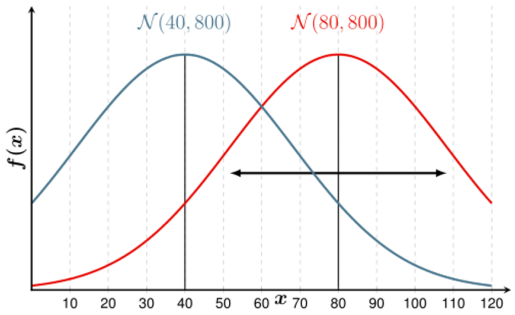
Testing Differences Between Two Independent Groups 02

EN5423 | Spring 2024

w09_two_independent_groups_02.pdf
(Week 9)

Contents



1 Estimating the Magnitude of Differences Between Two Groups

• Once a hypothesis test is completed to compare two data groups, the next step is to quantify the difference between their central tendencies.

• This difference should be evaluated against the effect size, which reflects the minimum change considered scientifically meaningful by the researcher.

• It is crucial to distinguish between statistical significance and practical significance, as they are not inherently synonymous.

• A significant result that shows a smaller difference than the effect size might suggest that the difference is not practically significant.

• Alternatively, such a result could be seen as an early indicator that the difference is beginning to approach practical significance.

• The relevance of the observed difference might vary for different applications; a difference deemed small for one purpose might be significant for another.

• Always reporting the observed difference is advisable, as it provides essential context for evaluating the significance of the results.

1.1 The Hodges-Lehmann Estimator of Difference in Medians

One nonparametric estimate of the difference between two independent groups is a Hodges-Lehmann estimator, $\widehat{\Delta}$ (Hodges and Lehmann, 1963; Hollander and Wolfe, 1999). This estimator is the median of all possible pairwise differences between the *x* values and *y* values:

$$\widehat{\Delta} = \text{median}\left[x_i - y_j\right] \text{ for } x_i, i = 1,2,\dots,n \text{ and } y_j, \text{ for } j = 1,2,\dots,m \qquad \text{Eq. (1)}$$

• In statistical analysis, there are $n \times m$ pairwise differences to consider when comparing two groups of data.

• The estimator, related to the rank-sum test, indicates that adjusting each *x* observation by the estimated value $\widehat{\Delta}$ would result in a rank-sum statistic W_{rs} that shows no significant difference between groups *x* and *y*.

• This shift, $\widehat{\Delta}$, effectively makes the data appear as if there is no difference between the two groups when analyzed with the rank-sum test.

• The estimator is a median unbiased estimator for the difference in medians between the populations *x* and *y*, balancing the probabilities of underestimation and overestimation at exactly one-half.

• When the populations are normal, this estimator is slightly less efficient than the parametric estimator $\bar{x} - \bar{y}$ for medians or means. However, it becomes more efficient, particularly in terms of lower variance, when dealing with non-normal populations.

• Another nonparametric method to estimate the difference in medians is by taking the difference between the sample medians ($x_{med} - y_{med}$) which in an example might be calculated as 10.5.

• This method, while straightforward, typically results in more variability and a larger confidence interval compared to the Hodges-Lehmann estimator.

Example 1 Precipitation nitrogen—Hodges-Lehmann estimator

```
# Perform two-sample t-test
def wilcoxon_rank_sum_test_with_ci(x, y, n_bootstrap=10000,
alpha=0.05):
    # Perform the Mann-Whitney U test with continuity correction
    stat, p = mannwhitneyu(x, y, alternative='two-sided',
use_continuity=True)

    # Calculate all pairwise differences for bootstrapping
    diff = np.array([i - j for i in x for j in y])

    # Bootstrap to compute the confidence interval
    estimates = []
    for _ in range(n_bootstrap):
        sample_diff = resample(diff)
        estimates.append(np.median(sample_diff))

    # Compute the Hodges-Lehmann estimator for the median of the
differences
    hl_estimate = np.median(diff)

    return {
        'W-statistic': stat,
        'p-value': p,
        'Hodges-Lehmann estimate': hl_estimate
    }

# Data from the question
X = np.array([0.59, 0.87, 1.10, 1.10, 1.20, 1.30, 1.60, 1.70, 3.20,
4.00])
Y = np.array([0.30, 0.36, 0.50, 0.70, 0.70, 0.90, 0.92, 1.00, 1.30,
9.70])

# Perform the test and output the results
results = wilcoxon_rank_sum_test_with_ci(X, Y)
print(f"W-statistic: {results['W-statistic']:.2f}")
print(f"P-value: {results['p-value']:.5f}")
print(f"Hodges-Lehmann Estimate: {results['Hodges-Lehmann
estimate']:.6f}")
>> W-statistic: 76.50
P-value: 0.04911
Hodges-Lehmann Estimate: 0.505000
```

