Paired Difference Tests of the Center

EN5423 | Spring 2024

w09_paired_diff_test_center_01.pdf
(Week 9)

Contents

Intro

Ex 1) To determine the effectiveness of an acid solution in developing wells in carbonate rock, yields of 20 wells were measured both before and after treatment of the wells with acid. Factoring out the differences in yield between wells, have the yields changed as a result of using the acid? What is the magnitude of this change?

Ex 2) Annual sediment loads are measured at two sites over a period of 24 years. Both drainage basins are of essentially the same size and have the same basin characteristics. However, logging has occurred in one basin during the period but not in the other. Can the portion of year-to-year variation in load due to differences in precipitation be compensated for in determining whether the site containing logging produced generally higher loads than the other?

Ex 3) Two laboratories are compared in a quality assurance program. Each lab is sent one of a pair of 30 samples split into duplicates in the field to determine if one lab consistently over- or under-estimates the concentrations of the other. If no difference between the labs is seen, then we should be able to do our analysis using data from both laboratories. The differences between labs must be discerned beyond the sample-to-sample differences.

• Each of the example situations mentioned above is addressed by using the matched-pair tests of this chapter. As opposed to the tests of ***"Testing Differences Between Two Independent Groups"***, we now consider data having a logical pairing of observations within each group.

• There may be a great deal of variability from one pair to another, as with the year-to-year pairs of sediment data in the second example above. Both basins may exhibit low yields in dry years and higher yields in wet years. This variability among pairs of observations is noise that would obscure the differences between the two groups being compared if the methods of week07 ***"Testing Differences Between Two Independent Groups"*** were used.

• Instead, blocking is used to eliminate the influence of this noise by basing the analysis on the pairwise differences between the groups. Tests are then conducted on the set of differences to determine whether the two groups differ significantly (table 1).

• Two nonparametric tests, the sign test and the signed-rank test, determine whether one group's paired observation is generally higher than the other group's paired observation. Also presented is the paired *t*-test, the parametric test of whether the mean difference between the groups equals zero. The *t*-test is used when the mean is of interest and requires that the differences between paired observations be normally distributed.

• A permutation test for determining whether the mean difference equals zero is also presented as a more powerful and flexible alternative to the paired *t*-test when differences do not follow a normal distribution.

• After surveying graphical methods to illustrate the test results, estimators for the difference between the two groups are discussed.

Note

Unpaired tests and matched-pair tests each have their own strengths and weaknesses, and their suitability depends on the specific context of the research question and the nature of the data. Neither can be universally considered "bad" or "good," but rather appropriate or inappropriate based on the situation.

Unpaired Tests:

- **Strengths:**
  - Useful when subjects in different groups are independent or unrelated.
  - Typically easier to conduct when pairing individuals is not feasible or practical.
  - Applicable when analyzing data from independent samples or groups.
- **Weaknesses:**
  - May not account for potential confounding variables or individual differences between groups.
  - Less effective at controlling for variability compared to matched-pair tests.
  - Assumes independence between observations in each group, which may not always hold true.

Matched-Pair Tests:

- **Strengths:**
  - Controls for individual differences by pairing similar subjects or units.
  - Reduces variability by comparing each subject or unit to its own matched counterpart.
  - More powerful for detecting differences when variability within pairs is smaller than variability between pairs.
- **Weaknesses:**
  - Requires careful matching of subjects or units, which can be challenging and time-consuming.
  - Limited to situations where logical pairings can be established between observations.
  - May not be feasible when pairing is not possible or when the matching criteria are subjective or difficult to define.

Table 1 Paired difference tests of this chapter and their characteristics.
[For the sign test and the signed-rank test, the data from one group are frequently higher than the data from the other group. For the paired *t*-test and the permutation test on mean difference, one group has a higher mean. HA is the alternative hypothesis, the signal to be found if it is present]

Characteristic	Sign test	Signed-rank test	Paired *t*-test	Permutation test on mean difference
Class of test	Nonparametric	Nonparametric	Parametric	Permutation
Distributional assumption for differences	None	Symmetry	Normal distribution	Symmetry
Estimator of difference	Median difference	Hodges-Lehmann estimate	Mean difference	Mean difference

For ***paired observations*** (x_i, y_i), their differences

$$D_i = x_i - y_i \qquad \text{Eq. (1)}$$

Where i = 1, 2, ..., n

• The tests in this chapter ***determine whether x_i and y_i are from the same population***—the null hypothesis—by analyzing D_i.

