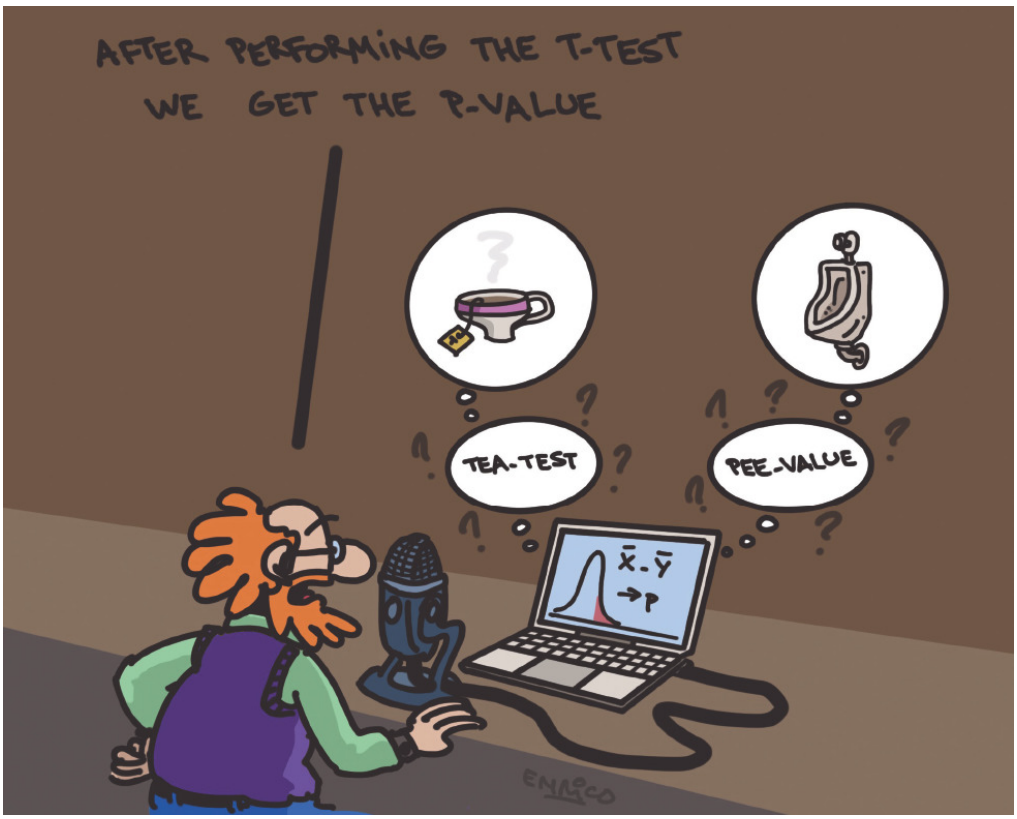
Testing Differences Between Two Independent Groups 01

EN5423 | Spring 2024

w07_two_independent_groups_01.pdf
(Week 7)

Contents



Testing Differences Between Two Independent Groups Intro

• Wells near a hazardous waste site are tested to see if wells downstream have higher levels of a toxic compound than those upstream, checking if there is a significant difference at a 0.01 significance level and if the cleanup cost is justified based on the difference.

• In another study, 16 streams are examined for biological diversity, comparing eight natural streams against eight affected by urban runoff, to assess if urban streams have lower biological quality.

• A third study investigates if bedrock fracturing impacts well yields in the Piedmont region by comparing yields between fractured and unfractured bedrock wells.

These examples illustrate comparisons between ***two unrelated groups of data to see if one group generally has higher values than the other***, without direct pairing between the groups' observations. They highlight situations without a natural pairing structure and emphasize that each group should represent different conditions, ensuring that observations and their populations are exclusive to each group.

This chapter focuses on various statistical tests—nonparametric, permutation, and parametric—to determine if there is a significant difference in the central tendency between two independent groups. It also covers graphical methods for presenting test results, estimating the ***difference's magnitude between groups***, and ***examining variations within the groups***. A summary of these test types and their applications is provided in table 1.

Table 1. Hypothesis test methods in this chapter and their characteristics. H_A is the alternative hypothesis, the signal to be found if it is present.

Objective (H_A)	Test	Class of test	Distributional assumption	Estimator of difference
Data values in one group are frequently higher than those in the other group	Wilcoxon rank-sum test	Nonparametric	None	Hodges-Lehmann estimate
One group has a higher mean	Two-sample *t*-test	Parametric	Normal distribution. Differences additive	Mean difference
	Two-sample permutation test	Permutation	Same distribution as in the other group	Mean difference
One group has higher variability	Fligner-Killeen	Nonparametric	None	Difference in median absolute distance from the median
	Levene's	Parametric	Normal distribution	Difference in group variance

1 The Rank-sum Test

• The rank-sum test, developed by Wilcoxon in 1945 and equivalent to Mann and Whitney's test from 1947, has several names, including Wilcoxon rank-sum test, Mann-Whitney test, Wilcoxon-Mann-Whitney rank-sum test, and Two-sample Wilcoxon test. The key points are whether it shows a significant difference between two groups at a specified significance level and the size of that difference.

