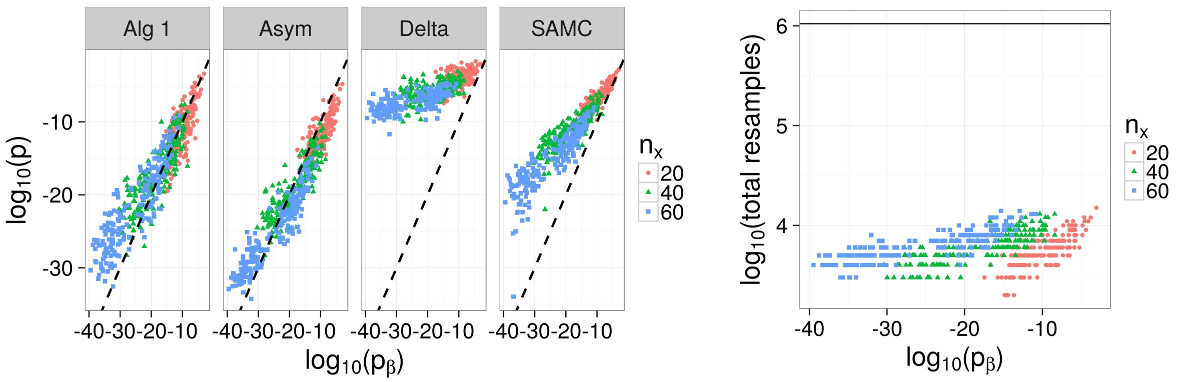
Figure S7: Simulation results using the statistic $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ with exponential data, $n = n_x = n_y = 20$, 40, 60, and rates $\lambda_y = 5, 10$ and $\lambda_x = 1$. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *Delta* is the delta method, *SAMC* is the SAMC algorithm, and p_β is the two-sided p-value from the beta prime distribution. The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods. The horizontal line in S7b shows the number of iterations used in the SAMC algorithm (set in advance, and independent of p-value). The SAMC algorithm did not produce values for 15 tests (points missing). This figure appears in color in the electronic version of this article.

simulated 100 datasets for each combination of parameters. We used the p-value from the beta prime distribution, denoted as p_β (see Web Appendix B), as an approximation to the true permutation p-value.

Results for equal and unequal sample size are shown in Figures S7 and S8, respectively. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *Delta* is the delta method, *SAMC* is the SAMC algorithm, and p_β is the two-sided p-value from the beta prime distribution. The number of resamples used by our resampling algorithm is shown in Figures S7b and S8b.

C.2.2 Under the null hypothesis $P_x = P_y$

We generated data $x_i, i = 1, \ldots, n_x$ and $y_j, j = 1, \ldots, n_y$ as realizations of the respective random variables $X_i \overset{\text{iid}}{\sim} \text{Exp}(1)$ and $Y_j \overset{\text{iid}}{\sim} \text{Exp}(1)$. For equal sample sizes, we set $n = n_x = n_y = 20, 40, 60$. For unequal sample sizes, we set $n_x = 20, 40, 60$ and $n_y = 100$. For both equal and unequal sample sizes, we simulated 1,000 datasets for each combination of parameters (we used 1,000 datasets instead of 100 to better investigate the type I error rate). We used the p-value from simple Monte Carlo resampling with 10^5 resamples, denoted as $\tilde{p}$,



(a) p-values

(b) Number of resamples in Alg 1

Figure S8: Simulation results using the statistic $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ with exponential data, $n_x = 20$, 40, 60, $n_y = 100$, and rates $\lambda_y = 5, 10$ and $\lambda_x = 1$. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *Delta* is the delta method, *SAMC* is the SAMC algorithm, and p_β is the two-sided p-value from the beta prime distribution. The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods. The horizontal line in S8b shows the number of iterations used in the SAMC algorithm (set in advance, and independent of p-value). This figure appears in color in the electronic version of this article.

as an approximation for the true permutation p-value.

Results for equal and unequal sample size are shown in Figures S9 and S10, respectively. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *Delta* is the delta method, *Beta prime* gives the p-value from the beta prime distribution, and $\tilde{p}$ is from simple Monte Carlo resampling with 10^5 resamples. Given the large p-values, using 10^5 Monte Carlo resamples should be sufficient to obtain reliable estimates of the true permutation p-value. Therefore, this comparison demonstrates that the permutation p-value is not exactly the same as the p-value from the beta prime distribution. However, it appears reasonably close, so we use it as an approximation to the truth in other simulations in which the p-values are much smaller and simple Monte Carlo methods are not feasible.

We do not show results from the SAMC algorithm, because as noted above, the `EXPERT` package (Yu et al., 2011) does not provide results for p-values $> 10^{-3}$.

Tables S3 and S4 show the Type I error rates under the null $H_0 : P_x = P_y$ for the equal and unequal sample size simulations, respectively. *MC* is the unadjusted p-value from simple Monte Carlo resampling and 10^5 resamples, *Beta prime* is the p-value from the beta prime distribution, *Alg 1* is our resampling algorithm, and *Asymptotic* is our asymptotic approximation.

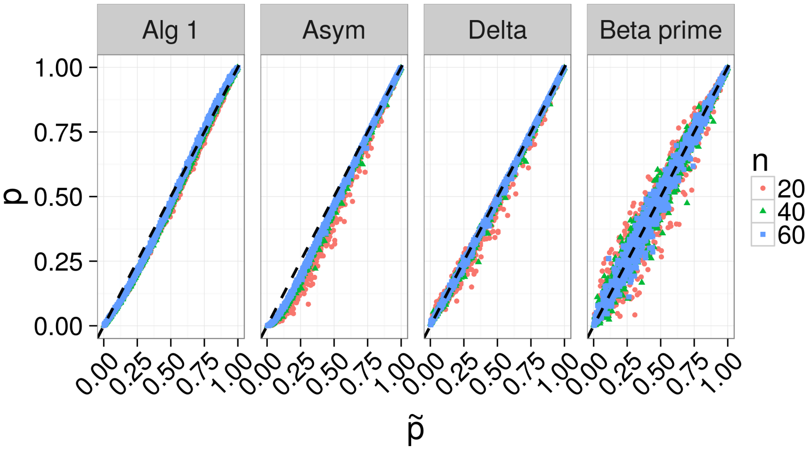


Figure S9: Simulation results using the statistic $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ with exponential data under the null of $P_x = P_y$ with equal sample sizes of $n = n_x = n_y =$ 20, 40, 60. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *Delta* is the delta method, *Beta prime* gives the p-value from the beta prime distribution, and $\tilde{p}$ is from simple Monte Carlo resampling with 10^5 resamples. The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods. This figure appears in color in the electronic version of this article.

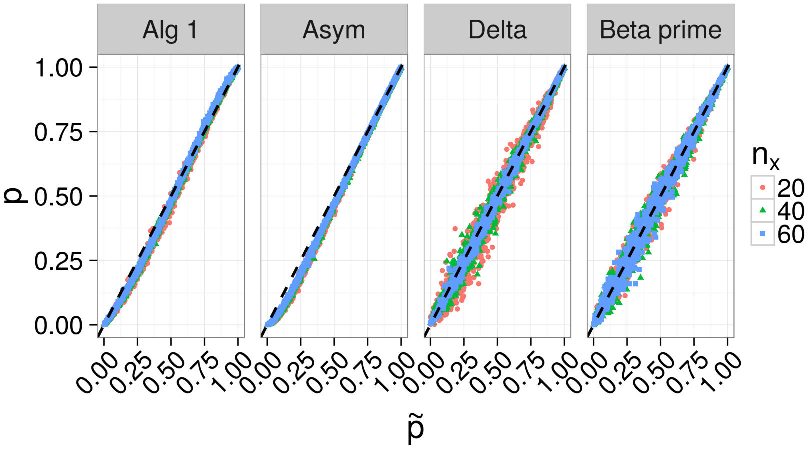


Figure S10: Simulation results using the statistic $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ with exponential data under the null of $P_x = P_y$ with unequal sample sizes of $n_x = 20$, 40, 60 and $n_y = 100$. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *Delta* is the delta method, *Beta prime* gives the p-value from the beta prime distribution, and $\tilde{p}$ is from simple Monte Carlo resampling with 10^5 resamples. The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods. This figure appears in color in the electronic version of this article.

Table S3: Type I error rates $\Pr(\text{p-value} \leq \text{signif level}|H_0)$ for $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ with exponential data and equal sample sizes $n = n_x = n_y$. *MC* is simple Monte Carlo resampling with 10^5 resamples, *Alg 1* is our resampling algorithm, *Asymptotic* is our asymptotic approximation, *Delta* is the delta method, and *Beta prime* is the the beta prime distribution.

signif level	n	MC	Alg 1	Asymptotic	Delta	Beta prime
0.01	20	0.010	0.016	0.066	0.003	0.009
	40	0.010	0.018	0.050	0.002	0.008
	60	0.013	0.013	0.031	0.006	0.015
0.05	20	0.064	0.084	0.14	0.045	0.058
	40	0.061	0.079	0.11	0.054	0.061
	60	0.051	0.063	0.091	0.050	0.047
0.10	20	0.11	0.15	0.21	0.12	0.11
	40	0.11	0.14	0.17	0.11	0.11
	60	0.093	0.11	0.14	0.095	0.092

Table S4: Type I error rates $\Pr(\text{p-value} \leq \text{signif level}|H_0)$ for $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ with exponential data and unequal sample sizes $n_x \neq n_y$ (n_x shown, and $n_y = 100$). *MC* is simple Monte Carlo resampling with 10^5 resamples, *Alg 1* is our resampling algorithm, *Asymptotic* is our asymptotic approximation, *Delta* is the delta method with, and *Beta prime* is the beta prime distribution.

signif level	n	MC	Alg 1	Asymptotic	Delta	Beta prime
0.01	20	0.011	0.016	0.054	0.008	0.012
	40	0.008	0.012	0.033	0.004	0.006
	60	0.012	0.016	0.035	0.007	0.014
0.05	20	0.061	0.082	0.127	0.065	0.056
	40	0.048	0.062	0.097	0.047	0.050
	60	0.047	0.065	0.083	0.044	0.051
0.10	20	0.12	0.16	0.19	0.14	0.12
	40	0.10	0.14	0.17	0.11	0.10
	60	0.091	0.12	0.14	0.093	0.088

C.3 Difference in means with gamma data

In this subsection, we use the statistic $T = |\bar{x} - \bar{y}|$ with data generated as gamma random variables.