• If the median or mean paired difference, D_i, ***significantly differs from zero***, the null hypothesis is rejected. As with the tests of previous week's methods, the most important determinant of which test to use is the study objective.

• If the study is trying to determine if the conditions represented by the two groups are the same, this is a frequency question and best answered by a nonparametric test. If groups are similar (the null hypothesis, H_0), then the observation from one group in the pair will be higher than the paired observation in the other group approximately half the time.

• If groups are not the same, the observation from one group will be higher than the paired observation from the other group at a frequency greater than 50 percent, and the nonparametric test will pick up on this difference.

• We discuss two nonparametric tests: **the sign** and **signed-rank tests**. The sign test examines whether within an (x, y) pair, does x tend to be higher (or lower, or different) than y? The sign test is very useful ***when the magnitude of the paired differences cannot be computed*** but one observation can be determined to be higher than the other, as when comparing a <1 to a 3.

• It is quite useful with censored data (exercises 4 and 5 at the end of this chapter). The sign test is often **less powerful** than the **signed-rank test** because the sign test uses only the algebraic sign (+ or −) of the difference, **ignoring the magnitude**. It treats a large difference as no different than a small difference.

• The signed-rank test is ***generally more powerful than the sign test*** because it uses the magnitudes of differences—a ***larger difference has more weight than a smaller difference***. The signed-rank test's null hypothesis is that the frequency of $x > y$ for pairs is 50 percent, and so the median of x equals the median of y.

• The alternative (two-sided) hypothesis is that the frequency of $x > y$ is not 50 percent, and therefore the medians of x and y differ.

• When the D_i, values follow a normal distribution, a paired t-test can evaluate a different null hypothesis: The mean of the differences $x_i - y_i$ = 0, and therefore the mean of x_i differs from the mean of y_i. This is also equivalent to testing that the sums of the x_i and the y_i are the same, as both groups of paired data have the same sample size.

• Permutation tests on the mean difference are often more powerful alternatives to the paired t-test, as permutation tests are not impaired when the shape of the D_i distribution fails to follow a normal distribution.

• Nonparametric tests determine whether differences in frequencies occur, such as the frequency of $x_i > y_i$. The paired *t*-test determines whether measures of mass (means) of two groups are the same or not.

• Tests on means and tests on frequencies have different goals. Imagine you want to know if one lab method usually reports higher concentrations in water samples compared to another method. For a bunch of water samples, Method A mostly gave moderate concentrations, while Method B mostly gave lower concentrations with a few high values. The average concentration from each method might seem similar. If we use tests like the t-test or permutation test on averages, they might say both methods are the same. However, if we use a nonparametric test to see which method tends to have higher concentrations more often, we might find a difference. In most cases, the method with moderate concentrations was higher than the one with lower concentrations, even though it had some extreme values.

1 The Sign Test

• For data pairs (x_i, y_i), i =1, 2, ..., *n*, the sign test determines whether *x* is frequently larger (or smaller, or different) than *y*, without regard to whether that difference is additive or to the distributional shape of the differences.

1.1 Null and Alternative Hypotheses

The null and alternative hypotheses may be stated as

H_0: Prob $[x > y] = 0.5$

versus one of the three possible alternative hypotheses:

H_{A1}: Prob $[x > y] \neq 0.5$ (Two-sided test—*x* might be larger or smaller than *y*). Reject H_0 when the two-sided *p*-value < α.

H_{A2}: Prob $[x > y] > 0.5$ (One-sided test—*x* is expected to be larger than *y*). Reject H_0 when the one-sided *p*-value < α.

H_{A3}: Prob$[x > y] < 0.5$ (One-sided test—*x* is expected to be smaller than *y*). Reject H_0 when the one-sided *p*-value < α.

1.2 Computation of the Exact Sign Test

• If the null hypothesis is true, about half of the differences (D_i) will be positive ($x_i > y_i$) and about half negative ($x_i < y_i$).

• If one of the alternative hypotheses is true instead, more than half of the differences will tend to be either positive or negative. The significance level, α, reported by software is the probability of obtaining the observed test statistic, or a result more extreme, when the null hypothesis is true.