Example 2 Hand computation of the Hodges-Lehmann estimator.

Suppose we had a sample of 3 values for the x group and they were values of 15, 17, and 25, and we had a sample of 4 values for the y group and they were the values 8, 27, 3, and 5. We can compute the Hodges-Lehmann estimate by hand by enumerating all possible values of the differences as shown here.

x_i	y_j	All possible differences ($x_i - y_j$)		
15	8	7	9	17
17	27	−12	−10	−2
25	3	12	14	22
	5	10	12	20

Ranked in order from smallest to largest, the 3·4 =12 pairwise differences are

−12, −10, −2, 7, 9, 10, 12, 12, 14, 17, 20, 22.

The median of these is the average of the 6th and 7th smallest values, or $\widehat{\Delta}$ =11. Note that the unusual *y* value of 27 could have been any number greater than 14 and the estimator $\widehat{\Delta}$ would be unchanged; thus $\widehat{\Delta}$ is resistant to outliers.

1.2 Confidence Interval for the Hodges-Lehmann Estimator, $\widehat{\Delta}$

• A nonparametric interval estimate for $\widehat{\Delta}$ illustrates how variable the median difference between groups might be.

• No distribution is assumed for this interval; it is computed using the process for the binomial confidence interval on the median described by finding appropriate rank positions from among the ordered *n·m* pairwise differences that represent the ends of the confidence interval.

• When ***the large-sample approximation*** to the rank-sum test is used, a critical value, $z_{\frac{\alpha}{2}}$, from a function for standard normal quantiles determines the upper and lower ranks of the pairwise differences corresponding to the ends of the confidence interval. Those ranks are:

$$R_1 = \frac{N - \mathrm{z}_{\alpha/2} \cdot \sqrt{\frac{N(n+m+1)}{3}}}{2} \qquad \text{Eq. (2)}$$

$$R_u = N - R_1 + 1 \qquad \text{Eq. (3)}$$

• When the exact test is used for smaller sample sizes, the quantiles for the rank-sum test statistics having a *p*-value nearest to $\alpha/2$ and $1-(\alpha/2)$ are used to find the lower and upper ends of the confidence limit for $\widehat{\Delta}$. The lower limit uses the lower $\alpha/2$ quantile. The upper limit uses the upper $\alpha/2$ quantile plus 1.

Example 3 Hand computation of the confidence interval for the Hodges-Lehmann estimator

The N =12 possible pairwise differences between x and y are
−12, −10, −2, 7, 9, 10, 12, 12, 14, 17, 20, 22.

To determine an $\alpha \cong 0.10$ confidence interval for $\widehat{\Delta}$, the quantiles for the rank-sum statistic at $\alpha/2 = 0.05$ and $1-(\alpha/2) = 0.95$ can be provided wilcoxon_rank_sum_test_with_hl_ci.

```
X = np.array([15, 17, 25])
Y = np.array([8,27,3,5])
wilcoxon_rank_sum_test_with_hl_ci(X, Y, alpha=0.05)
```

```
{'W-statistic': 9.0,
 'p-value': 0.4,
 'Hodges-Lehmann estimate': 11.0,
 '95% confidence interval': (-12, 22)}
```

1.3 Estimate of Difference Between Group Means

• Where group means are of interest, the difference between the means of the two groups $\bar{x} - \bar{y}$ is the most efficient estimator of the mean difference between groups.

• For the precipitation nitrogen data from example 5.1, the output in the *t*-test example in Section 3.3 in *w09_two_independent_groups_02.pdf* shows that an estimated difference in group means equals 1.666–1.638 = 0.028, which was not significantly different from zero.