1.1 Null and Alternate Hypotheses for the Rank-sum Test

• In its most general form, the rank-sum test is a test for whether one group tends to produce larger observations than the second group. It has as its null hypothesis:

$$H_0 \text{Prob}\,(x_i > y_i) = 0.5, i = 1,2,\dots,n; j = 1,2,\dots,m$$

where the x_i are from one group and the y_i are from a second group. In words, this states that the probability of an x value being higher than any given y value is one-half. The alternative hypothesis is one of three statements:

H_{A1}: Prob $(x_i > y_i) \neq 0.5$ (Two-sided test, x might be larger or smaller than y)
H_{A2}: Prob $(x_i > y_i) > 0.5$ (One-sided test, x is expected to be larger than y)
H_{A3}: Prob $(x_i > y_i) < 0.5$ (One-sided test, x is expected to be smaller than y)

The rank-sum test is commonly used to ***compare medians between groups***, especially when the groups have similar distribution shapes. However, its application is broader.

For instance, if the lower halves of two sites' concentration distributions are alike, but a contaminant raises the upper 40% of one site's concentrations, medians might not differ much. Yet, the rank-sum test could still detect a significant ***difference because it considers the entire distribution***, including how the upper 40% of concentrations at the contaminated site exceed those at the clean site.

1.2 Assumptions of the Rank-sum Test

There are three assumptions for the rank-sum test (Conover, 1999):

1) Data in both groups are random samples from their respective populations.
Example) Imagine a study comparing the heights of men and women. The men's group consists of randomly selected men from a city, and the women's group consists of randomly selected women from the same city. Each group is a random sample because every man or woman in the city had an equal chance of being selected for the study.

2) In addition to independence of data within each group, there is mutual independence between the two groups. For example, data from the same sampling unit (and certainly the exact same observations) should never be present in both groups.
Example) A research project investigates the effect of a new fertilizer on plant growth. Group 1 consists of plants grown without the fertilizer (control group), and Group 2 consists of plants grown with the fertilizer (treatment group). Each plant is grown in its own pot and soil, ensuring no plant's growth influences another's. Additionally, plants from the control group are kept

separate from those in the treatment group to prevent any possible cross-contamination of fertilizer, ensuring mutual independence.

3) The measurement scale is at least ordinal.

Example) A customer satisfaction survey categorizes responses into "Very Unsatisfied," "Unsatisfied," "Neutral," "Satisfied," and "Very Satisfied." These categories represent an ordinal scale because they have a specific order where "Very Unsatisfied" is less than "Unsatisfied," and so on, but the exact differences between each category are not defined. A study might compare two products by asking customers to rate their satisfaction with each product using this scale, aiming to determine which product has higher satisfaction levels.

• The rank-sum test does not require equal variances or specific distributions for the data, allowing for various distributions including normal, lognormal, and exponential, among others.

• It is used to check if one group generally has higher values than another, ***even if their distributions differ***. The test primarily aims to see if the two groups originate from the same population, focusing on whether they share the same median and other percentiles, or if they only differ by their central value.

• An example illustrates that even with different shapes and variabilities in distributions, as long as the data are transformed similarly (e.g., using logarithms), the rank-sum test can still be valid. This test is flexible, accommodating different types of data transformations like logarithms or square roots, and can identify multiplicative differences between groups.

• Unlike the *t*-test, the rank-sum test is useful in a wide range of scenarios because it is adaptable to various distributions and transformations.

1.3 Computation of the Rank-sum Test

• For sample sizes n and m where $n < m$, and xi , i =1, 2, ..., n and $y\,j$, j =1, 2, ..., m are the two data groups, compute the joint ranks R_k:

$R_k = 1$ to $(N = n + m)$, using average ranks in case of ties. Then the test statistic:
$W_{rs} = sum\ of\ ranks\ for\ the\ group\ having\ the\ smaller\ sample\ size, or$
$= \sum R_i\ \ from\ i = 1,2, \dots, n$ (using either group with equal sample sizes $n = m$)

• A one-sided or one-tailed alternative should be chosen when one group is expected to be higher or lower (but not both!) than the second group prior to observing the data. For example, y is a background site with lower concentrations expected than for a possibly higher-concentration site x. Determine the p-value associated with W_{rs}. Reject H_0 when $p < \alpha$.

• The code we used in the previous chapter provides the exact p-value for small to moderate sample sizes unless there are ties, in which case the large-sample approximation is provided.

Example 1
Precipitation quality was compared at sites with different land uses by Oltmann and Shulters (1989). Ten concentrations of organic plus ammonia nitrogen (NH4orgN) at each site are listed below, along with their group location as the variable "where". Note that three pairs of concentrations (at 0.7, 1.1, and 1.3 milligrams per liter [mg/L]) are tied, and so are assigned tied ranks equal to the average of their two individual ranks.

