# Worked Examples of the Five Reference Distributions in Nonparametric Hypothesis Testing

*A Companion to "Approaches to Nonparametric Statistical Analysis"*

## Introduction

The companion paper "Approaches to Nonparametric Statistical Analysis" introduced five probability distributions that serve as reference distributions for nonparametric hypothesis tests: the chi-square distribution, the F-distribution, the standard normal distribution (used as a large-sample approximation), the discrete uniform distribution of ranks, and the binomial distribution. This paper works through one complete, fully-computed numerical example for each distribution — real data, ranks or signs computed by hand, the test statistic built up step by step, the relevant formula shown in general form and then with the actual numbers substituted in, and a graph showing exactly where the resulting statistic falls relative to the reference distribution and the rejection region.

Each example follows the same four-part structure: (1) a short scenario and dataset, (2) the calculation of the test statistic, (3) the general formula alongside the same formula with the computed numbers plugged in, and (4) a decision based on comparing the statistic to a critical value at α = 0.05, illustrated graphically.

## Example 1: Chi-Square Distribution — Kruskal-Wallis H Test

Scenario: A researcher compares three tutoring methods (A, B, C) by recording the exam scores of five students taught under each method. Because the sample sizes are small and scores are not assumed to be normally distributed, the Kruskal-Wallis H test — whose test statistic follows a chi-square distribution — is used instead of a one-way ANOVA.

### Step 1: The Data

| Student | Method A | Method B | Method C |
|---|---|---|---|
| 1 | 78 | 88 | 65 |
| 2 | 82 | 95 | 70 |
| 3 | 85 | 90 | 72 |
| 4 | 74 | 92 | 68 |
| 5 | 91 | 85 | 75 |

### Step 2: Rank All 15 Scores Together and Sum Ranks by Group

| Group | Ranks (within combined sample of 15) | Rank Sum (Rj) |
|---|---|---|
| Method A | 7.0, 8.0, 9.5, 5.0, 13.0 | 42.5 |
| Method B | 11.0, 15.0, 12.0, 14.0, 9.5 | 61.5 |
| Method C | 1.0, 3.0, 4.0, 2.0, 6.0 | 16.0 |

Tied scores are not an issue here — all 15 exam scores are distinct, so ranks 1 through 15 are assigned without any ties.

### Step 3: Compute the H Statistic

With N = 15 total observations, k = 3 groups, and nj = 5 observations per group:

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(N+1)$$

$$H = \frac{12}{15(16)} \left( \frac{42.5^2}{5} + \frac{61.5^2}{5} + \frac{16^2}{5} \right) - 3(16) = 10.445$$

## Step 4: Compare to the Chi-Square Distribution

The H statistic is compared to a chi-square distribution with k − 1 = 2 degrees of freedom. At α = 0.05, the critical value is $\chi^2(0.05, 2) = 5.991$. The computed H = 10.445 exceeds this critical value, corresponding to a p-value of approximately 0.0054.

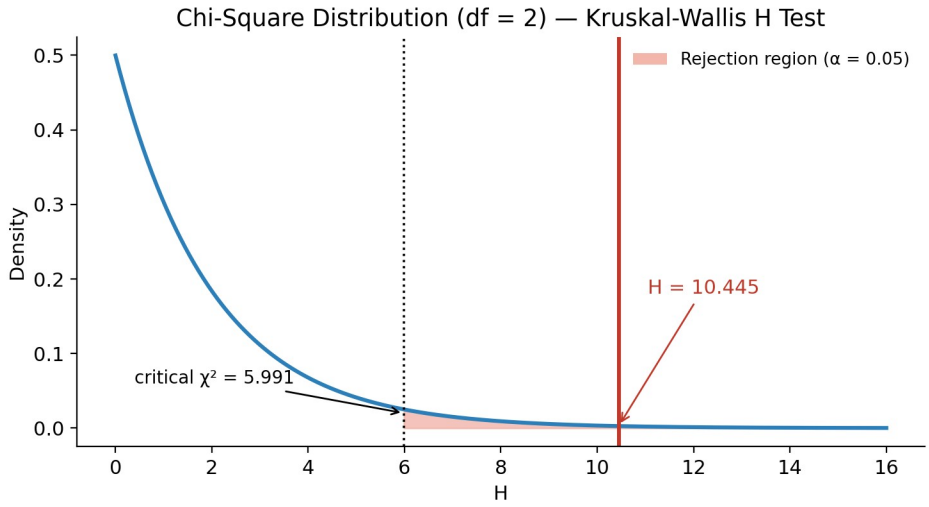


*Figure 1. The computed H = 10.445 falls well inside the rejection region of the $\chi^2$(df = 2) distribution, so H0 (equal distributions across methods) is rejected at α = 0.05.*

Conclusion: There is a statistically significant difference in exam scores among the three tutoring methods (H = 10.445, df = 2, p ≈ 0.005).