C.3.1 Small sample sizes

We generated data $x_i, i = 1, \ldots, n_x$ and $y_j, j = 1, \ldots, n_y$ as realizations of the respective random variables $X_i \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \lambda_x)$ and $Y_j \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \lambda_y)$, where $\alpha = 0.5, 3, 5$, $\lambda_x = 1$, and λ is the rate parameter. For equal sample sizes, we set $n = n_x = n_y = 20, 40, 60$, and for unequal sample sizes we set $n_x = 20, 40, 60$ and $n_y = 100$. For $\alpha = 0.5$, we set $\lambda_y = 2.5, 3$ for all n or n_x. For $\alpha = 3$, we set $\lambda_y = 1.5, 1.75$ for all n or n_x. For $\alpha = 5$, we set $\lambda_y = 1.25, 1.5$ for all n or n_x. For both equal and unequal sample sizes, we simulated 100 datasets for each combination of parameters.

Results for equal and unequal sample size are shown in Figures S11 and S12, respectively. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *t-test* is a t-test with unequal variance, and *Saddle* is the saddlepoint approximation (see Web Appendix B). SAMC results are not shown, as the `EXPERT` package does not provide p-values larger than 10^{-3}. We use the p-values from simple Monte Carlo resampling, denoted as $\tilde{p}$, with 10^5 resamples as a basis of comparison, and only show values for which $\tilde{p} > 10^{-3}$ to ensure that the $\tilde{p}$ are reliable (1,023 values shown in Figure S11, and 573 values shown in Figure S12).

We use a t-test with unequal variance because we anticipate that this is the test that would be used in practice, though we note that it tests a more general null hypothesis ($H_0 : \mu_x = \mu_y$) than the permutation test ($H_0 : P_x = P_y$). This puts our methods at a disadvantage.

Overall, Figures S11 and S12 suggest that our methods work well in this setting, though our resampling algorithm might be liberal for equal sample sizes and $\alpha = 0.5$. The t-test performs well in some scenarios, but tends to be too conservative, particularly for unequal sample sizes. Overall, the Saddlepoint approximation with fixed α and the MLE $\hat{\lambda}$ from the pooled data appears to have more variance than the other methods. Comparison with Figures S22 and S23 in Web Appendix D suggests that our resampling algorithm might be more reliable in this setting than moment corrected correlation (MCC) (Zhou and Wright, 2015) under the alternative and for unequal sample sizes.

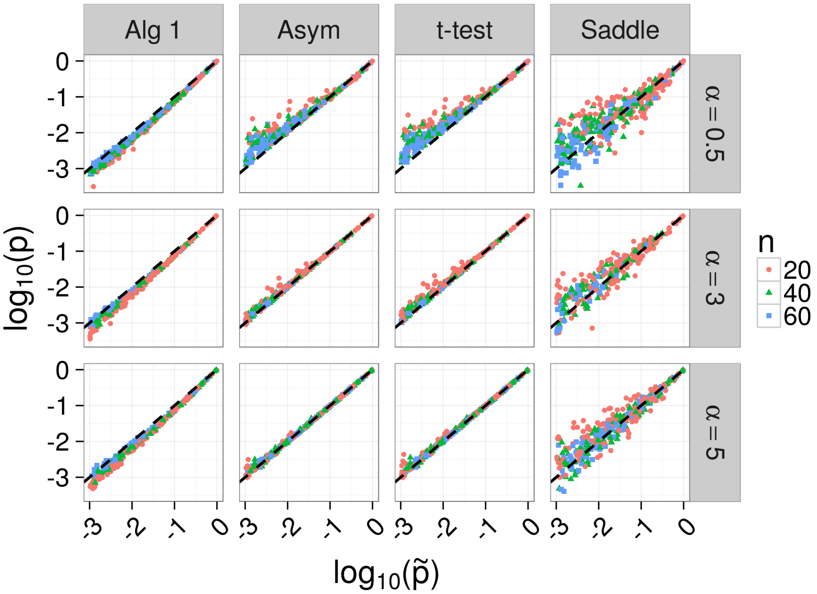


Figure S11: Simulation results using the statistic $T = |\bar{x} - \bar{y}|$ with gamma data and equal sample sizes of $n = n_x = n_y = 20$, 40, 60. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *t-test* is a t-test with unequal variance, and *Saddle* is the saddlepoint approximation (see Web Appendix B). $\tilde{p}$ is the p-values from simple Monte Carlo resampling with 10^5 resamples. SAMC results are not shown, as the EXPERT package does not produce p-values larger than 10^{-3}. Only simulations with $\tilde{p} > 10^{-3}$ shown (1,023 values shown). The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods. This figure appears in color in the electronic version of this article.

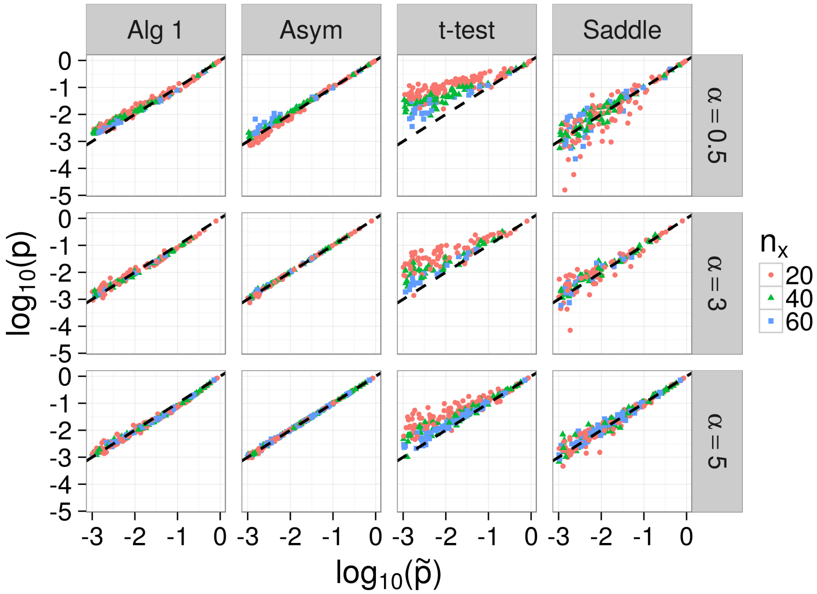


Figure S12: Simulation results using the statistic $T = |\bar{x} - \bar{y}|$ with gamma data and unequal sample sizes of $n_x = 20, 40, 60$ and $n_y = 100$. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *SAMC* is the SAMC algorithm, and p_t is a two-sided t-test with equal variance. SAMC results are not shown, as the EXPERT package does not produce p-values larger than 10^{-3}. Only simulations with $\tilde{p} > 10^{-3}$ shown (573 values shown). The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods. This figure appears in color in the electronic version of this article.

C.3.2 Under the null hypothesis $P_x = P_y$

We generated data $x_i, i = 1, \ldots, n_x$ and $y_j, j = 1, \ldots, n_y$ as realizations of the respective random variables $X_i \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \lambda)$ and $Y_j \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \lambda)$ for $\alpha = 0.5, 3, 5$ and $\lambda = 1, 5$, where λ is the rate parameter. For equal sample sizes, we set $n = n_x = n_y = 20, 40, 60$, and for unequal sample sizes we set $n_x = 20, 40, 60$ and $n_y = 100$. For both equal and unequal sample sizes, and for each each n or n_x and combination of α and λ, we simulated 1,000 datasets (we used 1,000 datasets instead of 100 to better investigate the type I error rate). We used the p-value from simple Monte Carlo resampling with 10^5 resamples, denoted as $\tilde{p}$, as an approximation for the true permutation p-value.

Results for equal and unequal sample size are shown in Figures S13 and S14, respectively. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *Saddle* is the saddlepoint approximation described in Web Appendix B, *t-test* shows the p-value from a two-sided t-test with unequal variance, and $\tilde{p}$ is from simple Monte Carlo resampling with 10^5 resamples. We do not show results from the SAMC algorithm, because the `EXPERT` package (Yu et al., 2011) does not provide results for p-values $> 10^{-3}$.

Figures S13 and S14 suggest that our methods work well in this setting, and have less variability than both the t-test and saddlepoint approximation (using fixed α fixed and the MLE $\hat{\lambda}$ from the pooled data).

Tables S5 and S6 show the Type I error rates under the null $H_0 : P_x = P_y$ for the equal and unequal sample size simulations, respectively. *MC* is the unadjusted p-value from simple Monte Carlo resampling and 10^5 resamples, *Saddle* is the saddlepoint approximation described in Web Appendix B, *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, and *t-test* shows the p-value from a two-sided t-test with equal variance.

C.4 Ratio of means with gamma data

In this subsection, we use the statistic $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ with data generated as gamma random variables.

C.4.1 Small sample sizes

We generated data $x_i, i = 1, \ldots, n_x$ and $y_j, j = 1, \ldots, n_y$ as realizations of the respective random variables $X_i \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \lambda_x)$ and $Y_j \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \lambda_y)$, where λ is the rate parameter, and $\alpha = 0.5, 3, 5$. For equal sample sizes, we set $n = n_x = n_y = 20, 40, 60$, and for unequal sample sizes, we set $n_x = 20, 40, 60$ and $n_y = 100$. For all simulations, we set $\lambda_x = 1$.

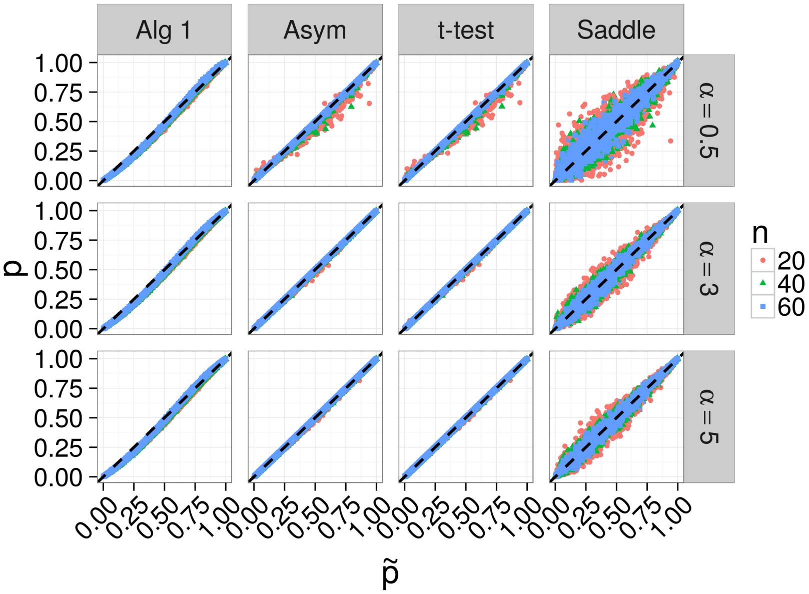


Figure S13: Simulation results using the statistic $T = |\bar{x} - \bar{y}|$ with gamma data under the null $P_x = P_y$ with equal sample sizes of $n = n_x = n_y = 20, 40, 60$. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *Saddle* is the saddlepoint approximation described in Web Appendix B, *t-test* shows the p-value from a two-sided t-test with unequal variance, and $\tilde{p}$ is from simple Monte Carlo resampling with 10^5 resamples. The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods. This figure appears in color in the electronic version of this article.