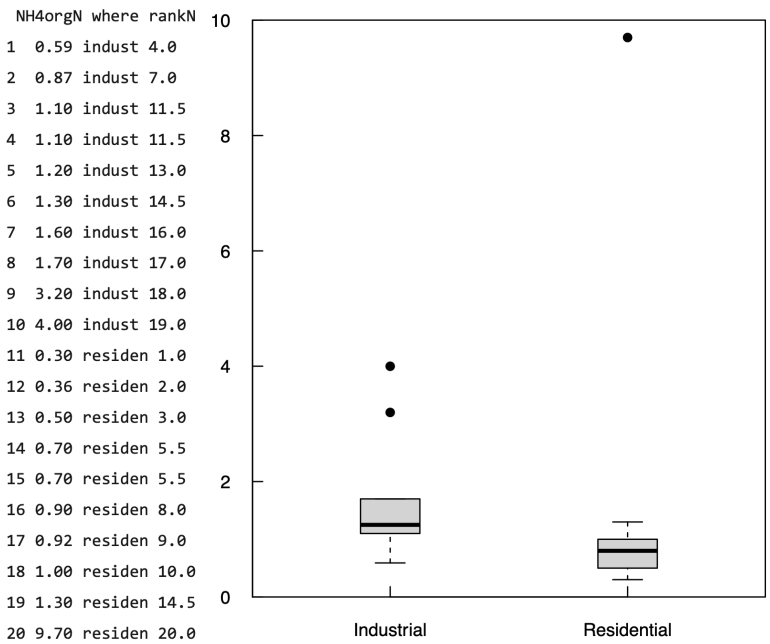


Figure 1. Boxplots of ammonia plus organic nitrogen. Data from Oltmann and Shulters (1989) by land use type: industrial or residential.

Median concentrations for the industrial and residential sites are 1.25 and 0.80 mg/L, respectively. The rank-sum test determines if ammonia plus organic nitrogen concentrations differ significantly ($\alpha = 0.05$) between the industrial and residential sites. The null (H_0) and alternate (H_A) hypotheses are:

H_0: Prob(concentration [industrial] ≥ concentration [residential]) = 0.5.
H_A: Prob(concentration [industrial] ≥ concentration [residential]) ≠ 0.5.

```
import numpy as np
import pandas as pd
from scipy import stats

def analyze_data_full(data1, data2, n_bootstrap=10000, ci=95,
ci_option='two-sided'):
    np.random.seed(42)  # Ensure reproducibility
    median_diffs = []

    # Bootstrap sampling for median differences
    for _ in range(n_bootstrap):
        sample1 = np.random.choice(data1, size=len(data1), replace=True)
        sample2 = np.random.choice(data2, size=len(data2), replace=True)
        median_diff = np.median(sample1) - np.median(sample2)
        median_diffs.append(median_diff)

    # Calculate confidence interval based on the selected option
    if ci_option == 'two-sided':
        lower_percentile = (100 - ci) / 2
        upper_percentile = 100 - lower_percentile
        confidence_interval = np.percentile(median_diffs,
[lower_percentile, upper_percentile])
```

```
    elif ci_option == 'lower':
        upper_percentile = ci
        confidence_interval = [-np.inf, np.percentile(median_diffs,
upper_percentile)]
    elif ci_option == 'upper':
        lower_percentile = 100 - ci
        confidence_interval = [np.percentile(median_diffs,
lower_percentile), np.inf]

    # Perform Mann-Whitney U test
    u_statistic, p_value = stats.mannwhitneyu(data1, data2,
alternative='two-sided')

    # Calculate actual median difference
    actual_median_diff = np.median(data1) - np.median(data2)

    # Print summary
    ci_text = f"{confidence_interval[0]:.6e} to
{confidence_interval[1]:.6e}" if ci_option == 'two-sided' else
confidence_interval[1] if ci_option == 'lower' else
confidence_interval[0]
    summary = f"""W = {u_statistic}, p-value = {p_value:.5f}
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
{ci_text}
sample estimates:
difference in location
{actual_median_diff}"""

    print(summary)

# Example usage with the provided data setup
data1 = np.array([0.59, 0.87, 1.10, 1.10, 1.20, 1.30, 1.60, 1.70, 3.20,
4.00])
data2 = np.array([0.30, 0.36, 0.50, 0.70, 0.70, 0.90, 0.92, 1.00, 1.30,
9.70])

# To use this function, simply call it with your data and specify the
desired confidence interval option
# For example:
analyze_data_full(data1, data2, ci_option='lower')
>>
W = 76.5, p-value = 0.04911
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
1.0200000000000002
sample estimates:
difference in location
```

```
0.44999999999999996
```

The test statistic is the sum of ranks in the group with fewer observations. Here either group could be used because sample sizes are equal. Choosing the residential group, the sum of ranks is 78.5. An exact test cannot be computed with a fractional test statistic, so the large-sample approximation form of the test will automatically be computed. Note that `stats.mannwhitneyu` subtracts the smallest possible test statistic prior to reporting the result. Here the smallest possible value for W_{rs} equals 2, so the test statistic reported by `stats.mannwhitneyu` equals 76.5.

The conclusion is that ammonia plus organic nitrogen concentrations from industrial precipitation differ significantly from those in residential precipitation at these locations by a median difference of 0.4499. This estimate is the Hodges-Lehmann estimate discussed later in section 1.5 and is presented along with its confidence interval when specifying the option is true.

1.4 The Large-sample Approximation to the Rank-sum Test

• For the rank-sum test, the distribution of the exact test statistic W_{rs} is closely approximated by a normal distribution when the sample size for each group is 10 or more (**Figure. 2**). With $n = m = 10$, there are 184,756 possible arrangements of the data ranks (this can be computed with the scipy.special.comb in Python). The sum of ranks for one of the two groups for all arrangements comprises the exact distribution of *Wrs*, shown as bars in **Figure 2,** with a mean of 105.

