A Deeper Look at the Permutation Test: Theory, Simulation, and Consistency

1 The permutation test explained

Suppose we have collected two data samples: a set of m iid observations from group A, and a set of n iid observations from another group, B. The sample values, $x_1, x_2, ..., x_m$ and $y_1, y_2, ..., y_n$, are observations of random variables X_A and Y_B having respective distributions F_X and F_Y. We want to determine whether these $m + n$ data points actually come from the *same* distribution. Formally, we consider the hypotheses

$$H_0 : \ F_X = F_Y$$
$$H_1 : \ F_X \neq F_Y$$

Suppose the nature of the two distributions is unknown to us. We certainly don't want our hypothesis test to hinge on dubious assumptions about the nature of the unknown distributions F_X and F_Y. How can we use our sample to make a population-level inference *without* making assumptions about the two distributions?

Our sample consists of $N = n + m$ observations. We pool this sample, and use the notation z_i, $1 \leq i \leq N$, for elements of the pooled sample. We will assume throughout that the pooled sample, $S = \{z_1, ..., z_N\}$, is an independent sample, which means we are now making the additional assumption that our two samples are independent of one another. As a consequence, under H_0, S is in fact iid, making S *exchangeable*: every permutation of the values of the random variable have the same probability. That is, if $\pi : [1, N] \rightarrow [1, N]$ is any permutation of the indices, then letting G be the joint CDF of the random variables Z_i,

$$G_{Z_1,...,Z_N}(z_1, ..., z_N) = G_{Z_{\pi(1)},...,Z_{\pi(N)}}(z_1, ..., z_N).$$

The proof that iid implies exchangeability follows immediately from the fact that the joint density of the iid sample factors as the product of the marginal densities, together with the fact that multiplication of real numbers is commutative.

Exchangeability suggests the following test. We choose a test statistic, T, for comparing F_X and F_Y; the difference of the sample means is a common choice. We then enumerate each of the $\binom{N}{n}$ permutations of the group labels among the fixed values z_i. Under H_0, the set of all such permutations has the discrete uniform distribution. For each permutation π, we compute the corresponding value of the test statistic, T_π. Under H_0, T satisfies

$$\mathbb{P}(T = t) = \frac{\#\{\text{permutations}, \ \ \pi : T_\pi = t\}}{\binom{N}{n}}$$

Provided only that we are able to enumerate all permutations, we thus specify the exact distribution of T under the null, called the *permutation distribution*. Notice that, by conditioning on the sample (i.e., holding the values z_i fixed and temporarily forgetting the distributions that generated these values) and building the null distribution for T on the discrete uniform distribution of permutations of group labels, we have ensured that the distribution of T assumes no knowledge of the distributions of F_X and F_Y—not their shape, moments, etc.[1] T owes its distribution to H_0 and the consequent exchangeability of the random variables Z_i: the stage is set for a non-parametric test.

If T^* is the observed value of T (calculated without permuting labels), then for a two-sided hypothesis test, the p-value is

$$p\text{ - value} = \frac{\#\{\pi : |T_\pi| \geq |T^*|\}}{\binom{N}{n}}.$$

For an α-level test, finding that $p < \alpha$ is evidence that our sample is *not* in fact exchangeable. Since we have assumed independence, we are led to reject H_0. Together, independence of the pooled sample and H_0 mean that the pooled sample is iid, and thus exchangeable:

$$\text{independence} \quad + \quad \text{common distribution} \quad \Longrightarrow \quad \text{exchangeability}$$

Technically, a small p-value is evidence against exchangeability. Since the independence assumption is not suspect, we are led to reject the hypothesis H_0 of a common distribution.

This is called a *permutation test*. Part of its allure is *exactness*: under H_0, we are in theory able to construct the exact discrete distribution of the test statistic; this is done below in power simulations for small samples (Figure 1). If our samples are representative of the distributions F_X and F_Y, the inference from lack of exchangeability within the sample to lack of distributional inequality is reasonable. However, representativeness may require fairly large samples, in which case enumerating all permutations becomes infeasible. In practice, we often randomly choose (with replacement) a large number of the $\binom{N}{n}$ possible permutations, calculate the associated test statistic value for each permutation, and *estimate* the distribution of the test statistic under H_0 using the resulting values.