• Perhaps it is obvious that when x and y are transformed prior to performing the *t*-test the difference in means in the transformed units does not estimate the difference between group means on their original scale. Less obvious is that the retransformation of the difference back to the original scale also does not estimate the difference between group means, but is closer to a function of group medians.

• For the log transformation, the difference in group means in log units when retransformed would equal the ratio of the geometric means of the two groups. How close any retransformation comes to estimating the ratio of group
medians depends on how close the data are to being symmetric in their transformed units.

1.4 Parametric Confidence Interval for Difference in Group Means

• A *t*-confidence interval around the mean difference between groups $\bar{x} - \bar{y}$ is output by the *t*-test command. It is appropriate in situations where the *t*-test may be used—when both data groups closely follow a normal distribution. For the most common situation, where the standard deviations of the two groups are dissimilar and should not be pooled, the confidence interval is:

$$CI = \bar{x} - \bar{y} \pm t_{\alpha/2,(df)} \cdot \sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}} \qquad \text{Eq. (4)}$$

where *df* is the degrees of freedom used in the Welch's *t*-test.

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} \qquad \text{Eq. (5)}$$

where n_1 and n_2 are the sample sizes of the two groups and s_1 and s_2 are the sample variances of the two groups.

1.5 Bootstrap Confidence Interval for Difference in Group Means

• Bootstrap Confidence Interval for Difference in Group Means allows computation of confidence intervals around the difference in group means, regardless of the distribution of data in either group.

• This method is preferred when using a permutation test to test for a difference in group means or when data do not appear to come from a normal distribution.

• Bootstrap intervals also work well for data that follow a specific distribution, and will be quite similar to t-intervals when data follow a normal distribution.

• Bootstrap confidence intervals were described earlier, where the percentile bootstrap method was introduced.

• To compute the bootstrap interval, observations are repeatedly and randomly resampled from the original data with replacement for each group.

• The process involves resampling thousands of times, each time estimating the difference in group means.

• A 95-percent bootstrap confidence interval is found by going to the 2.5 and 97.5 percentiles of the thousands of resampled differences. This method is called the percentile bootstrap method.

• A bootstrap confidence interval on the difference between group means is computed using the bootstrap_confidence_interval function.

1.6 Graphical Presentation of Results

• Overlapping and side-by-side histograms and dot-and-line plots of means and standard deviations were found to inadequately portray the complexities commonly found in Environmental data.

• Probability plots and quantile plots, which plot a point for every observation, can show complexity but often provide too much detail for visual summarization of hypothesis test results.

• Two methods, side-by-side boxplots and Q-Q plots, are very well suited to describing the results of hypothesis tests and visually allowing a judgment of whether data fit the assumptions of the test being employed.

1.7 Side-by-side Boxplots

• The best method for illustrating results of the rank-sum test is side-by-side boxplots.

• With boxplots, only a few quantiles are compared, but the loss of detail is compensated for by greater clarity.

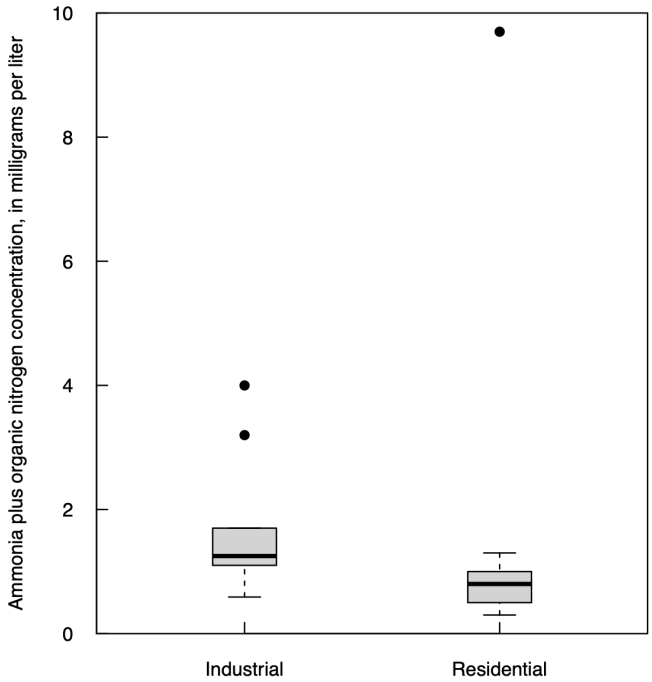


Fig. 1. Boxplots of ammonia plus organic nitrogen from the precipn data. Data from Oltmann and Shulters (1989) by land use type: industrial or residential.