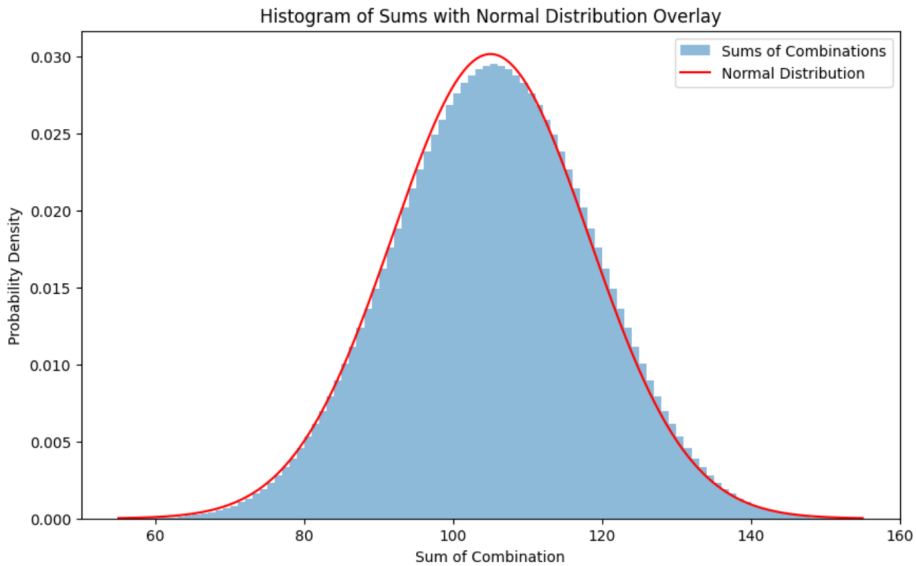


Figure 2. Histogram showing the distribution of the exact test statistic W_{rs} and its fitted normal approximation for n =10 and m =10.

• Superimposed on the exact distribution is the normal distribution that closely approximates the exact values. This demonstrates how well the *p*-values can be approximated even for relatively small sample sizes. The approximation does not imply that the data are, or must be, normally distributed. Rather, it is based on the near normality of the test statistic at large sample sizes. If there are no ties and the assumptions of H_0 are valid, W_{rs} has a mean, μ_W, and standard deviation, σ_W , of:

$$\mu_W = n \cdot (N + 1)/2 \qquad \text{Eq. (1)}$$

$$\sigma_W = \sqrt{n \cdot m \cdot (N+1)/12} \quad \text{Eq. (2)}$$

where $N = n + m$

• The *p*-value from the large-sample approximation is computed by standardizing W_{rs} and making a continuity correction. The continuity correction shifts the normal distribution to fit halfway through the top of the bars of the exact test statistic distribution. The correction moves the probability of occurrence from the outer edge of each bar to its center prior to using the normal curve. It therefore equals *d*/2, where *d* is the minimum difference between possible values of the test statistic (the bar width). For the rank-sum test $d = 1$, as the test statistic values change by units of one. *Zrs*, the standardized form of the test statistic, is therefore computed as:

$$Z_{rs} = \begin{cases} \dfrac{W_{rs} - \frac{d}{2} - \mu_W}{\sigma_W} \text{ if } W_{rs} > \mu_W \\ 0 \text{ if } W_{rs} = \mu_W \\ \dfrac{W_{rs} + \frac{d}{2} - \mu_W}{\sigma_W} \text{ if } W_{rs} < \mu_W \end{cases} \quad \text{Eq. (3)}$$

Z_{rs} is the quantile of the standard normal distribution from which the *p*-value is computed. For the precipitation nitrogen in example 1 the approximate *p*-value is 0.0491 (see the Python output in the previous section). Reporting the *p*-value shows how close the risk of type I error is to 0.05.

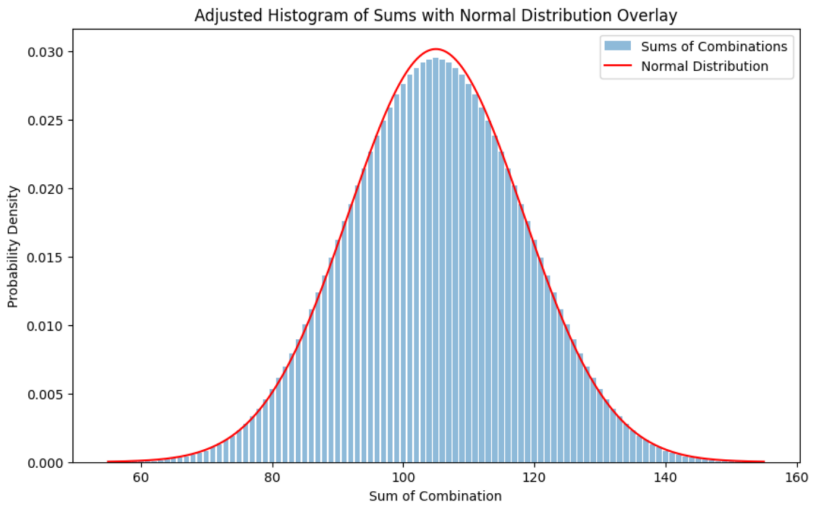


Figure 3. Adjusted histogram showing the distribution of the exact test statistic W_{rs} and its fitted normal approximation for $n = 10$ and $m = 10$.