2 Power simulation

Figure 1 illustrates power simulations using normal samples from $\mathrm{N}(0, 1)$ and $\mathrm{N}(\mu_i, 1)$, where μ_i is the mean difference (effect size).[2] Each power estimation is based on 1000 tests. As expected, power appears to be an increasing function of both effect size and sample size.

Notice that for both the orange and the red power curves, we have $n+m = 14$. However, power appears to be nearly uniformly greater for the orange graph, when $n = m$. To see why, consider the test statistic,

$$\overline{X}_A - \overline{Y}_B.$$

Assuming equal variance, we have

$$\mathrm{Var}(\overline{X}_A - \overline{Y}_B) = \left(\frac{1}{n} + \frac{1}{m}\right)\sigma^2.$$

[1] Stated differently, had the same sample values arisen from a *different* pair of distributions, F'_X and F'_Y, the distribution of T would remain unchanged.

[2] The normal distribution was chosen for convenience. To better appreciate the workings of the test, power simulation function was built from scratch, and is included in Section 4.

Consider the functions $f, g : \mathbb{R}_+^2 \to \mathbb{R}$ defined by

$$f\begin{pmatrix} t_1 \\ t_2 \end{pmatrix} = \frac{1}{t_1} + \frac{1}{t_2},$$
$$g\begin{pmatrix} t_1 \\ t_2 \end{pmatrix} = t_1 + t_2.$$

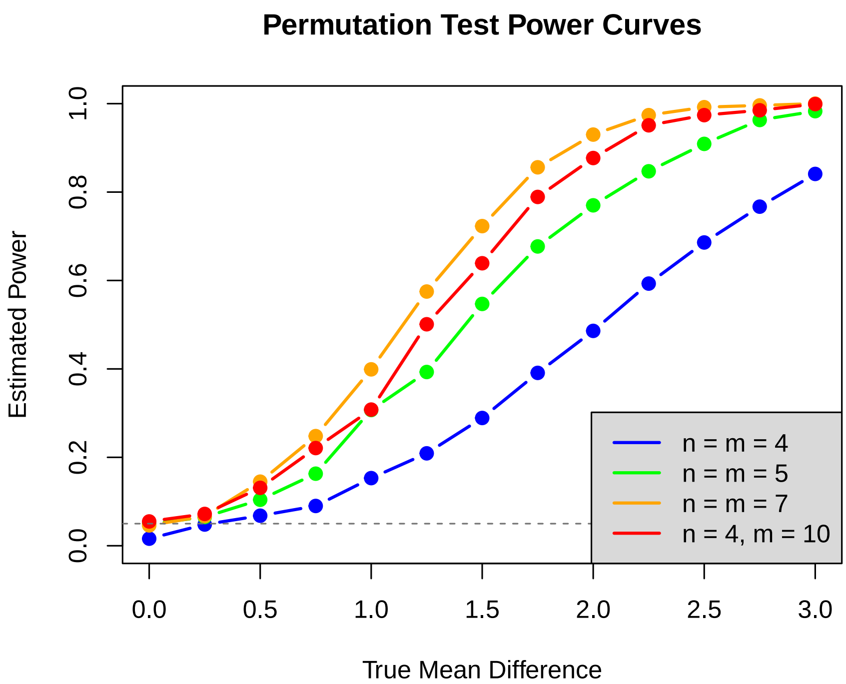

Figure 1: Power simulations for various sample sizes and effect sizes

To solve the constrained optimization problem

$$\min \ f(\mathbf{t}) \ \text{ subject to } \ g(\mathbf{t}) = C,$$

we use Lagrange multipliers:

$$Df(\mathbf{t}) = \lambda Dg(\mathbf{t})$$
$$\begin{bmatrix} -t_1^{-2} & -t_2^{-2} \end{bmatrix} = \lambda \begin{bmatrix} 1 & 1 \end{bmatrix}.$$

Hence, $t_1^2 = t_2^2$, since both are equal to $-\frac{1}{\lambda}$. Since both components are positive by assumption, we get $t_1 = t_2 = \frac{C}{2}$. To see that this extremum is a local minimum, consider that the the Hessian matrix at every point $t_1, \ t_2 > 0$ is positive definite:

$$\mathcal{H}(\mathbf{t}) = 2\begin{bmatrix} t_1^{-3} & 0 \\ 0 & t_2^{-3} \end{bmatrix}, \quad t_1, \ t_2 > 0.$$