• The difference in medians is clearly displayed, as well as the similarity in spread (IQR).

• The rejection of normality by Shapiro-Wilk tests is evident in the presence of skewness (industrial group) and the one large outlier (residential group).

• Side-by-side boxplots are an effective and concise method for illustrating the basic characteristics of data groups and of differences between those groups.

1.8 Q-Q Plots

• Another method for illustrating rank-sum results is the quantile-quantile (Q-Q) plot, where quantiles from one group are plotted against quantiles from a second group.

• When sample sizes of the two groups are identical, a Q-Q plot is a simple scatterplot of the ordered data pairs from each sample.

• If sample sizes are not equal, quantiles from the smaller dataset are used as is, while quantiles for the larger dataset are interpolated.

• A Q-Q plot typically includes the line $y = x$, indicating identical values for x and y, which helps compare the data directly.

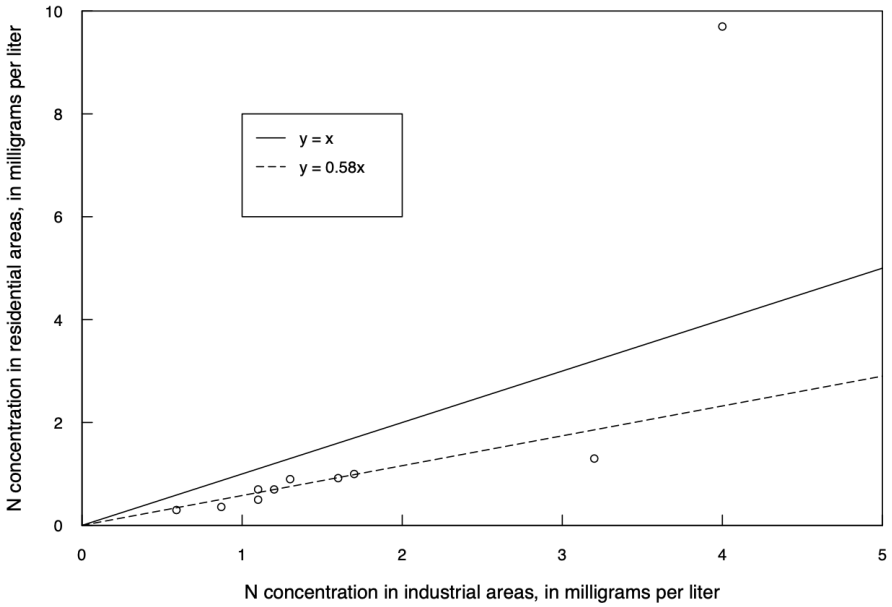


Fig. 2. Q-Q plot of the precipitation nitrogen data

• The Q-Q plot for the precipitation nitrogen data above shows that the data are not parallel to the $y = x$ line, indicating that quantiles differ not by an additive constant but increasingly depart from this line, suggesting a multiplicative relation.

• This indicates that a *t*-test would not be suitable without transformation as it assumes an additive difference, whereas the rank-sum test does not assume this and is applicable when differences are multiplicative.

• The magnitude of the relation between two sets of quantiles on a Q-Q plot can be estimated using the median of all possible ratios, which is a type of Hodges-Lehmann estimator; for the discussed data, the median ratio equals 0.58.

• The data are crowded together at low concentrations but spread further apart at higher concentrations, indicating right-skewness, which was addressed by choosing a natural logarithmic transform for the data.

• The transformed data show (see below) a constant variance across concentrations, indicating decreased skewness, and the slope of the quantiles becomes parallel to the $y = x$ line, turning a multiplicative relation on the original scale into an additive relation on the logarithmic scale.

• The Hodges-Lehmann estimate of the difference between the natural logarithms of x and y, $\widehat{D}$, is equal to -0.5447.