Note that a tie correction for the standard deviation of the large-sample test statistic σ_W is necessary when ties occur and tied ranks are assigned (Conover, 1999). The formula below for σ_W should be used for computing the large-sample approximation rather than the uncorrected σ_W whenever ties occur.(***HW07 #1***)

$$\sigma_{Wt} = \sqrt{\frac{nm}{N(N-1)} \sum_{k=1}^{N} R_k^2 - \frac{nm(N+1)^2}{4(N-1)}} \quad \text{Eq. (4)}$$

where $N = n + m$.

Note (*Example 1*)

What Do Lower P-values Mean?

- A **low *p*-value** (usually set below a threshold like 0.05) indicates that the observed rank sum is unlikely to have occurred by random chance under the null hypothesis. In simpler terms, if the *p*-value is low, it suggests that there is a statistically significant difference between the medians of the two groups being compared.
- In the context of your scenario, a low *p*-value would mean that the evidence suggests a significant difference in medians, supporting the alternative hypothesis that the group with the lower sum of ranks (implying lower median) is indeed different from the other group.

If the sum of ranks for the *n* samples is significantly lower, resulting in a low *p*-value, it indicates that the observations in the nn group tend to be lower than those in the mm group, leading to the conclusion that the medians between the two groups differ significantly.

2 The Permutation Test of Difference in Means

• Permutation tests solve the long-standing riddle of how to test for differences between means for skewed, moderately sized datasets. They compute the *p*-value using computer-intensive methods (see section 2.2.) rather than assuming data follow normal distributions.

• Permutation tests are also called resampling methods (Good, 2001), randomization tests (Manly, 2007), and observation randomization tests (Brown and Rothery, 1993). Although they were conceived of in the early 1900s, software for quickly computing them became available around the late 1980s.

2.1 Assumptions of the Permutation Test of Difference in Means

• A two-sample permutation test for differences in means avoids the assumptions of the parametric *t*-test (section 3). ***The t-test requires that the data from each group follow a normal distribution*** and that the groups have the same variance.

• Violation of these assumptions leads to a loss of power, raising *p*-values and failing to find differences between group means when they occur. These assumptions are avoided by using a ***permutation test***.

• The permutation test assumes only that the data from the two are exchangeable (Good, 2001). The exchangeable assumption is that any value observed in one group may belong in the population of either group.

2.2 Computation of the Permutation Test of Difference in Means

• Permutation tests calculate either all of the possible test results that could be computed for the observed data or a large random selection of those results, and ***then determine what proportion of the computed results are equal to or more extreme than the one result obtained using the dataset tested***. That proportion is the *p*-value of the test.

• For a two-sample permutation ***test of means***, the test statistic is the observed difference in the two group means, $\bar{x} - \bar{y}$. If the null hypothesis is true, the group assignment is arbitrary, as there is no difference in the means and the data in essence come from the same population.

• Therefore, the data are rearranged regardless of group assignment in either all possible rearrangements or in several thousand randomly selected rearrangements. This produces a different set of numbers assigned to the two groups in each rearrangement.

• The difference in group means is computed and stored after each rearrangement, representing the distribution of differences to be expected when the null hypothesis is true.

• The proportion of differences from the rearrangements that equal or exceed the one observed difference from the original data ***is the permutation p-value of the test***.

Example 2

With $n = m = 10$, there are 184,756 possible rearrangements of assigning data to groups. As an example, one rearrangement for the precipitation nitrogen data is found in the third column of panda's data frame output below. Instead of computing all of these assignments, the permutation procedure will randomly rearrange the group assignment many thousands of times and compute the difference in the resulting means.

	Value	Source	Random_Rearrangement
0	0.59	Indust	Residien
1	0.87	Indust	Indust
2	1.10	Indust	Indust
3	1.10	Indust	Residien
4	1.20	Indust	Residien
5	1.30	Indust	Indust
6	1.60	Indust	Indust
7	1.70	Indust	Indust
8	3.20	Indust	Indust
9	4.00	Indust	Residien
10	0.30	Residien	Residien
11	0.36	Residien	Residien
12	0.50	Residien	Residien
13	0.70	Residien	Residien
14	0.70	Residien	Indust
15	0.90	Residien	Indust
16	0.92	Residien	Residien
17	1.00	Residien	Indust
18	1.30	Residien	Indust
19	9.70	Residien	Residien

```
Permutation Test of Difference Between 2 Group Means
Data: NH4orgN by where
Number of Possible Permutations is greater than 10000
R = 10000 p-value = 0.9964
Alt Hyp: true difference in means is not equal to 0
Sample estimates:
mean of Indust = 1.666, mean of Residien = 1.638
Diff of means = 0.028
95 percent confidence interval
-1.530 1.528
```

Out of 10,000 possible rearrangements of the where column, 99.5 percent of the absolute value of the estimated differences equaled or exceeded the observed difference of 0.028 (**Figure. 4**). Therefore, the observed difference in means ***is not unusual*** at all and the permutation *p*-value is far greater than any reasonable significance level. The conclusion is

to fail to reject H_0. ***There is little evidence that the group means differ***. The advantage of this test over a *t*-test is that there is no concern that the nonsignificant result might be a result of the unfulfilled requirement that the input data follow a normal distribution.