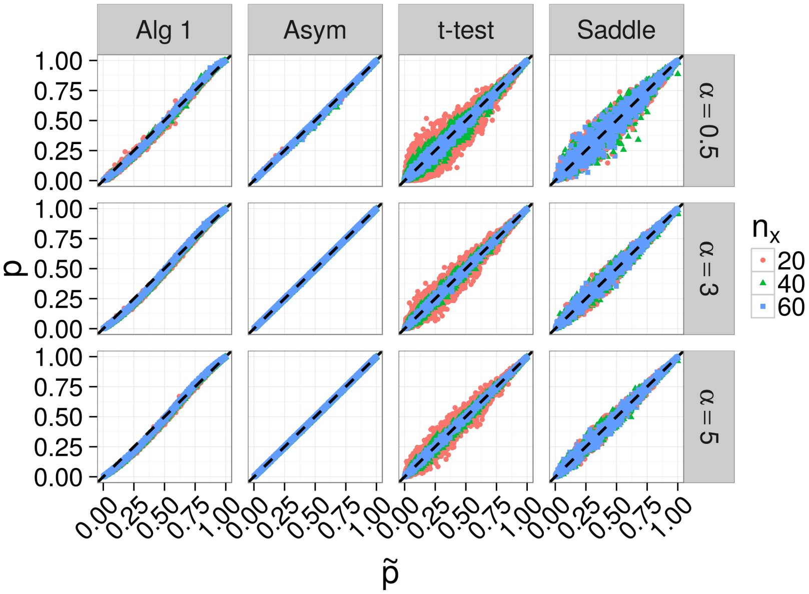


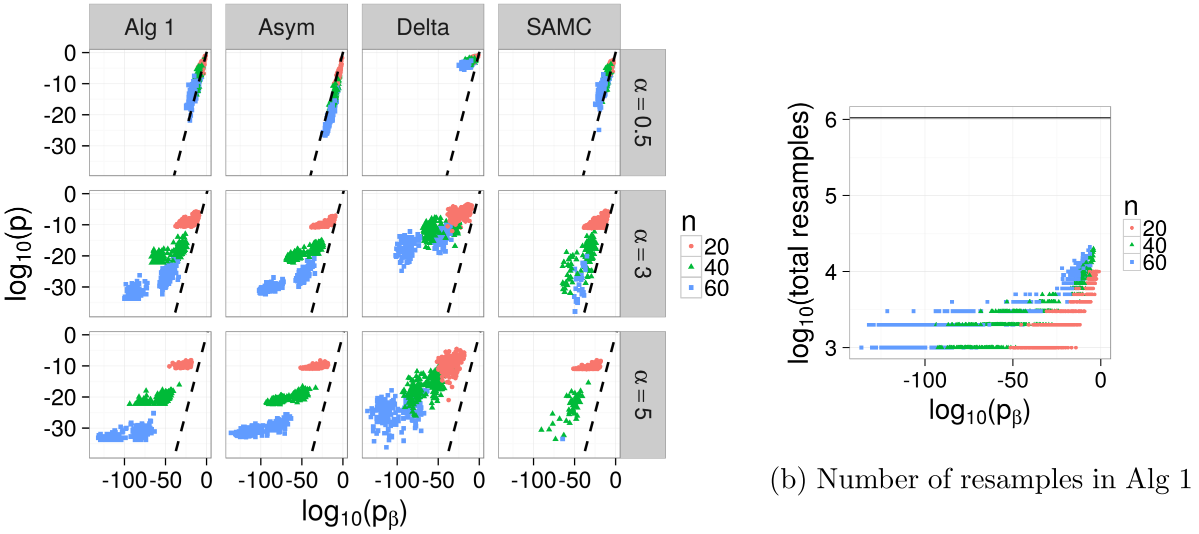
Figure S14: Simulation results using the statistic $T = |\bar{x} - \bar{y}|$ with gamma data under the null $P_x = P_y$ with unequal sample sizes of $n_x = 20, 40, 60$ and $n_y = 100$. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *Saddle* is the saddlepoint approximation described in Web Appendix B, *t-test* shows the p-value from a two-sided t-test with unequal variance, and $\tilde{p}$ is from simple Monte Carlo resampling with 10^5 resamples. The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods. This figure appears in color in the electronic version of this article.

Table S5: Type I error rates Pr(p-value $\leq$ signif level$|H_0$) for $T = |\bar{x} - \bar{y}|$ with gamma data and equal sample sizes $n = n_x = n_y$. *MC* is the unadjusted p-value from simple Monte Carlo resampling with 10^5 resamples, *Saddle* is the saddlepoint approximation described in Web Appendix B, *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, and *t-test* is a two-sided t-test with equal variance.

α	signif level	n_x	MC	Saddle	Alg 1	Asym	t-test
0.5	0.01	20	0.0110	0.0100	0.0165	0.0060	0.0045
		40	0.0125	0.0110	0.0150	0.0090	0.0085
		60	0.0115	0.0085	0.0140	0.0105	0.0105
	0.05	20	0.0495	0.0560	0.0665	0.0460	0.0410
		40	0.0515	0.0490	0.0660	0.0520	0.0485
		60	0.0455	0.0450	0.0595	0.0435	0.0425
	0.1	20	0.1000	0.1020	0.1280	0.1020	0.0945
		40	0.0995	0.0950	0.1260	0.1020	0.0975
		60	0.0980	0.0950	0.1230	0.0990	0.0965
3	0.01	20	0.0115	0.0070	0.0165	0.0095	0.0095
		40	0.0120	0.0115	0.0150	0.0120	0.0120
		60	0.0075	0.0075	0.0080	0.0070	0.0070
	0.05	20	0.0510	0.0465	0.0715	0.0515	0.0495
		40	0.0545	0.0575	0.0680	0.0560	0.0525
		60	0.0470	0.0475	0.0665	0.0480	0.0475
	0.1	20	0.0940	0.0990	0.1280	0.0980	0.0940
		40	0.0990	0.1000	0.1320	0.0990	0.0980
		60	0.0980	0.0985	0.1230	0.0980	0.0980
5	0.01	20	0.0115	0.0095	0.0175	0.0115	0.0115
		40	0.0090	0.0065	0.0130	0.0080	0.0080
		60	0.0045	0.0055	0.0085	0.0040	0.0040
	0.05	20	0.0525	0.0525	0.0675	0.0525	0.0505
		40	0.0525	0.0545	0.0715	0.0535	0.0520
		60	0.0460	0.0445	0.0580	0.0470	0.0470
	0.1	20	0.0965	0.0960	0.1220	0.0980	0.0955
		40	0.1070	0.1060	0.1370	0.1080	0.1080
		60	0.0925	0.0905	0.1300	0.0940	0.0915

Table S6: Type I error rates Pr(p-value $\leq$ signif level$|H_0$) for $T = |\bar{x} - \bar{y}|$ with gamma data and unequal sample sizes $n_x \neq n_y$ (n_x shown and $n_y = 100$). α is the shape parameter in the gamma distribution, *MC* is the unadjusted p-value from simple Monte Carlo resampling with 10^5 resamples, *Saddle* is the saddlepoint approximation described in Web Appendix B, *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, and *t-test* is a two-sided t-test with equal variance.

α	signif level	n_x	MC	Saddle	Alg 1	Asym	t-test
0.5	0.01	20	0.0095	0.0095	0.0105	0.0085	0.0245
		40	0.0090	0.0060	0.0105	0.0070	0.0140
		60	0.0130	0.0160	0.0170	0.0105	0.0135
	0.05	20	0.0460	0.0465	0.0675	0.0440	0.0740
		40	0.0455	0.0470	0.0620	0.0445	0.0540
		60	0.0505	0.0500	0.0670	0.0495	0.0530
	0.1	20	0.0915	0.0930	0.1260	0.0845	0.1220
		40	0.0980	0.0945	0.1280	0.0960	0.1040
		60	0.1100	0.1080	0.1410	0.1100	0.1080
3	0.01	20	0.0085	0.0095	0.0155	0.0085	0.0135
		40	0.0135	0.0120	0.0185	0.0135	0.0140
		60	0.0070	0.0055	0.0090	0.0070	0.0070
	0.05	20	0.0440	0.0440	0.0665	0.0435	0.0480
		40	0.0480	0.0555	0.0695	0.0485	0.0530
		60	0.0470	0.0495	0.0635	0.0485	0.0460
	0.1	20	0.0875	0.0885	0.1260	0.0885	0.1000
		40	0.1050	0.1040	0.1350	0.1060	0.0975
		60	0.1040	0.1080	0.1370	0.1040	0.1040
5	0.01	20	0.0140	0.0110	0.0200	0.0140	0.0145
		40	0.0090	0.0100	0.0155	0.0090	0.0100
		60	0.0105	0.0090	0.0120	0.0110	0.0075
	0.05	20	0.0540	0.0535	0.0845	0.0540	0.0620
		40	0.0530	0.0525	0.0730	0.0525	0.0555
		60	0.0520	0.0510	0.0635	0.0520	0.0500
	0.1	20	0.1140	0.1160	0.1520	0.1140	0.1130
		40	0.0995	0.1000	0.1300	0.0995	0.1040
		60	0.1040	0.0985	0.1320	0.1050	0.1060