• The exact form of the sign test is given below. It is the form appropriate when comparing 20 or fewer pairs of samples. With larger sample sizes, the large-sample approximation may be used. (R defaults to using exact tests for small sample sizes). Unfortunately, some software generally performs large sample approximations regardless of sample size.

Computation: Compute $D_i = x_i - y_i$. Ignore all tied data pairs (all $D_i = 0$). Reduce the sample size of the test to the number of nonzero differences $n = N - [number\ of\ D_i = 0]$. Assign a + for all $D_i > 0$, and a - for all $D_i < 0$.

Test statistic: $\mathbf{S^+}$ **= the number of pluses**, the number of times $x_i > y_i$, $i = 1, 2,..., n$.
Decision rule: To reject H_0: $Prob[x > y] = 0.5$, either

1. H_{A1}: Prob $[x > y] \neq 0.5$ (the x measurement tends to be either larger or smaller than the y measurement). Reject H_0 when the two-sided p-value associated with $S^+ < \alpha$.

2. H_{A2}: Prob $[x > y] > 0.5$ (the x measurement tends to be larger than the y measurement). Reject H_0 when the one-sided p-value associated with $S^+ < \alpha$.

3. H_{A3}: Prob$[x > y] < 0.5$ (the x measurement tends to be smaller than the y measurement). Reject H_0 when the one-sided p-value associated with $S^+ < \alpha$.

Example 1. Mayfly nymphs—Exact sign test, small samples.

Counts of mayfly nymphs were recorded in 12 small streams at low flow above and below industrial outfalls. The mayfly nymph is an indicator of ***good water quality***. The question to be considered is whether effluents from the outfalls decreased the number of nymphs found on the streambeds of that region. A type I risk level α of 1 percent is set as acceptable.

```
def binomial_test_results(successes, trials, alternative='greater',
alpha=0.05):
    p_value = binom_test(successes, trials, alternative=alternative)
    sample_estimate = successes / trials

    # Calculate Wilson score interval
    z_critical = np.abs(np.round(norm.ppf(alpha / 2), 4))
    p_hat = successes / trials
    interval_half_width = z_critical * np.sqrt((p_hat * (1 - p_hat))
/ trials + (z_critical**2) / (4 * trials**2))
    lower_bound = max(0, p_hat - interval_half_width)
    upper_bound = min(1, p_hat + interval_half_width)

    results = f"Exact binomial test\n"
    results += f"data: {successes} and {trials}\n"
    results += f"number of successes = {successes}, number of trials
= {trials}, p-value = {p_value:.6f}\n"
    results += f"alternative hypothesis: true probability of success
is {alternative} than 0.5\n"
    results += f"95 percent confidence interval:\n"
    results += f"{lower_bound:.7f} {upper_bound:.7f}\n"
    results += f"sample estimates:\n"
```

```
    results += f"probability of success\n"
    results += f"{sample_estimate:.7f}"

    return results

# Example usage:
results = binomial_test_results(11, 12, alternative='greater')
print(results)
```

```
Exact binomial test
data: 11 and 12
number of successes = 11, number of trials = 12, p-value = 0.003174
alternative hypothesis: true probability of success is greater than 0.5
95 percent confidence interval:
0.6928901 1.0000000
sample estimates:
probability of success
0.9166667
```

The exact one-sided p-value for $S^+ = 11$ is 0.003. Therefore reject that counts above and below the outfall are the same for the stated α of 0.01.