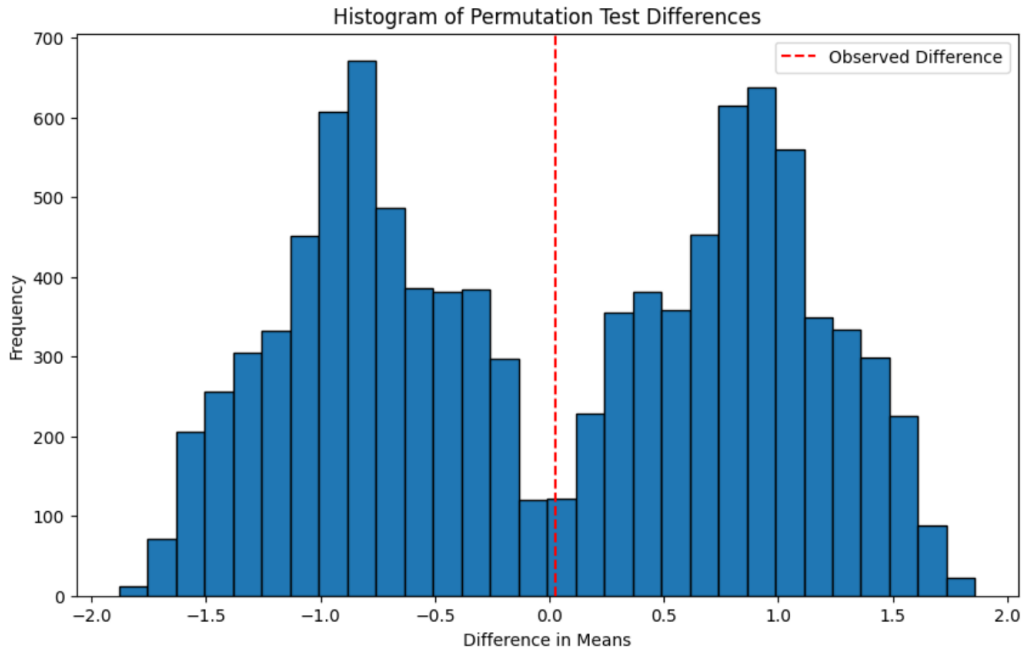


Figure 4. Histogram showing 10,000 permuted differences in group means for the dataset, computed by rearrangement of the group assignments. The observed difference in means from the original data is the solid vertical line.

3 The *t*-test

• The *t*-test has been the most widely used method for comparing two independent groups of data and is familiar to most Environmental scientists. ***However, there are five often overlooked problems with the t-test that make it less applicable*** for general use than the rank-sum or permutation tests. These are:

1. **Lack of power when applied to skewed data,**
    a. The *t*-test assumes that the data follow a normal distribution. When the data are highly skewed, meaning they are not symmetrically distributed around the mean, the *t*-test may not perform well and can lack statistical power. In other words, it might fail to detect a true difference between the groups because it's not designed to handle skewed distributions effectively.

2. **Dependence on an additive model,**
    a. The *t*-test relies on the assumption that the group means differ by a constant amount. This is known as an additive model. If the relationship between the groups is not additive, meaning the difference between groups varies across the range of values, the *t*-test may not provide accurate results.
3. **Lack of applicability for censored data,**

a. Censored data are observations where the true value is known to fall within a certain range but the exact value is unknown or "censored." The *t*-test does not handle censored data well because it assumes complete information about the data points. When dealing with censored data, other statistical methods like survival analysis or methods specific to censored data are more appropriate.

4. **Assumption that the mean is a good measure of central tendency for skewed data,**
   a. The *t*-test assumes that the mean accurately represents the central tendency of the data. However, for skewed distributions, the mean may be influenced by extreme values, leading to a biased estimate of central tendency. In such cases, the median, which is less affected by extreme values, may be a better measure of central tendency.

5. **Difficulty in detecting non-normality and inequality of variance for the small sample sizes common to Environmental data.**
   a. The *t*-test assumes that the data are normally distributed and that the variances of the two groups being compared are equal. However, in practice, it can be challenging to determine if these assumptions hold, especially with small sample sizes common in Environmental data. Violations of these assumptions can lead to inaccurate results. Permutation tests, which are distribution-free and nonparametric, can be more robust in such situations as they do not rely on distributional assumptions.

These problems were discussed in detail by Helsel and Hirsch (1988).

3.1 Assumptions of the *t*-test

• In order to compute an accurate *p*-value the *t*-test assumes that ***both groups of data are normally distributed around their respective means***.

• The test originally also assumed that the two groups have the same variance—a correction for unequal variance was added later.

• The *t*-test is a test for differences in central location only, and assumes that there is an additive difference between the two means, if any difference exists. ***These assumptions of normality and equal variance are rarely satisfied with Environmental data***. The null hypothesis is stated as:

H_0: $\mu_x = \mu_y$ the means for groups *x* and *y* are identical.

If rejected, the alternative hypothesis is either two-sided or one-sided:

H_0: $\mu_x \neq \mu_y$ (two-sided)
H_0: $\mu_x > \mu_y$ (one-sided)

3.2 Computation of the Two-sample *t*-test Assuming Equal Variances

• Two independent groups of data are to be compared. Each group is assumed to be normally distributed around its respective mean value, with each group having the same variance.