The same is true in the discrete case, when t_1 and t_2 are both required to be positive integers. Thus, variance of the test statistic is minimized when we have a balanced sample, with $n = m$. Smaller variance means that, for a fixed value of $N = n + m$, our observed test statistic is a more precise estimator of the true *non-zero* effect size (under H_1); since the permutation distribution of the test statistic is symmetric about zero when $n = m$, this can be expected to yield a more extreme test statistic, and thus higher power. Of course, this is all immediately evident from the fact that that f is symmetric with respect to its two arguments.

3 Consistency of the permutation test

Power simulation suggests that as sample sizes n and m grow without bound, the power of the test approaches 1. The proof of this fact reveals what is happening under the hood of the permutation test.

Suppose F_X and F_Y are any two distinct distributions with finite variances σ_1^2 and σ_2^2 (we are not here assuming common variance). Suppose, as well, that the means of F_X and F_Y, μ_X and μ_Y, are not equal, so that the test statistic $T = \overline{X}_A - \overline{Y}_B$ can, with large enough sample sizes, detect a difference between the two distributions. We want to prove that, given $\alpha \in (0,1)$, given the non-zero difference $\mu_X - \mu_Y$ and given $\delta > 0$, there exists a sample size, M, such that $n, m \geq M$ implies that the power of the α-level permutation test will be at least $1 - \delta$.

The intuition is simply this: T is converging in probability to the non-zero difference $\Delta := \mu_X - \mu_Y$, while the probability of the permutation distribution under H_0 becomes increasingly concentrated around its zero mean. For large enough sample sizes, T should, with high probability, lie near Δ, which eventually lies beyond most of the probability mass of the permutation distribution under H_0.

Before seeing the full argument, consider the following condensed argument. Given $\delta > 0$ and $\epsilon = \frac{|\Delta|}{2}$, suppose there exists $N_1 > 0$ such that $n, m \geq N_1$ implies

$$\mathrm{P}(|T - \Delta| \leq \epsilon) \geq 1 - \delta, \tag{1}$$

so that $\mathrm{P}(|T| > 2^{-1}|\Delta|) \geq 1 - \delta$.[3] Suppose there also exists $N_2 > 0$ such that $n, m \geq N_2$ implies

$$\mathrm{P}_{\mathrm{perm}}(|T_\pi| > \epsilon) < \alpha, \tag{2}$$

where P_{perm} is the permutation distribution under H_0. Let $M := \max(N_1, N_2)$. If $n.m \geq N$, then with probability at least $1 - \delta$ we have $|T| > 2^{-1}|\Delta|$ and we *also* have

$$\begin{aligned} \mathrm{P}(|T_\pi| \geq |T|) &< \mathrm{P}(|T_\pi| > \epsilon) \\ &< \alpha. \end{aligned}$$

That is, there exists $M > 0$ such that we reject the null at the α level with probability no less than $1 - \delta$ for sample sizes $n, m \geq M$; the power of our test is thus no less than $1 - \delta$. All that remains is to show that there exist N_1 and N_2 such that Equation 1 and Equation 2 hold.

Equation 1 is the simpler of the two. By the Weak Law of Large Numbers, we have

$$\overline{X}_A \xrightarrow{P} \mu_X \quad \text{and} \quad \overline{Y}_B \xrightarrow{P} \mu_Y,$$

so that

$$\overline{X}_A - \overline{Y}_B \xrightarrow{P} \mu_X - \mu_Y.$$

[3]This follows from the fact that $|T - \Delta| \leq \epsilon \implies |T| \geq |\Delta| - \epsilon = \frac{|\Delta|}{2}$.

Therefore, given $\delta > 0$ and $\epsilon > 0$, there exists $N_1 > 0$ such that $n, m \geq N_1$ implies

$$\mathrm{P}\Big(|T - \Delta| < \epsilon\Big) > 1 - \delta, \tag{3}$$

which is Equation 1.