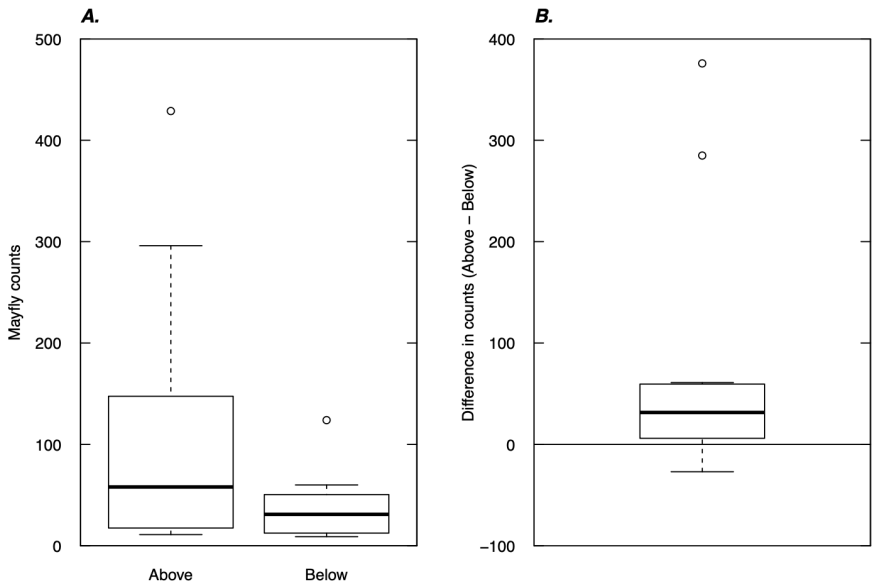


Figure 1. Boxplots of (A) mayfly nymph counts at two different sites, Above and Below, and (B) the differences ($D_i = Above_i - Below_i$)

1.3 The Large-sample Approximation to the Sign Test

• For sample sizes of n >20, the exact sign test statistic can be modified so that its distribution closely follows a normal or chi-square distribution, depending on the form of the approximation used. Again, this does not mean that the data or their differences require normality. It is only the modified test statistic that approximately follows a standard distribution.

The large-sample approximation for the sign test using a standard normal distribution takes the form

$$Z^+ \begin{cases} \dfrac{S^+ - \frac{1}{2} - \mu_{S^+}}{\sigma_{S^+}} & \text{if } S^+ > \mu_{S^+} \\ 0 & \text{if } S^+ = \mu_{S^+} \\ \dfrac{S^+ + \frac{1}{2} - \mu_{S^+}}{\sigma_{S^+}} & \text{if } S^+ < \mu_{S^+} \end{cases} \qquad \text{Eq. (2)}$$

where

$\mu_{S^+} = \frac{n}{2}$ and
$\sigma_{S^+} = \frac{1}{2}\sqrt{n}$

The 1/2 in the numerator of Z^+ is a continuity correction. Z^+ is compared to ***quantiles of the standard normal distribution*** to obtain the approximate *p*-value. The square of *Z* is used when compared to a chi-square distribution with 1 degree of freedom.

Example 2. Mayfly nymphs—Large-sample approximation for the sign test.

```
def sign_test_approximation(S_plus, n):
    # Calculate mean and standard deviation of S_plus
    mu_S_plus = n / 2
    sigma_S_plus = 0.5 * np.sqrt(n)

    # Calculate the test statistic Z_plus
    if S_plus > mu_S_plus:
        Z_plus = (S_plus - mu_S_plus - 0.5) / sigma_S_plus
    elif S_plus < mu_S_plus:
        Z_plus = (S_plus - mu_S_plus + 0.5) / sigma_S_plus
    else:
        Z_plus = 0

    return Z_plus

def sign_test_p_value(Z_plus):
    # Calculate the p-value using the standard normal distribution
    p_value = 1 - norm.cdf(Z_plus)

    return p_value

def sign_test_results(successes, trials, alpha=0.05):
    n = trials
    S_plus = successes
```

```
    # Calculate test statistic Z_plus
    Z_plus = sign_test_approximation(S_plus, n)

    # Calculate p-value
    p_value = sign_test_p_value(Z_plus)

    # Construct results string
    results = f"1-sample proportions test with continuity
correction\n"
    results += f"data: {successes} out of {trials}, null probability
0.5\n"
    results += f"X-squared = {Z_plus**2:.2f}, df = 1, p-value =
{p_value:.6f}\n"
    results += f"alternative hypothesis: true p is greater than
0.5\n"
    results += f"95 percent confidence interval:\n"
    results += f"sample estimates:\n"
    results += f"p = {successes/trials:.7f}"

    return results

# Example usage:
results = sign_test_results(11, 12)
print(results)
```

```
1-sample proportions test with continuity correction
data: 11 out of 12, null probability 0.5
X-squared = 6.75, df = 1, p-value = 0.004687
alternative hypothesis: true p is greater than 0.5
95 percent confidence interval:
sample estimates:
p = 0.9166667
```

The approximation p-value of 0.0047 is reasonably close to the exact p = 0.003, though the exact test should be preferred with these small sample sizes. The availability of exact p-values in Python makes approximate methods far less necessary than in the past—there is no reason to use them if an exact test can be computed.