• The sole difference between the groups is that their means may not be the same—one is an additive shift from the other.

• The test statistic (*t* in eq. 5) is the difference in group means, divided by a measure of noise:

$$t = \frac{\bar{x} - \bar{y}}{s\sqrt{\frac{1}{n} + \frac{1}{m}}} \quad \text{Eq. (4)}$$

where

$\bar{x}$ is the sample mean of data in the first group x_i from i =1, 2, ..., n, and
$\bar{y}$ is the sample mean of data in the second group yj from j =1, 2, ..., m.
s is the pooled sample standard deviation and is estimated by assuming that each group's standard deviation is identical (eq. 6).

$$s = \sqrt{\frac{(n-1)s_x^2 + (m-1)s_y^2}{n + m - 2}} \quad \text{Eq. (5)}$$

• If the null hypothesis of equal group means, H_0: $\mu_x = \mu_y$, is rejected because the two-sided *p*-value < α, the two-sided alternative that the group means do not differ is H_A: $\mu_x \neq \mu_y$.

• Similarly, the one-sided alternative H_A: $\mu_x \neq \mu_y$ is employed when the mean of group *X* is expected to be greater than the mean of group Y prior to seeing any data.

• The null hypothesis is rejected in favor of the one-sided alternative when the one-tailed *p*-value is less than α.

• These *p*-values are accurate if the data from each group follow the test's assumptions. If data are skewed or of unequal variance the *p*-values are expected to be too large and a false tendency to not find differences occurs.

3.3 Adjustment of the t-test for Unequal Variances

• When two groups have unequal variances, the *t*-test's degrees of freedom should be adjusted using Satterthwaite's approximation (here called the Welch's *t*-test), which was developed in the 1940s.

• The degrees of freedom will be lowered, changing the *p*-value and penalizing the test because it is being applied to data that do not meet the *t*-test's assumptions. Statistics software correctly performs the Welch/Satterthwaite version of the test by default.

• Unless you have a clear reason for doing so (and we doubt that there is one), do not remove this adjustment by performing the *t*-test using the pooled standard deviation or with the option to assume equal variance.
• Always assume unequal variances. There is no benefit to performing the pre-1940s unadjusted test, as the adjustment goes to zero when sample variances are identical.

• Using the unadjusted test on data with unequal variance will likely provide an incorrect *p*-value that may be either too small or too large.

Example 3

The Shapiro-Wilk test of normality for each of the two groups of ammonia plus organic nitrogen from example 1 show that neither group follows a normal distribution at the $\alpha = 0.05$ level.

```
group1 = df[df['Source'] == 'Indust']['Value'].values
group2 = df[df['Source'] == 'Residien']['Value'].values

# Shapiro-Wilk test for Indust
stat_a, p_a = shapiro(group1)
print(f'Indust: Statistics={stat_a}, p={p_a}')

# Shapiro-Wilk test for Residien
stat_b, p_b = shapiro(group2)
print(f'Residien: Statistics={stat_b}, p={p_b}')
>> Indust: Statistics=0.8034604787826538, p=0.015972202643752098
Residien: Statistics=0.4675414562225342, p=1.517449732091336e-06
```

As this dataset is small and nowhere near the requirement for the Central Limit Theorem to hold, we should expect some loss of power, inflating the *p*-value of the *t*-test.

Testing for unequal variance, both the parametric Levene's and nonparametric Fligner-Killeen tests (Aho, 2016)—discussed in section 2.1 (the second lecture on week07)— find no difference in the variances of the two groups, though 10 observations is a small amount of data to work with.

```
# Perform Levene's test
statistic, p_value = levene(group1, group2, center='median')
print("Levene's test statistic:", statistic)
print("p-value:", p_value)

# Perform Fligner-Killeen test
statistic, p_value = fligner(group1, group2)

print("Fligner-Killeen test statistic:", statistic)
print("p-value:", p_value)
>> Levene's test statistic: 0.2242391747820643
p-value: 0.6415212057122008
Fligner-Killeen test statistic: 0.06754838244965784
p-value: 0.7949403585559228
```

The *t*-test is computed with the default two-sided alternative using the scipy.stats.ttest_ind in Python. The Welch's *t*-test is used by default:

```
# Perform two-sample t-test
statistic, p_value = ttest_ind(group1, group2)

print("Two-sample t-test statistic:", statistic)
print("p-value:", p_value)
```

```
>> Two-sample t-test statistic: 0.029044191616409778
p-value: 0.9771489515202003
```

From the *p*-value of 0.977, the null hypothesis of no difference cannot be rejected. There is essentially no evidence that the means differ using the *t*-test.

3.4 The *t*-test After Transformation Using Logarithms

• A *t*-test on logarithms of data has been a popular approach to use when data are skewed. Environmental data more often appear closer to the shape of a skewed lognormal distribution than to a normal distribution.

• In log units, skewed data often appear close to a normal distribution with equal group variance. Not all who use it realize that by transforming with logs, the test determines whether the geometric means, and not arithmetic means, of the two groups differ.