For Equation 2, we first find the mean and variance of the permutation distribution, which ultimately depends on the mean and variance of our finite population, S. Call the mean $\overline{z}$. The permuted groups are created by sampling, at random and without replacement, n values from the pooled sample for group A; the other m values are assigned to the group B permuted sample. The random variable $T_\pi = \overline{X}_A - \overline{Y}_B$ has expected value

$$\mathbb{E}_{\text{perm}}(T_\pi) = \mathbb{E}_{\text{perm}}(\overline{X}_A) - \mathbb{E}_{\text{perm}}(\overline{Y}_B).$$

Consider $\overline{Y}_B$. This random variable is the average of m randomly chosen values Y_{B1}, $Y_{B2}, \ldots,$ Y_{Bm}. (The random selection of n values forming the random variable $\overline{X}_A$ induces a random selection of m values forming the random variable $\overline{Y}_B$, and conversely.) Each Y_{Bk} has mean

$$\begin{aligned}
\mathbb{E}_{\text{perm}}(Y_{Bk}) &= \sum_{i=1}^{N} z_i \mathrm{P}_{\text{perm}}(Y_{Bk} = z_i) \\
&= \sum_{i=1}^{N} \left(z_i \cdot \frac{1}{N}\right) \\
&= \overline{z}.
\end{aligned}$$

By linearity, we have

$$\begin{aligned}
\mathbb{E}_{\text{perm}}(\overline{Y}_B) &= \frac{1}{m} \sum_{k=1}^{m} \mathbb{E}_{\text{perm}}(Y_{Bk}) \\
&= \overline{z}.
\end{aligned}$$

An analogous argument shows that $\mathbb{E}_{\text{perm}}(\overline{X}_A) = \overline{z}$, as well. Thus, $\mathbb{E}_{\text{perm}}(T_\pi) = 0$, and this mean is independent of the relative group sizes n and m.

The variance, σ_z^2, of this finite population is

$$\sigma_z^2 = \frac{1}{N-1} \sum_{i=1}^{N} (z_i - \overline{z})^2.$$

Because the overall mean $\overline{z}$ is fixed, we can express T_π purely in terms of $\overline{X}_A$. That is, $n\overline{X}_A + m\overline{Y}_B = N\overline{z}$ implies

$$\overline{Y}_B = \frac{N\overline{z} - n\overline{X}_A}{m}.$$

Thus,

$$\begin{aligned}
T_\pi &= \overline{X}_A - \overline{Y}_B \\
&= \frac{m\overline{X}_A - (N\overline{z} - n\overline{X}_A)}{m} \\
&= \frac{N}{m}(\overline{X}_A - \overline{z}).
\end{aligned}$$

Hence,

$$\begin{aligned}\mathrm{Var}_{\mathrm{perm}}(T_\pi) &= \left(\frac{N}{m}\right)^2 \mathrm{Var}_{\mathrm{perm}}(\overline{X}_A) \\ &= \left(\frac{N}{m}\right)^2 \cdot \sigma_z^2 \frac{N-n}{n(N-1)} \qquad (*) \\ &= \frac{N^2\sigma_z^2}{mn(N-1)},\end{aligned}$$

where the known formula for the sample mean when drawing without replacement from a finite population is used in line $(*)$.

We see that $\mathrm{Var}_{\mathrm{perm}}(T_\pi) \to 0$ as $n, m \to \infty$. Given $\epsilon = \frac{|\Delta|}{2}$ and $\alpha \in (0, 1)$, let

$$N_2 > \left(\frac{N^2\sigma_z^2}{(N-1)\epsilon^2\alpha}\right)^{\frac{1}{2}}.$$

Then $n, m \geq N_2$ implies that

$$\underbrace{\mathrm{P}(|T_\pi| \geq \epsilon) < \frac{\mathrm{Var}_{\mathrm{perm}}(T_\pi)}{\epsilon^2}}_{\text{Chebyshev's Inequality}} < \alpha,$$

and Equation 2 above holds. We have therefore proved the following theorem:

Theorem (Consistency of permutation test with mean difference): Let F_X and F_Y be two distributions having finite variance and different means μ_X and μ_Y, and suppose we have iid samples of sizes n and m drawn from F_X and F_Y, respectively. For any $\delta > 0$, there exists $N > 0$ such that if $n, m \geq N$, then the α-level permutation test for comparing hypotheses

$$\begin{aligned} H_0 &: \; F_X = F_Y \\ H_1 &: \; F_X \neq F_Y \end{aligned}$$

based on the test statistic of sample mean difference has power greater than $1 - \delta$.