• Q-Q plots illustrate the level of adherence to the assumptions of hypothesis tests and provide insight into which test procedures might be appropriate, demonstrating properties like skewness, the presence of outliers, and variance inequality.

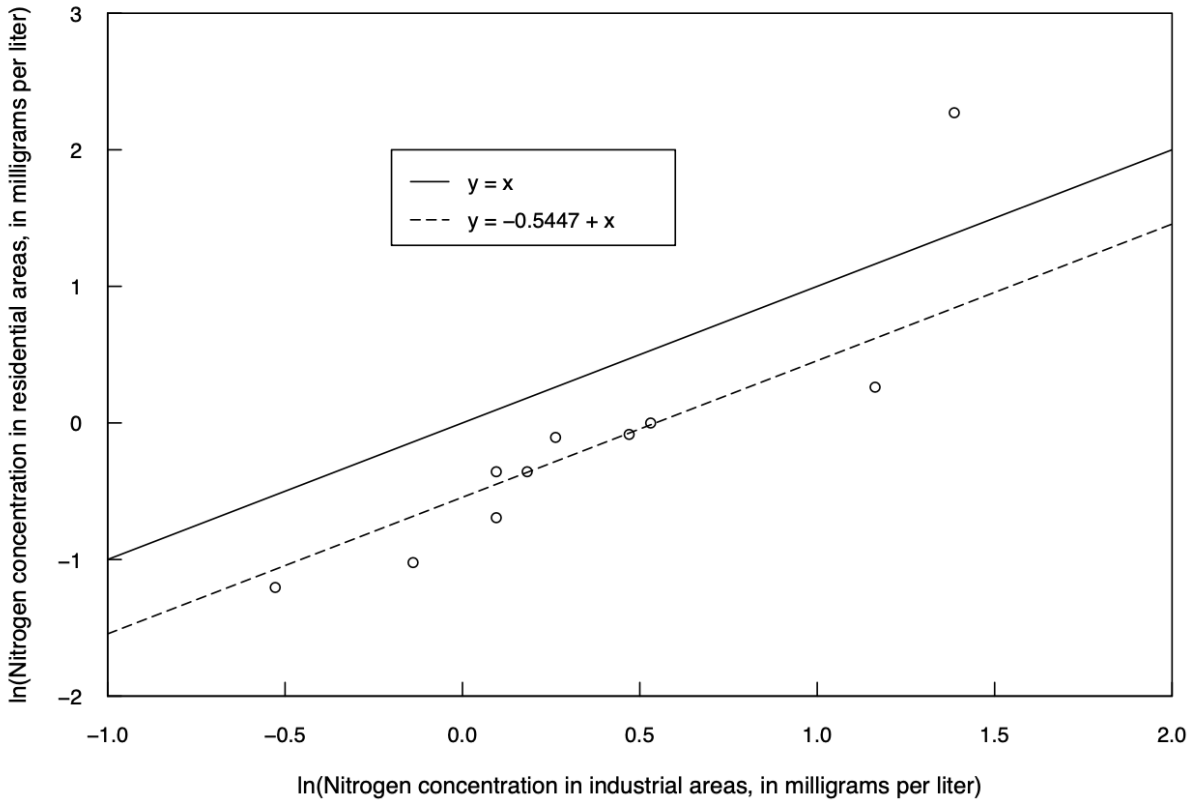


Fig. 3. Q-Q plot of the log-transfored precipitation nitrogen data

1.9 Two-group Tests for Data with Nondetects

• The two-sample *t*-test often substitutes one-half the reporting limit for ***nondetects*** in Environmental studies.

• This substitution method may miss real differences between groups or falsely detect differences that do not exist.

• The rank-sum test on data with one reporting limit is more effective than substitution followed by a *t*-test.

• When there are multiple reporting limits, data is recensored to treat all values below the highest limit as equivalent.

• The rank-sum test is then applied to these recensored values, providing a more accurate analysis than the *t*-test after substitution.

• An example of this is seen in the rank-sum test performed on trichloroethylene (TCE) concentrations in groundwater, using data recensored to show all values below 5 micrograms per liter as "<5."