(b) Number of resamples in Alg 1

(a) p-values

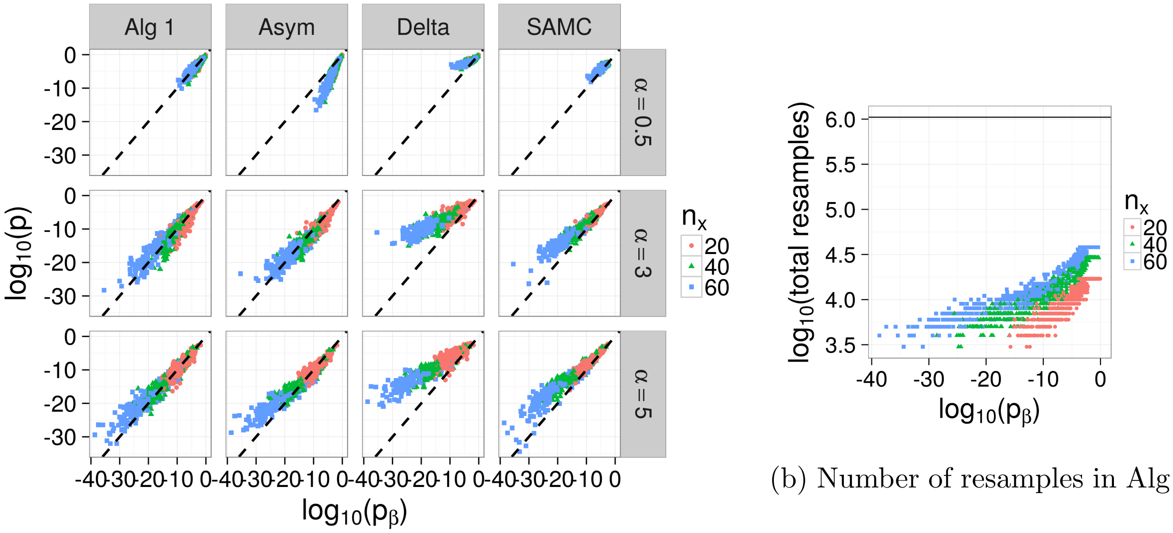
Figure S15: Simulation results using the statistic $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ with gamma data and equal sample sizes of $n = n_x = n_y = 20$, 40, 60. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *Delta* is the delta method, *SAMC* is the SAMC algorithm, and p_β is the two-sided p-value from the beta prime distribution. The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods. The horizontal line in S15b shows the number of iterations used in the SAMC algorithm (set in advance, and independent of p-value). The SAMC algorithm did not produce values for 652 tests (points missing). This figure appears in color in the electronic version of this article.

For equal samples sizes, we set $\lambda_y = 7, 12.5$ for each n. For unequal sample sizes, we set $\lambda_y = 2.25, 2.75$ for all n_x for $\alpha = 0.5$, $\lambda_y = 2, 2.5$ for all n_x for $\alpha = 3$, and $\lambda_y = 1.75, 2.25$ for all n_x for $\alpha = 5$. We simulated 100 datasets for each combination of parameters.

Results for equal and unequal sample size are shown in Figures S15 and S16, respectively. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *Delta* is the delta method, *SAMC* is the SAMC algorithm, and p_β is the two-sided p-value from the beta prime distribution. Figures S15b and S16b show the number of resamples used by our resampling algorithm.

C.4.2 Under the null hypothesis $P_x = P_y$

We generated data $x_i, i = 1, \ldots, n_x$ and $y_j, j = 1, \ldots, n_y$ as realizations of the respective random variables $X_i \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, 1)$ and $Y_j \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, 1)$ for $\alpha = 0.5, 3, 5$. For equal sample sizes, we set $n = n_x = n_y = 20, 40, 60$. For unequal sample sizes, we set $n_x = 20, 40, 60$



(b) Number of resamples in Alg 1

(a) p-values

Figure S16: Simulation results using the statistic $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ with gamma data and unequal sample sizes of $n_x = 20$, 40, 60, $n_y = 100$, and rates $\lambda_y = 5, 10$, and $\lambda_x = 1$. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *Delta* is the delta method, *SAMC* is the SAMC algorithm, and p_β is the two-sided p-value from the beta prime distribution. The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods. The horizontal line in S16b shows the number of iterations used in the SAMC algorithm (set in advance, and independent of p-value). The SAMC algorithm did not produce values for 304 tests (points missing). This figure appears in color in the electronic version of this article.

and $n_y = 100$. For both equal and unequal sample sizes, we simulated 1,000 datasets for each combination of parameters (we used 1,000 datasets instead of 100 to better investigate the type I error rate). We used the p-value from simple Monte Carlo resampling with 10^5 resamples, denoted as $\tilde{p}$, as an approximation for the true permutation p-value.

Results for equal and unequal sample size are shown in Figures S17 and S18, respectively. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *Delta* is the delta method, *Beta prime* gives the p-value from the beta prime distribution, and $\tilde{p}$ is from simple Monte Carlo resampling with 10^5 resamples. Given the large p-values, using 10^5 Monte Carlo resamples should be sufficient to obtain reliable estimates of the true permutation p-value. Therefore, this comparison demonstrates that the permutation p-value is not exactly the same as the p-value from the beta prime distribution. However, it appears reasonably close, so we use it as an approximation to the truth in other simulations in which the p-values are much smaller and simple Monte Carlo methods are not feasible.

We do not show results from the SAMC algorithm, because as noted above, the `EXPERT` package (Yu et al., 2011) does not provide results for p-values $> 10^{-3}$.

Tables S7 and S8 show the Type I error rates under the null $H_0 : P_x = P_y$ for the equal and unequal sample sizes, respectively. *MC* is the unadjusted p-value from simple Monte Carlo resampling with 10^5 resamples, *Beta prime* is the p-value from the beta prime distribution, *Alg 1* is our resampling algorithm, and *Asym* is our asymptotic approximation.

D Comparison with additional methods

D.1 Moment-corrected correlation

Moment-corrected correlation (MCC) (Zhou and Wright, 2015) is an analytical approximation to the permutation p-value, which is applicable in multiple testing situations in which the test statistic is permutationally equivalent to a single inner product. Where applicable, this approach is fast, as it does not involve resampling. However, if the test statistic of interest is not permutationally equivalent to an inner product, the MCC approach cannot be used.

The statistic $T = \bar{x} - \bar{y}$ fits into this setting, whereas, to the best of our knowledge, $T = \bar{x}/\bar{y}$ does not. To see this, let $\boldsymbol{z} = (\boldsymbol{x}', \boldsymbol{y}')'$ and $\boldsymbol{w} = (\underbrace{1/n_x, \ldots, 1/n_x}_{n_x}, \underbrace{-1/n_y, \ldots, -1/n_y}_{n_y})'$. Then $\bar{x} - \bar{y} = \boldsymbol{z}'\boldsymbol{w}$. In contrast, $\bar{x}/\bar{y}$ cannot be written in this form, and we conjecture that it is not permutationally equivalent to any statistic that can be.

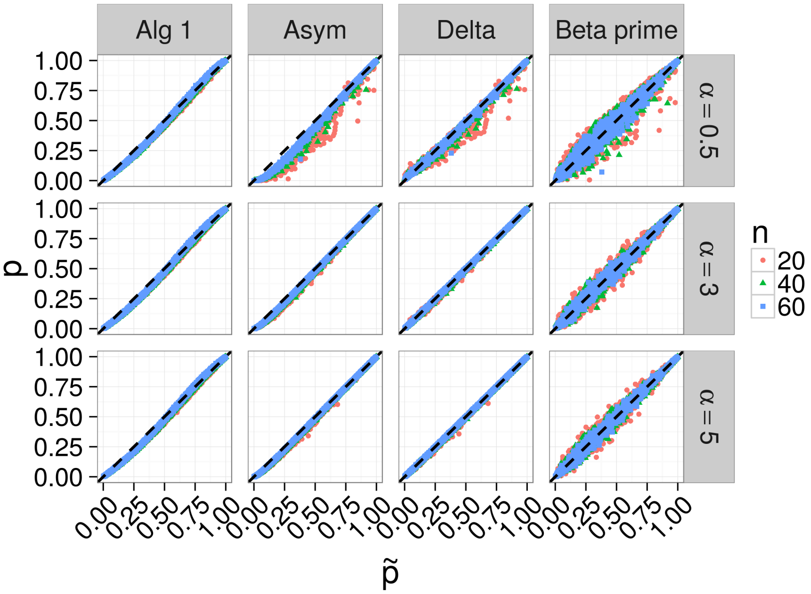


Figure S17: Simulation results using the statistic $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ with gamma data under the null of $P_x = P_y$ with equal sample sizes of $n = n_x = n_y = 20$, 40, 60. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *Delta* is the delta method, *Beta prime* gives the p-value from the beta prime distribution, and $\tilde{p}$ is from simple Monte Carlo resampling with 10^5 resamples. The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods. This figure appears in color in the electronic version of this article.

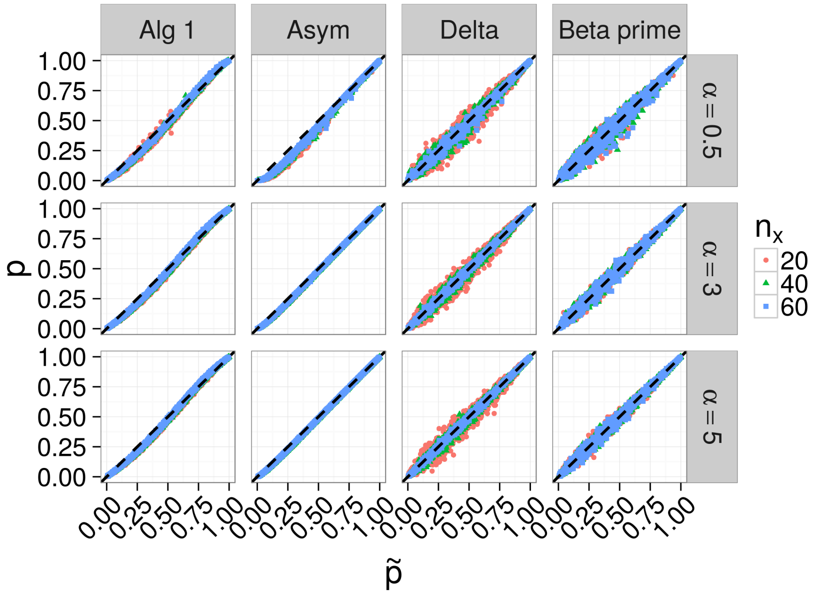


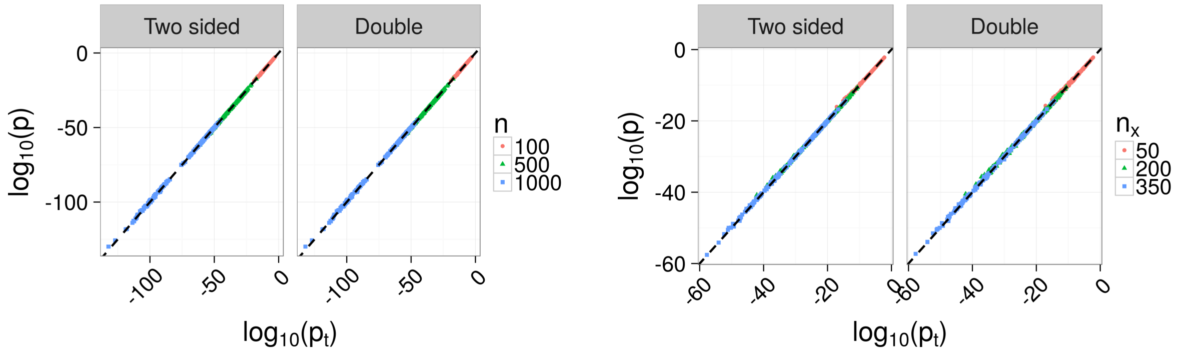
Figure S18: Simulation results using the statistic $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ with gamma data under the null of $P_x = P_y$ with unequal sample sizes of $n_x = 20$, 40, 60 and $n_y = 100$. *Alg 1* is our resampling algorithm with $B_{\text{pred}} = 10^3$ resamples in each partition, *Asym* is our asymptotic approximation, *Delta* is the delta method, *Beta prime* gives the p-value from the beta prime distribution, and $\tilde{p}$ is from simple Monte Carlo resampling with 10^5 resamples. The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods. This figure appears in color in the electronic version of this article.