# Example 2: F-Distribution — Two-Sample Variance Ratio Test

Scenario: A quality-control engineer wants to know whether two manufacturing lines produce items with equally consistent weight, or whether one line is more variable than the other. Eight items are sampled from each line and weighed in grams.

## Step 1: The Data

| Item | Line 1 (g) | Line 2 (g) |
|---|---|---|
| 1 | 50.1 | 50.0 |
| 2 | 49.8 | 51.2 |
| 3 | 50.3 | 48.9 |
| 4 | 49.9 | 50.5 |
| 5 | 50.2 | 49.3 |
| 6 | 50.0 | 51.0 |
| 7 | 49.7 | 49.0 |
| 8 | 50.4 | 50.8 |

## Step 2: Compute Each Sample Variance

Sample variance (n − 1 in the denominator): Line 1 variance $s1^2 = 0.0600$; Line 2 variance $s2^2 = 0.8527$. Because Line 2 has the larger variance, it is placed in the numerator so that F ≥ 1, which simplifies comparison to the standard upper-tail critical value.

## Step 3: Compute the F Statistic

$$F = \frac{s_1^2}{s_2^2}, \qquad d_1 = n_1 - 1, \;\; d_2 = n_2 - 1$$

$$F = \frac{0.8527}{0.0600} = 14.211$$

## Step 4: Compare to the F-Distribution

With d1 = n2 − 1 = 7 and d2 = n1 − 1 = 7 degrees of freedom, the two-tailed critical value at α = 0.05 (α/2 = 0.025 in each tail) is F(0.025, 7, 7) = 4.995. The computed F = 14.211 far exceeds this critical value, giving a two-tailed p-value of approximately 0.0024.

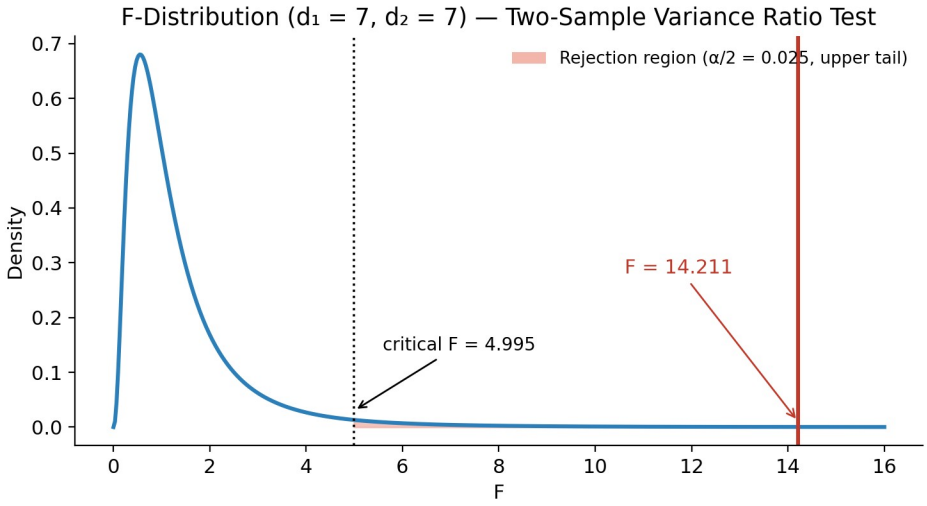


*Figure 2. The computed F = 14.211 falls deep in the upper rejection region of the F(7, 7) distribution.*

Conclusion: The two manufacturing lines do not produce items with equal variability — Line 2's weights are significantly more variable than Line 1's (F = 14.211, d1 = d2 = 7, p ≈ 0.002).