2 The Signed-rank Test

• The signed-rank test was developed by Wilcoxon (1945) and is sometimes called the Wilcoxon signed-rank test.

• It is used to determine whether the median difference between ***paired observations equals zero***. It may also be used to test whether the median of a single dataset is significantly different from zero.

2.1 Null and Alternative Hypotheses for the Signed-rank Test

For $D_i = x_i - y_i$, the null hypothesis for the signed-rank test is sated as

$H_0: median[D] = 0.$

The alternative hypothesis is one of three statements:

$H_{A1}: median[D] \neq 0$ (Two-sided test—x might be larger or smaller than y)

$H_{A2}: median[D] > 0$ (One-sided test—x might be larger than y)

$H_{A3}: median[D] < 0$ (One-sided test—x might be smaller than y)

• The signed-rank test checks if two groups' data share the same median or differ in location.

• If both groups come from the same population, their differences are split equally above and below zero.

• When the data above zero mirrors that below, the differences become symmetric with enough data.

• The test requires symmetric differences but not necessarily a normal distribution.

• If both groups have the same shape, only differing in median, it is called an additive difference.

• Boxplots of the two groups look similar, except for one being shifted by the median difference.

• The signed-rank test checks if this shift is significant, having more power than the sign test for additive differences.

• In addition, the signed-rank test is also appropriate when the differences are not symmetric in the original scale, but symmetry in the differences can be achieved by a transformation of both datasets.

• If the transformation is with logarithms, a multiplicative relation on the original scale results in an additive relation in the logs. The y group has a higher median and variance, whereas the x (background) group has a lower median and variance. This is quite common in water-resources data.

• In the original scale, the differences between pairs are asymmetric. By taking logs prior to calculating differences, a symmetric distribution of data often results. The log transformation changes a multiplicative relation $y_i = cx_i$, to an additive one: $\log(y_i) = \log(c) + \log(x_i)$

• The transformation makes the variations in log values similar, so the logs of both groups mainly differ in median. This leads to more symmetric differences in log units compared to the original scale. Then, the median difference in logs can be changed back to estimate the median ratio in the original scale.

$$\hat{c} = \text{median}\left[\frac{y}{x}\right] = \exp(\text{median}[\log(y_i) - \log(x_i)]) \qquad \text{Eq. (3)}$$

Note
The signed-rank test is more powerful than the sign test because it takes into account the magnitude of the differences between paired observations, not just their signs. This allows it to detect more subtle differences in the data, making it more sensitive to variations between the groups being compared. Additionally, the signed-rank test utilizes information about the ranks of the observations, which can provide more statistical power compared to simply considering the signs of the differences.

2.2 Computation of the Exact Signed-rank Test

The exact form of the signed-rank test is the best form for comparing 15 or fewer pairs of samples. With larger sample sizes, the large-sample approximation (section 2.3) may be used.

Computation: Compute the absolute value of the differences $|D_i|$, i =1, 2, ..., N. Rank the $|D_i|$ from smallest to largest. Delete any $|D_i| = 0$ and adjust the sample size to $n = N$ - [number of $|D_i| = 0$]. Compute the signed rank $R_i = i = 1,2, \dots, n$

R_i = rank of $|D_i|$ for $D_i > 0$, and
$= -($ rank of $|D_i|)$ for $D_i < 0$

When two nonzero differences are tied, assign the average of the ranks involved to all tied values.

Test statistic: The exact test statistic W^+ is the sum of all signed ranks R_i having a positive sign:

$$W^+ = \sum_{i=1}^{n} (R_i \mid R_i > 0) \qquad \text{Eq. (4)}$$

Where

| signifies "given that."

Decision rule: To reject H_0: $median[D] = 0$ when the p-value for W^+, either one- or two-sided as appropriate, is less than α.