Table S7: Type I error rates Pr(p-value $\leq$ signif level$|H_0$) for $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ with gamma data and equal sample sizes $n = n_x = n_y$. α is the shape parameter in the gamma distribution, *MC* is simple Monte Carlo resampling with 10^5 resamples, *Alg 1* is our resampling algorithm, *Asym* is our asymptotic approximation, *Delta* is the delta method, and *Beta prime* is the the beta prime distribution.

α	signif level	n	MC	Alg 1	Asym	Delta	Beta prime
0.5	0.01	20	0.013	0.018	0.093	0.002	0.015
		40	0.007	0.014	0.055	0.001	0.007
		60	0.007	0.010	0.047	0.002	0.011
	0.05	20	0.050	0.076	0.182	0.026	0.053
		40	0.050	0.072	0.135	0.037	0.055
		60	0.048	0.068	0.114	0.043	0.050
	0.1	20	0.110	0.136	0.243	0.106	0.108
		40	0.106	0.135	0.196	0.114	0.104
		60	0.096	0.127	0.178	0.101	0.097
3	0.01	20	0.007	0.012	0.027	0.003	0.006
		40	0.012	0.016	0.025	0.010	0.010
		60	0.012	0.015	0.025	0.012	0.008
	0.05	20	0.043	0.067	0.088	0.046	0.044
		40	0.053	0.062	0.073	0.052	0.051
		60	0.059	0.075	0.080	0.061	0.049
	0.1	20	0.095	0.126	0.143	0.103	0.090
		40	0.098	0.133	0.147	0.104	0.103
		60	0.095	0.115	0.116	0.097	0.093
5	0.01	20	0.009	0.015	0.023	0.009	0.009
		40	0.008	0.013	0.025	0.008	0.011
		60	0.012	0.012	0.019	0.012	0.013
	0.05	20	0.046	0.063	0.082	0.054	0.052
		40	0.048	0.063	0.066	0.050	0.043
		60	0.055	0.078	0.079	0.057	0.057
	0.1	20	0.093	0.130	0.139	0.106	0.099
		40	0.091	0.134	0.138	0.094	0.093
		60	0.115	0.138	0.136	0.116	0.112

Table S8: Type I error rates Pr(p-value $\leq$ signif level$|H_0$) for $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$ with gamma data and unequal sample sizes $n_x \neq n_y$ (n_x shown and $n_y = 100$). α is the shape parameter in the gamma distribution, *MC* is simple Monte Carlo resampling with 10^5 resamples, *Alg 1* is our resampling algorithm, *Asym* is our asymptotic approximation, *Delta* is the delta method, and *Beta prime* is the beta prime distribution.

α	signif level	n_x	MC	Alg 1	Asym	Delta	Beta prime
0.5	0.01	20	0.011	0.015	0.065	0.006	0.011
		40	0.015	0.018	0.053	0.003	0.013
		60	0.008	0.011	0.042	0.003	0.012
	0.05	20	0.043	0.069	0.128	0.047	0.053
		40	0.057	0.072	0.133	0.048	0.056
		60	0.052	0.071	0.112	0.045	0.050
	0.1	20	0.098	0.121	0.179	0.109	0.091
		40	0.113	0.141	0.195	0.119	0.108
		60	0.106	0.126	0.172	0.109	0.098
3	0.01	20	0.011	0.016	0.023	0.012	0.011
		40	0.005	0.011	0.027	0.005	0.009
		60	0.011	0.013	0.017	0.011	0.011
	0.05	20	0.047	0.070	0.073	0.059	0.039
		40	0.058	0.065	0.069	0.057	0.054
		60	0.053	0.066	0.070	0.050	0.052
	0.1	20	0.088	0.128	0.135	0.104	0.087
		40	0.094	0.124	0.124	0.101	0.089
		60	0.094	0.119	0.117	0.097	0.097
5	0.01	20	0.010	0.014	0.022	0.007	0.009
		40	0.011	0.011	0.017	0.011	0.009
		60	0.015	0.020	0.025	0.015	0.018
	0.05	20	0.058	0.074	0.085	0.066	0.054
		40	0.046	0.057	0.059	0.048	0.052
		60	0.059	0.081	0.085	0.061	0.062
	0.1	20	0.110	0.145	0.143	0.121	0.114
		40	0.081	0.114	0.108	0.085	0.088
		60	0.113	0.145	0.138	0.118	0.115



(a) $n_x = n_y$

(b) $n_x \neq n_y$

Figure S19: MCC with large sample sizes for $T = |\bar{x} - \bar{y}|$ with normal data and equal sample sizes of $n = n_x = n_y = 100, 500, 1,000$, and unequal sample sizes of $n_x = 50, 200, 350$ with $n_y = 500$. In both cases, data were simulated as normal random variables with $\mu_y = 0$, $\mu_x = 0.75, 1$ and $\sigma_x^2 = \sigma_x^2 = 1$. p_t is the p-value from a t-test with equal variance. The diagonal dashed line has a slope of 1 and an intercept of 0, and indicates agreement. This figure appears in color in the electronic version of this article.

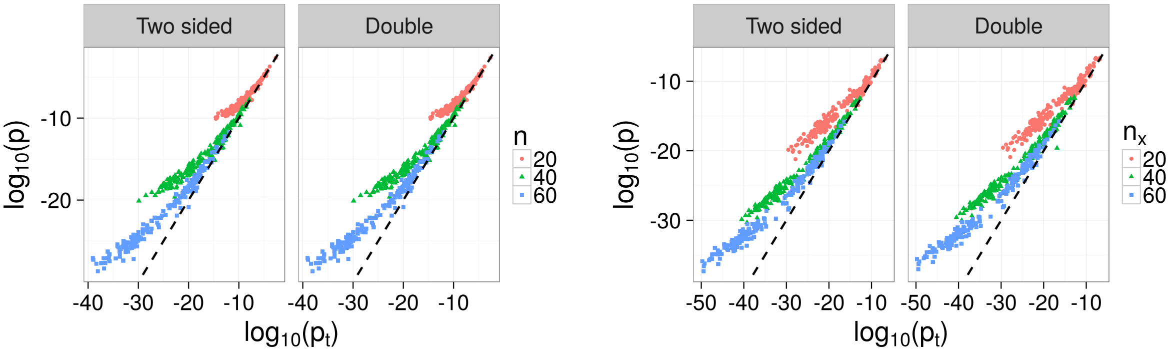
Figures S19 through S21 show simulation results for two-sided and doubled p-values, as described by Zhou and Wright (2015), using the `mcc` package (Zhou, 2014) under the same normal data settings as in Section C.1. While MCC is more reliable for large sample sizes (Figure S19), MCC appears to suffer from the same bias as our methods for small sample sizes (Figure S20). Furthermore, we do not think that MCC can be used to obtain p-values for the statistic $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$.

Figures S22 and S23 show simulation results for two-sided and doubled p-values for small sample sizes and under the null, respectively, using the `mcc` package (Zhou, 2014) under the same gamma data settings as in Section C.3. In Figure S22, we used $B = 10^5$ resamples to obtain the Monte Carlo estimate $\tilde{p}$ of the true permutation p-value, and only show results for $\tilde{p} > 10^{-3}$ to ensure reliable estimates (1,019 values shown in Figure S22a, and 705 values shown in Figure S22b).

As seen in Figure S22, in many cases the MCC method substantially underestimated the permutation p-value for equal sample sizes $n_x = n_y$ and $\alpha = 0.5$. We did not observe this tendency with our resampling algorithm (see Figures S11 and S12).