# Example 3: Standard Normal Distribution — Mann-Whitney U (z-Approximation)

Scenario: A researcher tests whether a new sleep medication increases hours of sleep compared to a placebo. Eight participants receive the drug and seven receive a placebo; nightly sleep hours are recorded. With combined sample size n1 + n2 = 15, the Mann-Whitney U statistic is large enough to use the normal approximation rather than exact tables.

## Step 1: The Data

| Group | Sleep Hours |
|---|---|
| Drug (n1 = 8) | 6.1, 7.0, 7.5, 6.8, 7.9, 8.2, 7.3, 8.0 |
| Placebo (n2 = 7) | 5.2, 5.8, 6.0, 5.5, 6.3, 5.9, 6.7 |

## Step 2: Rank All 15 Observations Together

| Group | Ranks | Rank Sum (R) |
|---|---|---|
| Drug | 6, 10, 12, 9, 13, 15, 11, 14 | 90 |
| Placebo | 1, 3, 5, 2, 7, 4, 8 | 30 |

## Step 3: Compute U

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U_1 = (8)(7) + \frac{8(9)}{2} - 90 = 2.0$$

Since U2 = n1n2 − U1 = 56 − 2.0 = 54.0, the smaller value U = min(U1, U2) = 2.0 is used as the test statistic.

## Step 4: Convert to a z-Score

$$z = \frac{U - \mu_U}{\sigma_U}, \quad \mu_U = \frac{n_1 n_2}{2}, \quad \sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

$$\mu_U = \frac{(8)(7)}{2} = 28.0, \quad \sigma_U = \sqrt{\frac{(8)(7)(16)}{12}} = 8.641, \quad z = \frac{2.0 - 28.0}{8.641} = -3.009$$

## Step 5: Compare to the Standard Normal Distribution

At α = 0.05 (two-tailed), the critical values are z = ±1.96. The computed z = -3.009 falls in the lower rejection region, corresponding to a two-tailed p-value of approximately 0.0026.

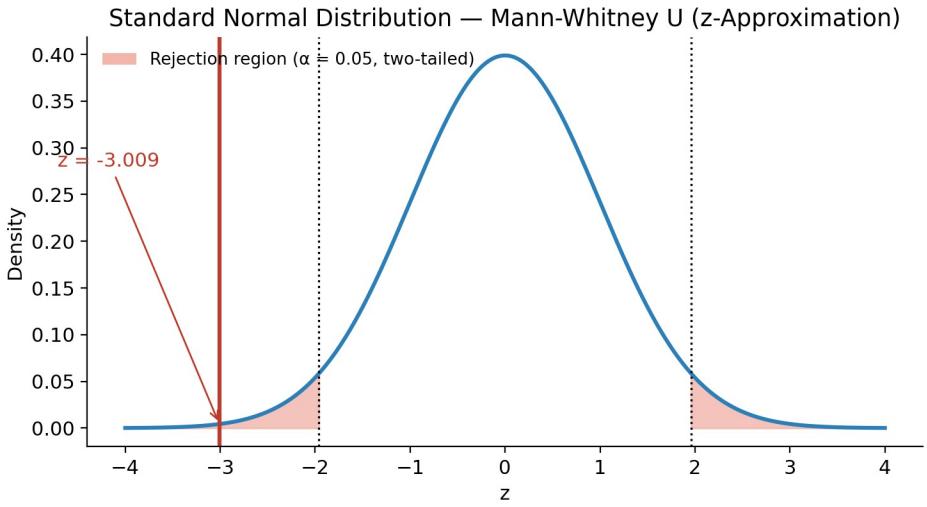


*Figure 3. The computed z = −3.009 falls in the lower rejection region of the standard normal distribution.*

Conclusion: Participants receiving the drug sleep significantly more hours than those receiving the placebo (U = 2.0, z = −3.009, p ≈ 0.003).

# Example 4: Discrete Uniform Distribution of Ranks — Exact Wilcoxon Rank-Sum Test

Scenario: A usability researcher measures reaction time (milliseconds) under two testing conditions with very small samples — three participants in Condition A and four in Condition B. Because the combined sample size (n = 7) is too small for the normal approximation to be reliable, the exact permutation distribution of the rank sum is used instead. This exact distribution is built directly from the fact that, under H0, every one of the C(7,3) = 35 ways of assigning ranks 1–7 to the three Condition-A observations is equally likely — the ranks are discrete-uniformly distributed over {1, …, 7}.