Example 3. Mayfly nymphs—Exact signed-rank test.

```
def wilcoxon_signed_rank_test(Above, Below, alternative='greater'):
    # Create DataFrame
    nymph_sr = pd.DataFrame({'Above': Above, 'Below': Below})

    # Calculate the differences
    nymph_sr['D.i'] = nymph_sr['Above'] - nymph_sr['Below']

    # Calculate the absolute differences
    nymph_sr['SR.i'] = abs(nymph_sr['D.i'])

    # Calculate the signed ranks
```

```
    nymph_sr['SR.i'] = nymph_sr['SR.i'].rank() * nymph_sr['D.i'] /
abs(nymph_sr['D.i'])

    # Perform Wilcoxon signed-rank test
    statistic, p_value = wilcoxon(Above, Below,
alternative=alternative, zero_method='pratt', correction=True)

    # Output the results

    print("V =", statistic, ", p-value =", p_value)
    print("alternative hypothesis: true location shift is",
alternative, "than 0")

    return nymph_sr[['Above', 'Below', 'D.i', 'SR.i']]

# Example usage:
Above = [12, 15, 11, 41, 106, 63, 296, 53, 20, 110, 429, 185]
Below = [9, 8, 38, 24, 48, 17, 11, 41, 14, 60, 53, 124]

nymph_sr = wilcoxon_signed_rank_test(Above, Below,
alternative='greater')
nymph_sr
```

```
V = 72.0 , p-value = 0.00341796875
alternative hypothesis: true location shift is greater than 0
```

	Above	Below	D.i	SR.i
0	12	9	3	1.0
1	15	8	7	3.0
2	11	38	-27	-6.0
3	41	24	17	5.0
4	106	48	58	9.0
5	63	17	46	7.0
6	296	11	285	11.0
7	53	41	12	4.0
8	20	14	6	2.0
9	110	60	50	8.0
10	429	53	376	12.0
11	185	124	61	10.0

The test statistic V, the sum of the positive SR_i values, is 72. The exact test will be computed for small sample sizes unless there are ties in the differences.

2.3 The Large-sample Approximation for the Signed-rank Test

• The large-sample approximation is computed by standardizing the exact test statistic; this is accomplished by subtracting its mean and dividing by its standard deviation.

• The distribution of the test statistic (not the data) was designed to be approximated by a standard normal distribution. This approximation is valid for sample sizes of n >15. The large-sample approximation for the signed-rank test takes the form

$$Z_{sr}^{+} = \begin{cases} \dfrac{W^{+} - \frac{1}{2} - \mu_{W^{+}}}{\sigma_{W^{+}}} & \text{if } W^{+} > \mu_{W^{+}} \\ 0 & \text{if } W^{+} = \mu_{W^{+}} \\ \dfrac{W^{+} + \frac{1}{2} - \mu_{W^{+}}}{\sigma_{W^{+}}} & \text{if } W^{+} < \mu_{W^{+}} \end{cases} \qquad \text{Eq. (5)}$$

where

$$\mu_{W^{+}} = \frac{n \cdot (n+1)}{4} \text{ and}$$

$$\sigma_{W^{+}} = \sqrt{\frac{n \cdot (n+1) \cdot (2n+1)}{24}}$$

The 1/2 in the numerator of Z_{sr}^{+} is the continuity correction. Z_{sr}^{+} is compared to quantiles of the standard normal distribution, a normal distribution with mean of 0 and standard deviation of 1, to obtain the approximate p-value for the signed-rank test.

Example 4. Mayfly nymphs—Large-sample approximation to the signed-rank test.

```
Above = np.array([12, 15, 11, 41, 106, 63, 296, 53, 20, 110, 429,
185])
Below = np.array([9, 8, 38, 24, 48, 17, 11, 41, 14, 60, 53, 124])
differences = Above - Below

# Check the number of observations
n = len(differences)

# Choose the method based on the sample size
if n < 20:
    method = 'exact'
else:
    method = 'approx'

# Perform the Wilcoxon signed-rank test with a specified method
stat, p_value = wilcoxon(differences, alternative='greater',
method=method)
print("Wilcoxon signed-rank test")
print(f"Test Statistic: {stat}, P-value: {p_value}, Method used:
{method}")

method = 'approx'
# Perform the Wilcoxon signed-rank test with a specified method
stat, p_value = wilcoxon(differences, alternative='greater',
method=method)
```

```
print("Wilcoxon signed-rank test")
print(f"Test Statistic: {stat}, P-value: {p_value}, Method used:
{method}")
Wilcoxon signed-rank test
Test Statistic: 72.0, P-value: 0.00341796875, Method used: exact
Wilcoxon signed-rank test
Test Statistic: 72.0, P-value: 0.004816487886294337, Method used: approx
```

The large-sample approximation p-value of 0.0048 is similar to the exact test results ($p = 0.003$). If sample sizes are small and p-values are close to where a decision may change if the p-value changes by small amounts, use the exact test. Generally, let Python perform the appropriate test automatically by not specifying the method = option.