D.2 Saddlepoint approximations

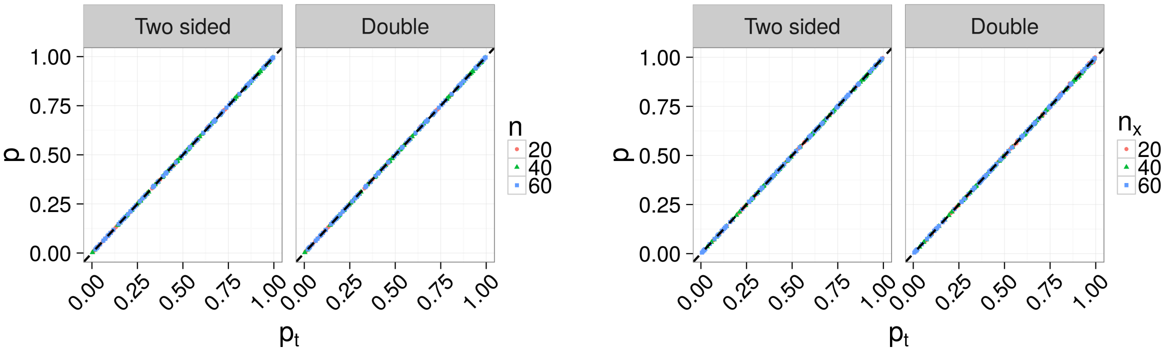
Saddlepoint approximations can be used to estimate permutation p-values (Robinson, 1982). As shown in Table S9, estimates from our methods are comparable to those from saddle-



(a) $n_x = n_y$

(b) $n_x \neq n_y$

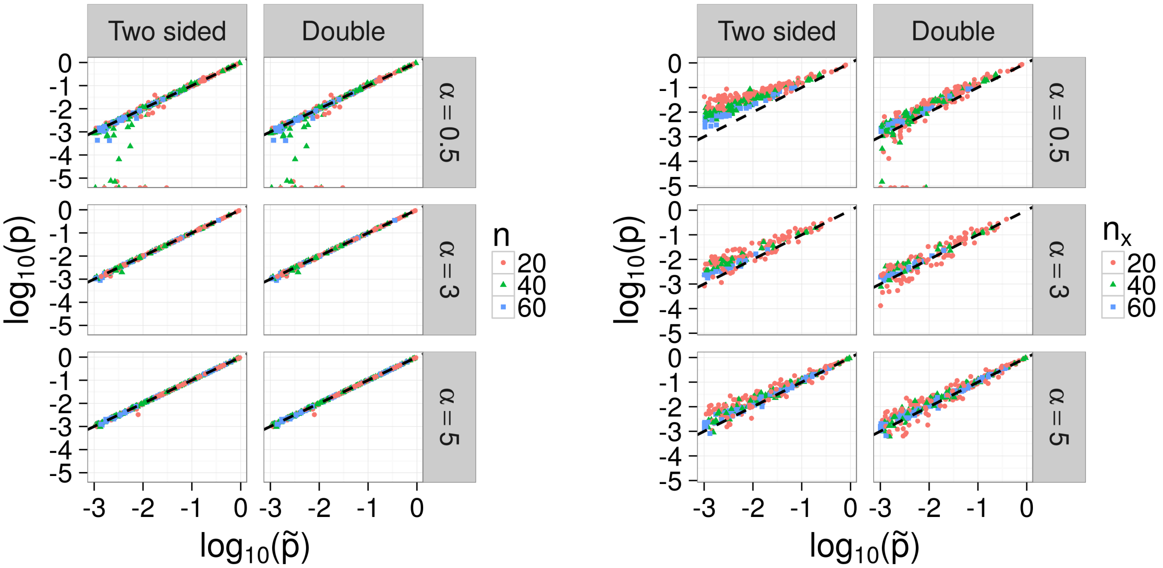
Figure S20: MCC with small sample size for $T = |\bar{x} - \bar{y}|$ with normal data and equal sample sizes of $n = n_x = n_y = 20, 40, 60$, and unequal sample sizes of $n_x = 20, 40, 60$ with $n_y = 100$. In both cases, data were simulated as normal random variables with $\mu_y = 0$, $\mu_x = 2, 3$ and $\sigma_x^2 = \sigma_x^2 = 1$. p_t is the p-value from a t-test with equal variance. The diagonal dashed line has a slope of 1 and an intercept of 0, and indicates agreement. This figure appears in color in the electronic version of this article.



(a) $n_x = n_y$

(b) $n_x \neq n_y$

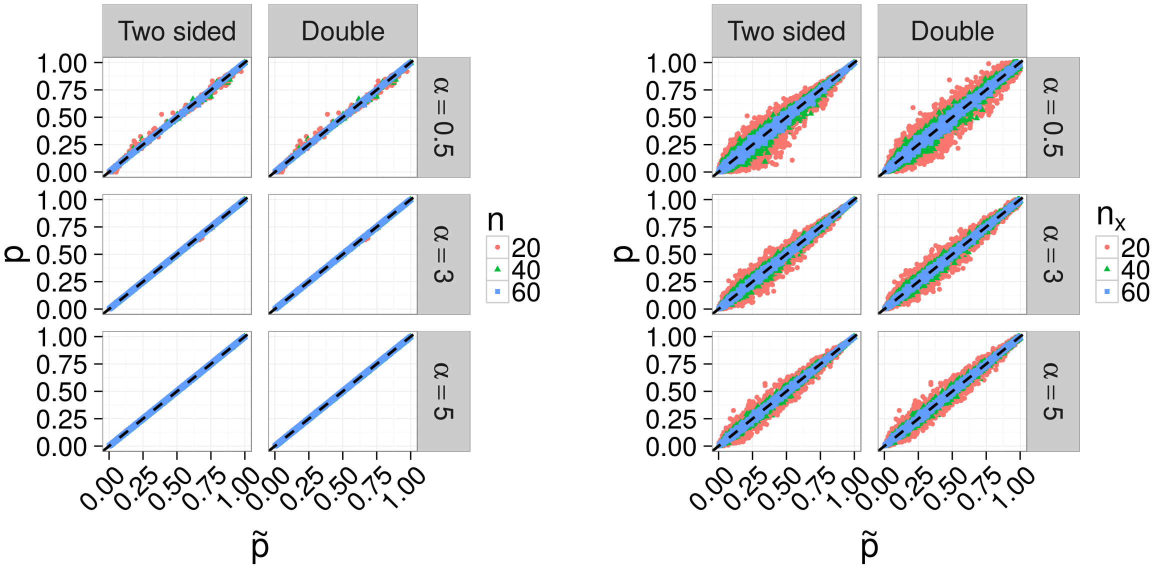
Figure S21: MCC under the null hypothesis for $T = |\bar{x} - \bar{y}|$ with normal data for equal sample sizes of $n = n_x = n_y = 20, 40, 60$, and unequal sample sizes of $n_x = 20, 40, 60$ with $n_y = 100$. In both cases, data were simulated as normal random variables with $\mu_y = \mu_x = 0$ and $\sigma_x^2 = \sigma_x^2 = 1$. p_t is the p-value from a t-test with equal variance. The diagonal dashed line has a slope of 1 and an intercept of 0, and indicates agreement. This figure appears in color in the electronic version of this article.



(a) $n_x = n_y$

(b) $n_x \neq n_y$

Figure S22: MCC with small sample size for $T = |\bar{x} - \bar{y}|$ with gamma data and equal sample size $n = n_x = n_y = 20, 40, 60$, and unequal sample sizes of $n_x = 20, 40, 60$ with $n_y = 100$. In both cases, data were simulated as gamma random variables, as described in Section C.3. $\tilde{p}$ is the p-value from simple Monte Carlo resampling with 10^5 resamples. The diagonal dashed line has a slope of 1 and an intercept of 0, and indicates agreement. This figure appears in color in the electronic version of this article.



(a) $n_x = n_y$

(b) $n_x \neq n_y$

Figure S23: MCC under the null hypothesis for $T = |\bar{x} - \bar{y}|$ with gamma data for equal sample sizes of $n = n_x = n_y = 20, 40, 60$, and unequal sample sizes of $n_x = 20, 40, 60$ with $n_y = 100$. In both cases, data were simulated as gamma random variables, as described in Section C.3. $\tilde{p}$ is the p-value from simple Monte Carlo resampling with 10^5 resamples. The diagonal dashed line has a slope of 1 and an intercept of 0, and indicates agreement. This figure appears in color in the electronic version of this article.

Table S9: Comparison with Saddlepoint approximations for $T = |\bar{x} - \bar{y}|$. Datasets are from Robinson (1982, Table 2), who obtained them from Lehman (1975). Dataset 1 pertains to hours of pain relief due to two different drugs ($n_x = n_y = 8$), and Dataset 2 pertains to the effect of an analgesia for two classes ($n_x = 7, n_y = 10$). The exact and saddlepoint p-values are from Robinson (1982). The the p-value from our resampling algorithm ($\tilde{p}_{\text{pred}}$) is the mean from 100 runs; the first and third quantiles were (0.080, 0.088) for dataset 1, and (0.011, 0.012) for dataset 2.

Method	Dataset 1	Dataset 2
Exact	0.102	0.012
First saddlepoint	0.089	0.010
Second saddlepoint	0.101	0.011
$\tilde{p}_{\text{pred}}$	0.083	0.012
$\hat{p}_{\text{asym}}$	0.092	0.013

point approximations when using the statistic $T = |\bar{x} - \bar{y}|$. However, unlike saddlepoint approximations, our resampling algorithm requires no derivations.

E Simulations under null hypotheses for single parameters

Neuhaus (1993), Janssen (1997), Chung et al. (2013), and others have extended permutation tests to be valid not only under the null $P_x = P_y$, but also under the more general null that $\theta(P_x) = \theta(P_y)$, where $\theta(P)$ is a single parameter. For example, for $X \sim N(\mu_x, \sigma_x^2), Y \sim N(\mu_y, \sigma_y^2)$, we might be interested in the alternative $H_1 : \mu_x \neq \mu_y$, even if $\sigma_x^2 \neq \sigma_y^2$.

As described by Chung et al. (2013), in order to obtain a test procedure that is asymptotically valid in the above setting where $\sigma_x^2 \neq \sigma_y^2$, we need to replace $T = |\bar{x} - \bar{y}|$ with the studentized statistic

$$T = \frac{|\bar{x} - \bar{y}|}{\sqrt{s_x^2/n_x + s_y^2/n_y}} \tag{12}$$

where $s_x^2 = (n_x - 1)^{-1} \sum_i (x_i - \bar{x})^2$ and $s_y^2 = (n_y - 1)^{-1} \sum_j (y_j - \bar{y})^2$ are the sample variances. For each permutation, we compute the quantities $\bar{x}^*, \bar{y}^*, {s_x^*}^2$, and ${s_y^*}^2$ with the permuted datasets. In this section, we conduct simulations using (12) when $P_x \neq P_y$ under the null $H_0 : \mu_x = \mu_y$ and alternative $H_1 : \mu_x \neq \mu_y$.