## Step 1: The Data and Ranks

| Condition | Reaction Times (ms) | Ranks (1–7) |
|---|---|---|
| A (n1 = 3) | 12, 15, 22 | 1, 2, 5 |
| B (n2 = 4) | 18, 20, 25, 28 | 3, 4, 6, 7 |

The observed rank sum for Condition A is R1 = 1 + 2 + 5 = 8.

## Step 2: The Discrete Uniform Basis of the Exact Test

$$R_i \sim \text{DiscreteUniform}\{1, 2, \ldots, n\} \text{ under } H_0, \qquad \binom{n}{n_1} \text{ equally likely arrangements}$$

Because every arrangement of ranks among the 7 positions is equally likely under H0, all 35 possible ways of choosing which 3 of the 7 ranks belong to Condition A are enumerated, and the rank sum R1 is computed for each. The result is the exact null distribution shown in Figure 4.

## Step 3: The Exact p-Value

$$\binom{7}{3} = 35 \text{ arrangements}, \quad R_1^{obs} = 8, \quad p_{\text{exact}} = \frac{8}{35} = 0.229$$

Counting all arrangements whose rank sum is at least as extreme (as far from the mean of 12.0) as the observed R1 = 8 gives 8 out of 35 arrangements, for an exact two-tailed p-value of 0.229.

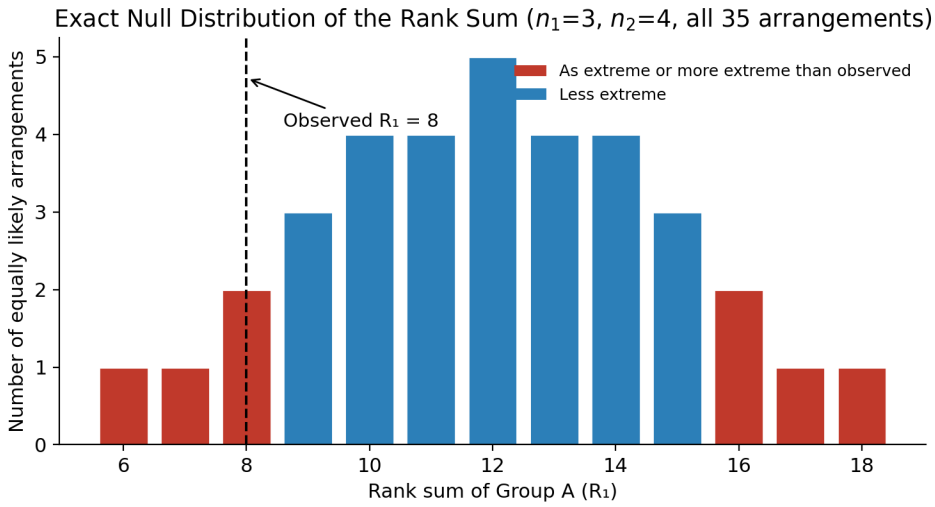


*Figure 4. The exact null distribution of the Condition-A rank sum, built from all 35 equally likely arrangements of ranks 1–7. Red bars mark outcomes as extreme or more extreme than the observed R1 = 8.*

Conclusion: With p = 0.229, which is greater than α = 0.05, there is not enough evidence at the 5% level to conclude that reaction times differ between the two conditions — though the result is suggestive given the very small sample size.

# Example 5: Binomial Distribution — Sign Test

Scenario: A company measures the productivity scores of 10 employees before and after adopting new software, to determine whether the software changed productivity. Because the differences are not assumed to be normally distributed and only the direction of each employee's change is considered, the Sign test is used.

## Step 1: The Data and Signs

| Employee | Before | After | Difference | Sign |
|---|---|---|---|---|
| 1 | 72 | 75 | 3 | + |
| 2 | 68 | 74 | 6 | + |
| 3 | 75 | 73 | -2 | – |
| 4 | 80 | 85 | 5 | + |
| 5 | 65 | 70 | 5 | + |
| 6 | 78 | 76 | -2 | – |
| 7 | 70 | 78 | 8 | + |
| 8 | 74 | 79 | 5 | + |

| Employee | Before | After | Difference | Sign |
|---|---|---|---|---|
| 9 | 69 | 68 | -1 | – |
| 10 | 77 | 82 | 5 | + |

Of the 10 employees, 7 showed an increase (+), 3 showed a decrease (−), and none showed no change. All n = 10 non-zero differences are used in the test.