The theorem is not purely of theoretical interest. Its proof has us compare the asymptotic behavior of T with that of the permutation distribution, and we learn something about each in the process. The proof also emphasizes the subtle way that comparative parametric information about F_X and F_Y is encapsulated in the observed value of the test statistic T^*—information which the non-parametric permutation test does not forget or ignore. True, the permutation distribution is built on the set of all $\binom{N}{n}$ group relabelings: under H_0, each of these group relabelings is equally likely to have been drawn from F_X and F_Y, so that this set of relabelings has the discrete uniform distribution. In a sense, this discrete uniform distribution acts as a veil which separates the resulting permutation distribution of T_π from any prior knowledge we may have had about F_X or F_Y.[4] That prior knowledge *is* retained, however: it is compressed into the single value T^*, so that by this one narrow window we may yet distinguish the two distributions. This explains why we *must* begin with informative, representative samples of F_X and F_Y: as the proof makes clear, the power of the test is a consequence of the precision with which each sample estimates its population mean.

[4] We begin with F_X and F_Y, and sample from each. The discrete uniform distribution, with its "every point receives equal weight" logic, effectively erases insights about F_X and F_Y held in the samples, and we emerge on the other side of the discrete uniform having forgotten all we knew about F_X and F_Y—*except* for the information smuggled in the test statistic.

4 R code

```
### Creating Figure 1 ###

## Power simulation for permutation test ##

# Normally distributed data, difference of means as test stat

simulate_power <- function(n, m,
                           alpha = 0.05,
                           effect_sizes = seq(0, 3, by = 0.25),
                           reps = 1000) {
  N <- n + m
  M <- combn(1:N, n)
  power_curve <- numeric(length(effect_sizes))

  for (i in seq_along(effect_sizes)) {
    mean_diff <- effect_sizes[i]
    num_reject <- 0

    for (j in 1:reps) {
      x <- rnorm(n, mean = 0, sd = 1)
      y <- rnorm(m, mean = mean_diff, sd = 1)
      z <- c(x, y)

      test_obs <- mean(x) - mean(y)
      perm_dist <- apply(M, 2, function(A_idx){
        B_idx <- setdiff(1:N, A_idx)
        mean(z[A_idx]) - mean(z[B_idx])
      })

      p_value <- mean(abs(perm_dist) >= abs(test_obs))
      if (p_value < alpha) {
        num_reject <- num_reject + 1
      }
    }

    power_curve[i] <- num_reject / reps
  }

  return(power_curve)
}

set.seed(2025)
effect_sizes <- seq(0, 3, by = 0.25)

powers <- list(
  k3 = simulate_power(4, 4, effect_sizes = effect_sizes),
```

```r
  k4 = simulate_power(5, 5, effect_sizes = effect_sizes),
  k5 = simulate_power(7, 7, effect_sizes = effect_sizes),
  k6 = simulate_power(4, 10, effect_sizes = effect_sizes)
)

colors <- c("blue", "green", "orange", "red")
names(colors) <- names(powers)

pdf("C:/Users/17402/Desktop/KCU/Fall2025/IntroStats/permtestpower.pdf",
    width = 6,
    height = 5)

# Base plot
plot(effect_sizes,
     powers$k3,
     type = "b",
     col = colors["k3"],
     lwd = 2,
     pch = 19,
     ylim = c(0,1),
     xlab = "True Mean Difference",
     ylab = "Estimated Power",
     main = "Permutation Test Power Curves")
lines(effect_sizes,
      powers$k4,
      type = "b",
      col = colors["k4"],
      lwd = 2,
      pch = 19)
lines(effect_sizes,
      powers$k5,
      type = "b",
      col = colors["k5"],
      lwd = 2,
      pch = 19)
lines(effect_sizes,
      powers$k6,
      type = "b",
      col = colors["k6"],
      lwd = 2,
      pch = 19)

abline(h = 0.05, col = "gray45", lty = 2)

legend("bottomright", legend = c("n = m = 4",
                                 "n = m = 5",
                                 "n = m = 7",
                                 "n = 4, m = 10"),
       col = colors,
```

```
         lty = 1,
         lwd = 2,
         bg = "gray85")

dev.off()
```