We generated data $x_i, i = 1, \ldots, n_x$ and $y_j, j = 1, \ldots, n_y$ as realizations of the respective random variables $X_i \overset{\text{iid}}{\sim} N(0, \sigma_x^2)$ and $Y_j \overset{\text{iid}}{\sim} N(0, \sigma_y^2)$, where $\sigma_x^2 = 9$ and $\sigma_y^2 = 1$. For equal sample sizes, we set $n = n_x = n_y = 20, 40, 60$, and for unequal sample sizes we set

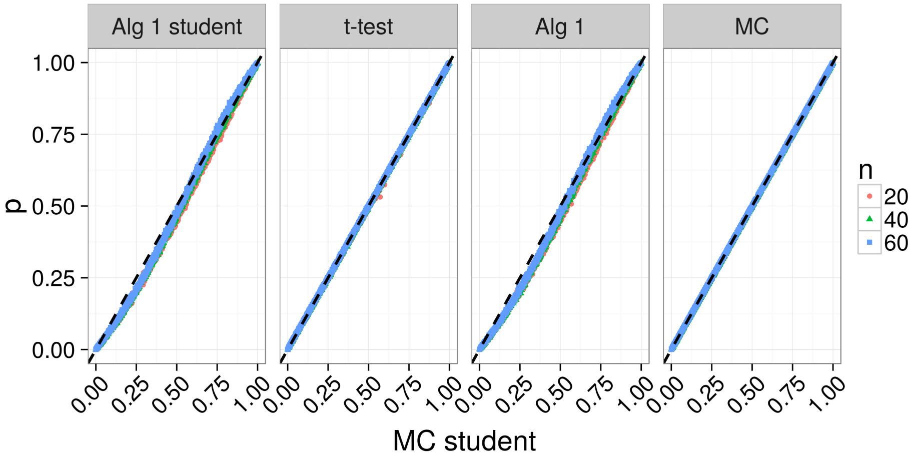


Figure S24: Simulation results under the null $\mu_x = \mu_y$ with normal data and unequal sample sizes of $n = n_x = n_y = 20, 40, 60$. *Alg 1 Student* and *Alg 1* are our resampling algorithm with the studentized (12) and unstudentized statistics, with $B_{\text{pred}} = 10^3$ resamples in each partition. *t-test* is the p-value from a two-sided t-test with unequal variance. *MC student* and *MC* are Monte Carlo estimates with the studentized (12) and unstudentized statistics, respectively, with 10^5 resamples. The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods. This figure appears in color in the electronic version of this article.

$n_x = 20, 40, 60$ and $n_y = 100$. For both equal and unequal sample sizes, we simulated 1,000 datasets for each combination of parameters. Figures S24 and S25 show the results with equal and unequal sample sizes, respectively.

As seen in Figures S24 and S25, the permutation test with the unstudentized statistic is relatively unaffected in our simulation under equal sample sizes, but is inaccurate for unequal sample sizes. This is as expected. By using a studentized statistic, our method is accurate even for unequal sample sizes. For comparison, Figures S24 and S25 also show the p-value from a t-test with unequal variance, as well as a Monte Carlo estimate using the unstudentized statistic $T = |\bar{x} - \bar{y}|$.

F Run time and sufficient sample size

In this section, we provide further details on the run-time of our resampling algorithm and guidance regarding the sample sizes necessary for our test to be reliable.

Our resampling algorithm runs in $O(B_{\text{pred}} m_{\text{stop}})$ time. In our current implementation, we set B_{pred} a priori. Regarding m_{stop}, we obtain the following approximation for small

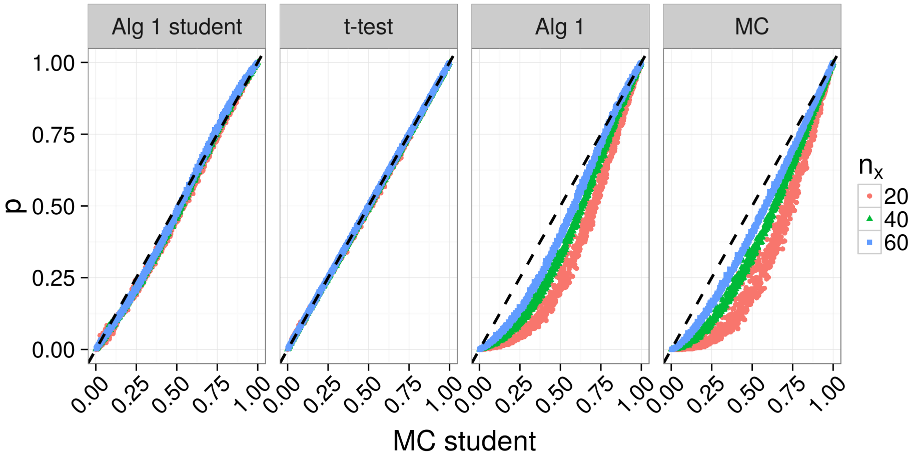


Figure S25: Simulation results under the null $\mu_x = \mu_y$ with normal data with unequal sample sizes of $n_x = 20, 40, 60$ and $n_y = 100$. *Alg 1 Student* and *Alg 1* are our resampling algorithm with the studentized (12) and unstudentized statistics, and with $B_{\text{pred}} = 10^3$ resamples in each partition. *t-test* is the p-value from a two-sided t-test with unequal variance. *MC student* and *MC* are Monte Carlo estimates with the studentized (12) and unstudentized statistics, respectively, with 10^5 resamples. The diagonal dashed line has slope of 1 and intercept of 0, and indicates agreement between methods. This figure appears in color in the electronic version of this article.

p-values, in which we assume that $1 - \Phi(\xi(m)) \gg 1 - \Phi(\xi^{\text{conj}}(m))$. From Algorithm 1,

$$\begin{aligned} m_{\text{stop}}^{\text{asym}} &= \min_m \{m \in \{1, \dots, m_{\max}\} : \boldsymbol{c}[m] < 1\} \\ &\approx \min_m \{m \in \{1, \dots, m_{\max}\} : \Pr(T(m) \geq t | \boldsymbol{x}, \boldsymbol{y}) < 1/B_{\text{pred}}\} \quad (\text{for large } B_{\text{pred}}) \\ &\approx \min_m \{m \in \{1, \dots, m_{\max}\} : 1 - \Phi(\xi(m)) < 1/B_{\text{pred}}\} \qquad (13) \\ &\approx \min_m \left\{m \in \{1, \dots, m_{\max}\} : \Phi^{-1}(1 - 1/B_{\text{pred}}) < \xi(m)\right\} \quad (\text{for large } n_x, n_y) \qquad (14) \\ &\equiv m_{\text{stop}}^{\text{asym}}, \end{aligned}$$

where (13) follows from (6) and the assumption that $1 - \Phi(\xi(m)) \gg 1 - \Phi(\xi^{\text{conj}}(m))$.

In the R package `fastPerm`, we provide functions for computing $m_{\text{stop}}^{\text{asym}}$, which can help an analyst to approximate run-time before running the algorithm. We emphasize that $m_{\text{stop}}^{\text{asym}}$ is based on asymptotic approximations, and may not be the same as the actual stopping partition; $m_{\text{stop}}^{\text{asym}}$ is not used in Algorithm 1. As shown in Figure S26, the expected stopping distribution $m_{\text{stop}}^{\text{asym}}$ appears to be a reasonable estimate of the actual stopping partition m_{stop} in our analysis of cancer genomic data.

We can also use $m_{\text{stop}}^{\text{asym}}$ to provide guidance on sample size. Note that $m_{\text{stop}}^{\text{asym}}$ is the ex-

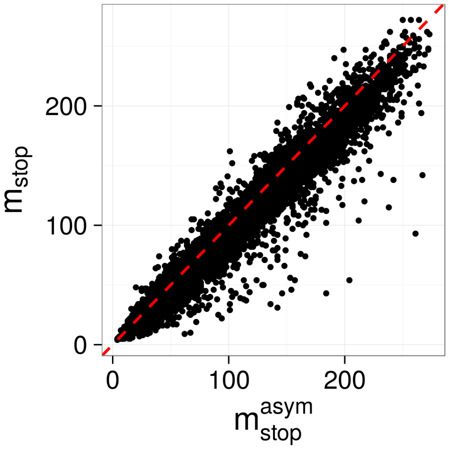


Figure S26: Comparison between $m_{\text{stop}}^{\text{asym}}$ and m_{stop} in the analysis of cancer genomic data. m_{stop} is the actual stopping partition, which our resampling algorithm determines dynamically. $m_{\text{stop}}^{\text{asym}}$ is our estimate of the stopping partition based on asymptotic approximations, and can be computed before running the algorithm. The dashed diagonal line has a slope of 1 and an intercept of 0, and indicates agreement. This figure appears in color in the electronic version of this article.

pected number of data points available to the Poisson regression in our resampling algorithm for estimating the overall p-value. Large values of $m_{\text{stop}}^{\text{asym}}$ imply more reliable but slower estimates, and smaller values of $m_{\text{stop}}^{\text{asym}}$ imply less reliable but faster estimates. To ensure that the results of the sampling algorithm are reliable, we recommend that $m_{\text{stop}}^{\text{asym}} \geq c$ for some constant c. For example, we use $c = 4$. Then for equal sample sizes $n = n_x = n_y$, we set

$$\hat{n} = \min_n \{n \in \mathbb{N} : m_{\text{stop}}^{\text{asym}} \geq c\}.$$

While not explicit in the above notation, we note that $m_{\text{stop}}^{\text{asym}}$, and thus $\hat{n}$, is a function of $\sigma_x^2, \sigma_y^2, \mu_x, \mu_y$, and B_{pred}. Tables S10 and S11 show $\hat{n}$ and $\hat{p}_{\text{asym}} = \hat{p}_{\text{asym}}(\hat{n}, \sigma_x^2, \sigma_y^2, \mu_x, \mu_y)$, the the p-value from our asymptotic approximation for the given set of parameter values and sample sizes. In Tables S10 and S11, we set $B_{\text{pred}} = 1,000$. As in Figure 1 in Section 3 and Figure S1 in Web Appendix A, to obtain $\hat{p}_{\text{asym}}$, we substituted parameter values for sample quantities, e.g. μ_x for $\bar{x}$ and σ_x^2 for $(n_x - 1)^{-1} \sum_{i=1}^{n_x} (x_i - \bar{x})^2$. As can be seen in Tables S10 and S11, $\hat{n}$ and $\hat{p}_{\text{asym}}$ have an inverse relationship.

In general, we recommend that researchers check the output from `fastPerm` to ensure that $m_{\text{stop}} \geq 4$, and we note that the sample sizes required to achieve $m_{\text{stop}} \geq 4$ increase as the p-value decreases. Based on Tables S10 and S11, at least 15-20 observations in each group appears sufficient for p-values near 1×10^{-6}, and at least 70-90 observations in each group appears sufficient for p-values near 1×10^{-30}.

Table S10: $\hat{n}$ for $T = \max(\bar{x}/\bar{y}, \bar{y}/\bar{x})$, equal samples sizes $n_x = n_y = \hat{n}$, $B_{\text{pred}} = 1,000$, and $c = 4$.