## Step 2: The Binomial Null Distribution

Under H0 (the software has no effect, so an increase or decrease is equally likely), the number of positive signs among n non-zero differences follows a binomial distribution with p = 0.5:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad p = 0.5$$

## Step 3: Compute the Exact p-Value

$$P(X \geq 7 \text{ or } X \leq 3) = \sum_{k=7}^{10} \binom{10}{k} (0.5)^{10} + \sum_{k=0}^{3} \binom{10}{k} (0.5)^{10} = 0.3438$$

This two-tailed sum includes all outcomes at least as extreme as the observed 7 positive signs out of 10 (i.e., 7, 8, 9, or 10 positive signs, and their mirror-image 3, 2, 1, or 0 positive signs).

## Step 4: Decision

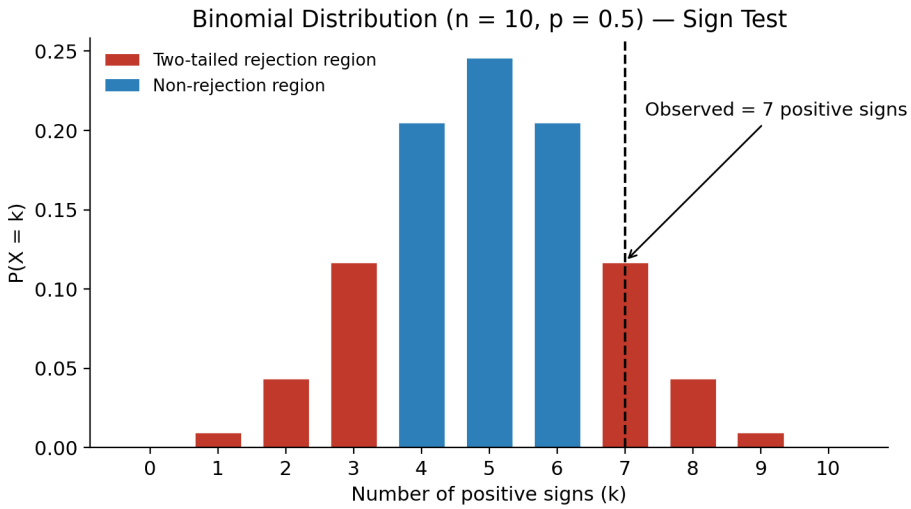


*Figure 5. The Binomial(n = 10, p = 0.5) distribution. Red bars mark the two-tailed rejection region; the observed 7 positive signs falls just inside it, but the total tail probability exceeds 0.05.*

Conclusion: With p = 0.3438, which is greater than α = 0.05, there is not enough evidence to conclude that the new software changed employee productivity, despite 7 of 10 employees showing an increase.

# Summary of All Five Examples

| Distribution | Test | Statistic | p-value | Decision at α = 0.05 |
|---|---|---|---|---|
| Chi-square (df = 2) | Kruskal-Wallis H | H = 10.445 | 0.0054 | Reject H0 |
| F (d1 = 7, d2 = 7) | Variance ratio test | F = 14.211 | 0.0024 | Reject H0 |
| Standard normal | Mann-Whitney U (z-approx.) | z = -3.009 | 0.0026 | Reject H0 |
| Discrete uniform (ranks) | Exact Wilcoxon rank-sum | R1 = 8 | 0.229 | Fail to reject H0 |
| Binomial (n = 10, p = 0.5) | Sign test | k = 7 | 0.3438 | Fail to reject H0 |

Together, these five worked examples show how each reference distribution connects a rank-, sign-, or variance-based test statistic computed from real data to a p-value: the chi-square and F-distributions arise as the limiting or exact distributions of sums of squared quantities, the standard normal arises as a large-sample approximation to a discrete rank-sum statistic, the discrete uniform distribution of ranks provides the exact combinatorial foundation for small-sample rank tests, and the binomial distribution governs the simplest sign-based test. In every case, the same logic applies: compute the statistic from the data, locate it on its reference distribution, and compare the resulting tail probability to the chosen significance level.