2.4 Permutation Version of the Signed-rank Test

There is no advantage to this over using the exact test, but it will likely be a better approximation than the large-sample approximation of the previous section.

Example 5. Mayfly nymphs—Permutation version of the signed-rank test.

```
def wilcoxon_signed_rank_test(above, below, alternative='greater',
n_resamples=1000000000):
    # Calculate the differences and filter out zeros
    differences = np.array(above) - np.array(below)
    non_zero_differences = differences[differences != 0]

    # Calculate ranks of the absolute differences
    ranks = rankdata(np.abs(non_zero_differences))

    # Compute signed ranks
    signed_ranks = ranks * np.sign(non_zero_differences)

    # Compute the sum of positive ranks for the actual data
    w_plus = np.sum(signed_ranks[signed_ranks > 0])

    # Number of all possible permutations
    total_permutations = 2 ** len(non_zero_differences)
    num_combinations = min(n_resamples, total_permutations)

    # Collect permutation stats
    permutation_stats = []
    for _ in range(num_combinations):
        # Randomly flip signs
        random_signs = np.random.choice([-1, 1],
size=len(non_zero_differences))
        permuted_signed_ranks = signed_ranks * random_signs
        permuted_stat =
np.sum(permuted_signed_ranks[permuted_signed_ranks > 0])
        permutation_stats.append(permuted_stat)
```

```
    # Calculate p-value
    if alternative == 'greater':
        p_value = np.mean([stat >= w_plus for stat in
permutation_stats])
    elif alternative == 'less':
        p_value = np.mean([stat <= w_plus for stat in
permutation_stats])
    else:  # two-sided
        p_value = np.mean([np.abs(stat) >= np.abs(w_plus) for stat in
permutation_stats])

    return w_plus, p_value

# Example data
Above = [12, 15, 11, 41, 106, 63, 296, 53, 20, 110, 429, 185]
Below = [9, 8, 38, 24, 48, 17, 11, 41, 14, 60, 53, 124]

# Running the test
stat, p_val = wilcoxon_signed_rank_test(Above, Below,
alternative='greater')
print(f"Test Statistic: {stat}, P-value: {p_val}")
```

```
Test Statistic: 72.0, P-value: 0.00341796875
```

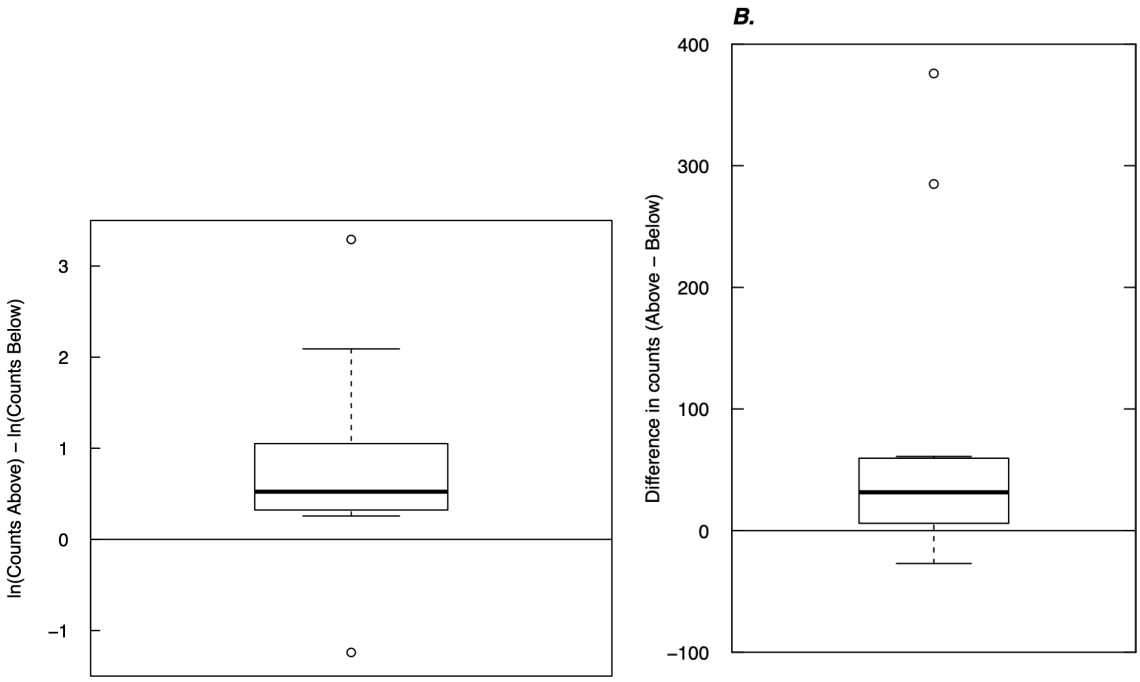
2.5 Assumption of Symmetry for the Signed-rank Test