$\mu_y = \sigma_y^2$	$\mu_x = \sigma_x^2$	$\hat{n}$	$\hat{p}_{\text{asym}}$
2	3	5	2.4×10^{-1}
	4	6	2.4×10^{-2}
	5	13	2.4×10^{-5}
	5.25	16	1.3×10^{-6}
	5.5	19	6.0×10^{-8}
	5.75	24	4.2×10^{-10}
	6	31	4.1×10^{-13}
	6.25	40	4.3×10^{-17}
	6.5	55	1.1×10^{-23}
	6.6	63	3.3×10^{-27}
	6.7	74	4.5×10^{-32}
	6.8	87	7.7×10^{-38}
	6.9	105	7.8×10^{-46}
	7	130	6.0×10^{-57}

Table S11: $\hat{n}$ for $T = |\bar{x} - \bar{y}|$, $\sigma_x^2 = \sigma_y^2 = 1$, equal samples sizes $n_x = n_y = \hat{n}$, $B_{\text{pred}} = 1,000$, and $c = 4$.

μ_y	μ_x	$\hat{n}$	$\hat{p}_{\text{asym}}$
0	1.5	5	5.4×10^{-2}
	2	9	7.7×10^{-4}
	2.2	13	2.1×10^{-5}
	2.25	15	3.7×10^{-6}
	2.3	18	3.1×10^{-7}
	2.4	32	4.0×10^{-12}
	2.45	53	2.3×10^{-19}
	2.475	80	1.3×10^{-28}
	2.48	89	1.1×10^{-31}
	2.49	115	1.5×10^{-40}
	2.5	165	1.4×10^{-57}

G Asymptotic test of the ratio of means via the delta method and application to cancer genomic data

Let $\bar{x}$ and $\bar{y}$ be the sample means, and $s_x^2 = (n_x - 1)^{-1} \sum_i (x_i - \bar{x})^2$ and $s_y^2 = (n_y - 1)^{-1} \sum_i (y_i - \bar{y})^2$ be the sample estimates of variance. By the central limit theorem, for n_x, n_y sufficiently large, and assuming independence between samples,

$$\begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix} \sim N\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \sigma_x^2/n_x & 0 \\ 0 & \sigma_y^2/n_y \end{bmatrix} \right).$$

Let $g(\bar{x}, \bar{y}) = (\bar{x}/\bar{y})$. Then $\nabla g = (1/\bar{y}, -\bar{x}/\bar{y}^2)'$, and by the delta method $\bar{x}/\bar{y} \to N(\theta, \tau_1^2)$, where $\theta = g(\mu_x, \mu_y) = \mu_x/\mu_y$ and

$$\tau_1^2 = \nabla g^T(\mu_x, \mu_y) \begin{bmatrix} \sigma_x^2/n_x & 0 \\ 0 & \sigma_y^2/n_y \end{bmatrix} \nabla g(\mu_x, \mu_y) = \frac{\sigma_x^2}{n_x} \frac{1}{\mu_y^2} + \frac{\sigma_y^2}{n_y} \frac{\mu_x^2}{\mu_y^4}.$$

Using unbiased estimates for the variance, we get

$$\hat{\tau_1}^2 = \frac{s_x^2}{n_x \bar{y}^2} + \frac{s_y^2 \bar{x}^2}{n_y \bar{y}^4}$$

where s_x^2 and s_y^2 are the sample variances for $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. Similarly, we estimate the variance of $\bar{y}/\bar{x}$ as

$$\hat{\tau_2}^2 = \frac{s_y^2}{n_y \bar{x}^2} + \frac{s_x^2 \bar{y}^2}{n_x \bar{x}^4}.$$

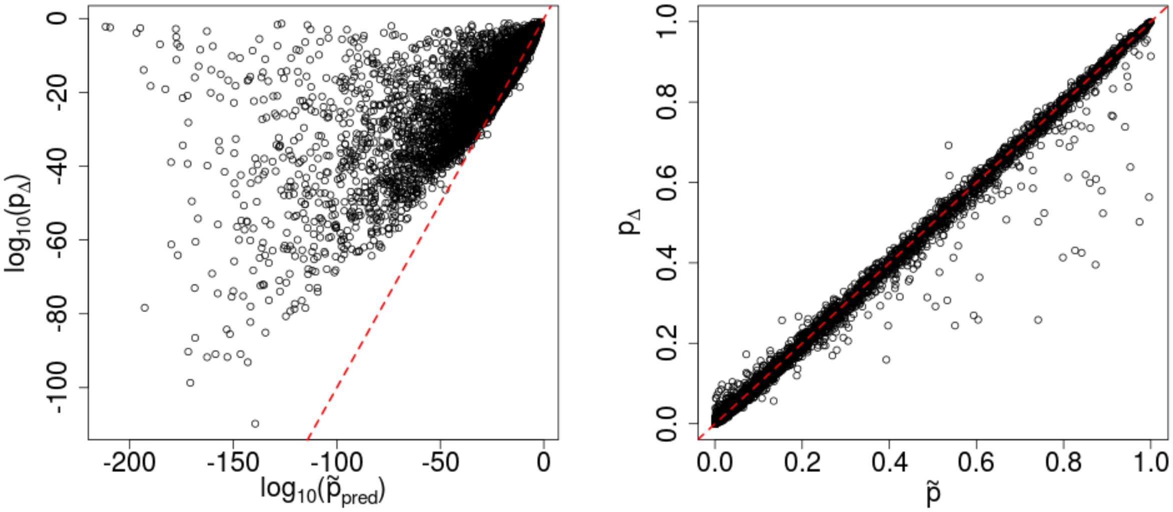
Therefore, to test the null $H_0 : \mu_x/\mu_y = 1$ versus the alternative $H_1 : \mu_x/\mu_y \neq 1$, the two-sided p-value using the delta method and unbiased estimates of variance is

$$p_\Delta = \begin{cases} \Pr(Z > \bar{x}/\bar{y}) + \Pr(U \leq \bar{y}/\bar{x}), & \bar{x}/\bar{y} \geq 1 \\ \Pr(U > \bar{y}/\bar{x}) + \Pr(Z \leq \bar{x}/\bar{y}), & \bar{x}/\bar{y} < 1 \end{cases},$$

where $Z \sim N(1, \hat{\tau_1}^2)$ and $U \sim N(1, \hat{\tau_2}^2)$. We use the Δ subscript in p_Δ to emphasize that the p-value is from the delta method. We note that p_Δ is potentially problematic, particularly if $\hat{\tau_1}^2$ or $\hat{\tau_2}^2$ are large, because the ratio is bounded below by zero, but the normal distribution is not.

We note that by allowing for unequal variance, we are testing a different null hypothesis than with the permutation test ($H_0 : P_x = P_y$). However, we expect that in practice, researchers would allow for unequal variance when using the delta method, which is why we use it as a basis for comparison. This comparison puts the permutation test at a disadvantage, but as shown in the simulations, the permutation test still performs better than the delta method.

Figure S27 compares estimates of the permutation p-values from our resampling algorithm ($\tilde{p}_{\text{pred}}$) to p_Δ for the cancer genomic data in Section 6. The dashed lines have an intercept of zero and slope of one, and indicate agreement. As seen in Figure S27, p_Δ tends to be an overestimate for small p-values, which is the same trend observed in the simulations. Out of the 100 genes with the smallest p_Δ, only three were identified by Zhan et al. (2015) as strongly distinguishing between LUAD and LUSC (*PVRL1*, *PERP*, and *ATP1B3*).



(a) Genes with $\tilde{p} \leq 1 \times 10^{-3}$ (10,302 genes)

(b) Genes with $\tilde{p} > 1 \times 10^{-3}$ (5,084 genes)

Figure S27: p-values for cancer genomic data: Comparison of results with the delta method (p_Δ) and our resampling algorithm ($\tilde{p}_{\text{pred}}$) with $B_{\text{pred}} = 10^3$ resamples within each partition, or with simple Monte Carlo ($\tilde{p}$) with a total of $B = 10^3$ resamples (see Section 6). The diagonal dashed lines have a slope of 1 and an intercept of 0, and indicate agreement between the methods. This figure appears in color in the electronic version of this article.

References

Becker, M. and Klößner, S. (2016). *PearsonDS: Pearson Distribution System.* R package version 0.98.

Butler, R. W. (2007). *Saddlepoint approximations with applications*, volume 22. Cambridge University Press.

Chung, E., Romano, J. P., et al. (2013). Exact and asymptotically robust permutation tests. *The Annals of Statistics* **41,** 484–507.

Hájek, J. (1961). Some extensions of the Wald-Wolfowitz-Noether theorem. *The Annals of Mathematical Statistics* pages 506–523.

Janssen, A. (1997). Studentized permutation tests for non-iid hypotheses and the generalized Behrens-Fisher problem. *Statistics & probability letters* **36,** 9–21.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). *Continuous univariate distributions*, volume 2. Wiley Series in Probability and Mathematical Statistics, 2nd edition.

Klar, B. (2015). A note on gamma difference distributions. *Journal of Statistical Computation and Simulation* **85,** 3708–3715.

Leemis, L. M. and McQueston, J. T. (2008). Univariate distribution relationships. *The American Statistician* **62,** 45–53.

Lehman, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks.* Holden-Day.

Lehmann, E. L. (1999). *Elements of large-sample theory.* Springer Science & Business Media.

Lugannani, R. and Rice, S. (1980). Saddle point approximation for the distribution of the sum of independent random variables. *Advances in applied probability* **12,** 475–490.

Mathai, A. (1993). On noncentral generalized laplacianness of quadratic forms in normal variables. *Journal of Multivariate Analysis* **45,** 239–246.

Neuhaus, G. (1993). Conditional rank tests for the two-sample problem under random censorship. *The Annals of Statistics* pages 1760–1779.

Robinson, J. (1982). Saddlepoint approximations for permutation tests and confidence intervals. *Journal of the Royal Statistical Society. Series B (Methodological)* pages 91–101.

Yu, K., Liang, F., Ciampa, J., and Chatterjee, N. (2011). Efficient p-value evaluation for resampling-based tests. *Biostatistics* pages 1–11.

Zhan, C., Yan, L., Wang, L., Sun, Y., Wang, X., Lin, Z., Zhang, Y., Shi, Y., Jiang, W., and Wang, Q. (2015). Identification of immunohistochemical markers for distinguishing lung adenocarcinoma from squamous cell carcinoma. *Journal of Thoracic Disease* **7,** 1398–1405.

Zhou, Y.-H. (2014). *MCC: Moment Corrected Correlation.* R package version 1.0.

Zhou, Y.-H. and Wright, F. A. (2015). Hypothesis testing at the extremes: fast and robust association for high-throughput data. *Biostatistics* pages 1–15.