• When the logarithms of data follow a normal distribution, the geometric mean estimates the sample median of the data. If a different transformation produces a distribution similar in shape to the normal, the *t*-test on transformed units can be considered a test for difference in group medians.

• Results of the *t*-test on data transformed to symmetry are often similar to those of the rank-sum test, as both are tests for differences in medians. However, the rank-sum test does not require the analyst to spend time determining what an appropriate transformation to symmetry might be.

Example 4

```
# Perform two-sample t-test
log_group1 = np.log(group1)
log_group2 = np.log(group2)
statistic, p_value = ttest_ind(log_group1, log_group2)

print("Two-sample t-test statistic:", statistic)
print("p-value:", p_value)
>> Alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.665563 0.721563
Sample estimates:
mean in group Indust: 0.352
mean in group Residen: -0.129
```

3.5 Conclusions as Illustrated by the Precipitation Nitrogen Example

• Several tests were run on the precipitation nitrogen data from **example 1**, with varying outcomes. This is because not all of the tests have the same objectives, and not all of the tests have the same requirements.

1. The most important decision to make, before running a statistical test, is to determine what the correct statistic is for the question being asked. Does one group have higher

values than the other? This is a frequency question and is best answered by a test on frequency measures (percentiles) such as medians. Concentrations in the industrial group are more often higher than those in the residential group. This is what was tested by the rank-sum test and seen in the boxplots of **Figure 2**. Running *t*-tests on logarithms may approximately test the same hypothesis, but there is no advantage to using them versus the actual rank-sum test.

2. When the interest is in means, because the objective is to test the cumulative amounts in each group (mass, volume, cumulative exposure), use a permutation test instead of the *t*-test. The lack of power encountered when a *t*-test is applied to non-normal and unequal variance data is overcome by permutation methods. Skewness and outliers inflate the sample standard deviation used in the *t*-test and it often fails to detect the differences present that could be seen with a permutation test.
3. Both permutation tests and *t*-tests determine whether the total amounts (standardized by sample size n) are the same or different. For the precipitation nitrogen data, the total in each group is about the same owing to the one large value in the residential group. Most of the nitrogen present came in that one precipitation event, such data often deserve closer scrutiny and may provide information about the processes occurring. Do not throw away outliers in order to meet the requirements of a substandard test. Use a better test and learn from the entire dataset.
4. Decide which type of test to use based on the study objectives rather than on the shape of the data distribution. For questions of whether one group has higher values than the other, compute the rank-sum test. For concerns about totals or mass, use a permutation test to judge differences in group means while protecting against the *t*-test's potential loss of power due to non-normal and unequal variance data.
5. A *t*-test cannot be easily applied to censored data, such as data below the detection limit. That is because the mean and standard deviation of such data cannot be computed without either substituting some arbitrary values or making a further distributional assumption about the data. Helsel (2012) provides several better methods for examining censored data. If the question is whether one group shows higher values than another, all data below the highest reporting limit can be assigned a tied rank and the rank-sum test computed, without making any distributional assumptions or assigning arbitrary values to the data.

• Determining which test is "better" between the Wilcoxon rank-sum test (often implemented as `ranksums` in Python) and permutation tests depends on several factors, including the specific characteristics of your data, your research question, and the assumptions you're willing to make. Here's a comparison between the two:

1. **Assumptions**:
    - **Rank-Sums Test**: The Wilcoxon rank-sum test assumes that the data in each group are independent and identically distributed (but not necessarily normal). It is a parametric test in the sense that it makes specific assumptions about the underlying distribution of the data, but it is less sensitive to violations of normality compared to parametric tests like the *t*-test.
    - **Permutation Test**: Permutation tests are non-parametric and distribution-free. They make fewer assumptions about the underlying distribution of the data but rely on the assumption of exchangeability under the null hypothesis (i.e., that the groups are exchangeable).
2. **Power**:

   - **Rank-Sums Test**: The Wilcoxon rank-sum test may have less power compared to permutation tests in certain situations, especially when the data are heavily skewed or the sample size is small.
   - **Permutation Test**: Permutation tests can be more powerful in detecting differences between groups, especially when the assumptions of parametric tests are violated. They are particularly robust against violations of distributional assumptions and are more flexible in terms of study design and data structure.
3. **Computational Complexity**:
   - **Rank-Sums Test**: The Wilcoxon rank-sum test involves ranking the data and calculating the test statistic based on the sum of ranks, which can be computationally efficient for moderate-sized datasets.
   - **Permutation Test**: Permutation tests require generating permutations of the data, which can be computationally intensive, especially for large datasets or a large number of permutations.
4. **Interpretability**:
   - **Rank-Sums Test**: The Wilcoxon rank-sum test provides a test statistic and a *p*-value, which can be interpreted similarly to other parametric tests.
   - **Permutation Test**: Permutation tests provide a *p*-value based on the distribution of permuted data, which may require a bit more explanation and interpretation compared to parametric tests.

In summary, the choice between the Wilcoxon rank-sum test and permutation tests depends on the specific characteristics of your data and the assumptions you are willing to make. Both tests have their strengths and limitations, and it's essential to consider these factors when selecting the appropriate test for your analysis. In general, if you are concerned about distributional assumptions or have small sample sizes, permutation tests may be a better choice. However, for moderately sized datasets with approximately symmetric distributions, the Wilcoxon rank-sum test can also be a robust option.