• The signed-rank test is used to determine if the median difference between two paired groups is zero (i.e., no difference). If the differences between the pairs are not symmetric around zero (asymmetric), the test may falsely suggest that there is a significant difference more often than it should. Essentially, if the distribution of differences leans one way (asymmetry), it can trick the test into rejecting the null hypothesis (no difference) when it should not.

• Some researchers argue that the signed-rank test can be seen as a test for whether differences are asymmetric. However, for asymmetry to significantly affect the test result (*p*-value), it needs to be quite pronounced. Small or moderate asymmetry might not have much impact on the outcome.

• In a *t*-test, which is another statistical test for comparing means, just a single outlier (a value that is much higher or lower than others in the dataset) can severely affect the test's results. However, in the signed-rank test, the differences need to be consistently skewed (most negative differences smaller than positive differences) for the test to wrongly indicate a significant difference solely due to this asymmetry.

• Unlike the *t*-test, outliers do not generally impact the signed-rank test as much. This is because the signed-rank test uses the rank (order) of the differences rather than their actual values, making it less sensitive to extreme values.



(Right) Figure 2. Boxplot of the differences of the natural logarithms of the mayfly data from example 1. (left) **Figure 1B**

Example 6. Mayfly nymphs—Signed-rank test on logarithms.

The Above–Below differences are asymmetric in figure 1B, violating one of the signed-rank test's assumptions and indicating that the differences between the two groups may not be an additive one. Asymmetry can be expected when large values tend to produce large differences and smaller values smaller differences. This indicates that a multiplicative relation between the data pairs is more realistic. Here the natural logs of the data are calculated, and a new set of differences $Dl_i = \log(x_i) - \log(y_i)$ are computed and shown in figure 2. Comparing figure 2 and 1B, note that differences in natural log units are much more symmetric than those in the original scale.

```
# Example 6. Mayfly nymphs—Signed-rank test on logarithms.

differences = np.log(Above) - np.log(Below)
# Perform the Wilcoxon signed-rank test with a specified method
method = 'exact'
stat, p_value = wilcoxon(differences, alternative='greater',
method=method)
print("Wilcoxon signed-rank test")
print(f"Test Statistic: {stat}, P-value: {p_value}, Method used:
{method}")
Wilcoxon signed-rank test
Test Statistic: 69.0, P-value: 0.008056640625, Method used: exact
```

• When using the signed-rank test, having asymmetrical data does not greatly affect the *p*-values. For instance, with the mayfly data, whether we use the original numbers or their natural logs (to make the data more symmetrical), the p-values are quite close (0.003 vs. 0.008). This

similarity is also seen when compared to the sign test, which does not require the data to be symmetrical at all.

• However, the real issue arises when we try to estimate the actual difference between the paired data. If the relationship between the data pairs is more multiplicative (think in terms of ratios or percentages) rather than additive (simple differences), then using an additive approach to measure the difference might lead to inaccurate estimates.

To ensure you are measuring the difference correctly:

- If you think the data operates on a multiplicative scale, transforming the data using logarithms can help model this relationship accurately.
- You can visually check the nature of the relationship (whether it is additive or multiplicative) by looking at scatterplots of the data.
- If the relationship seems additive, and you want to avoid problems caused by asymmetry, ***you could use a permutation version of the signed-rank test***.

This approach will help you decide the best method to apply, based on the actual characteristics of your data.