• The *t*-test on values substituted for nondetects did not find a significant difference between groups due to non-normality and unrealistic assumptions about data uniformity. The rank-sum test, however, indicated a significant difference between groups, particularly at the higher end of the distribution. The difference in test results between the t-test and rank-sum test provides strong evidence of the latter's superiority for data with nondetects.

• The rank-sum test is recommended over the *t*-test for analyzing data with a single or the highest reporting limit, especially in cases of censored data.

2 Tests for Differences in Variance Between Groups

• Differences in central location (mean, median) are not the only types of differences between groups that are of interest; variability in data is also crucial.

• Differences in variance violate the assumptions of the standard uncorrected t-test.
Some analysts use Bartlett's test to check for unequal variances before deciding between an uncorrected t-test and the Welch t-test, which adjusts for unequal variances.

• Using Welch's t-test is recommended over the uncorrected t-test even if variances are not significantly different, as Welch's correction minimizes when variances are similar.

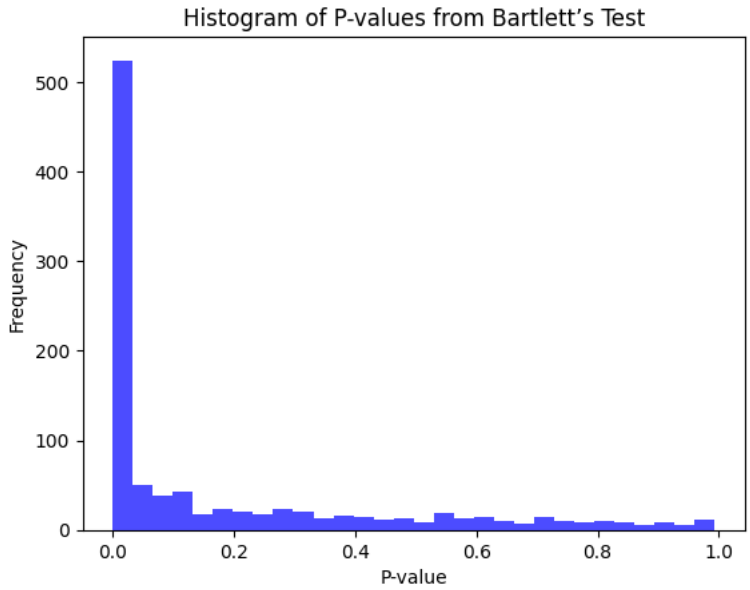
• Bartlett's test, which tests for unequal variance, is highly sensitive to deviations from a normal distribution and is considered non-robust.

• Conover and Iman (1981) noted that Bartlett's test often leads to incorrect conclusions about variance differences and is not recommended unless populations are known to be normal. There are better tests for detecting heteroscedasticity (changing variance) than Bartlett's test.

Example 4 Simulating Bartlett's Test on Lognormal Data

```
def simulate_bartlett_test(mean_log, sd_log, num_samples, num_repetitions):
    """Simulate Bartlett's test on lognormal data across multiple repetitions.

    Args:
        mean_log (float): Mean of the logarithm of the distribution.
        sd_log (float): Standard deviation of the logarithm of the distribution.
        num_samples (int): Number of observations in each simulated group.
        num_repetitions (int): Number of repetitions for the simulation.

    Returns:
        list: p-values from each repetition's Bartlett's test.
    """
    p_values = []
```

```
    for _ in range(num_repetitions):
        # Generate two groups of lognormal data
        data1 = np.random.lognormal(mean_log, sd_log, num_samples)
        data2 = np.random.lognormal(mean_log, sd_log, num_samples)

        # Perform Bartlett's test on the two groups
        _, p_value = bartlett(data1, data2)
        p_values.append(p_value)

    return p_values

# Parameters
mean_log = 0.7
sd_log = 1
num_samples = 50
num_repetitions = 1000

# Run simulation
p_values = simulate_bartlett_test(mean_log, sd_log, num_samples, num_repetitions)

# Plotting the histogram of p-values
plt.hist(p_values, bins=30, color='blue', alpha=0.7)
plt.title('Histogram of P-values from Bartlett's Test')
plt.xlabel('P-value')
plt.ylabel('Frequency')
plt.show()

# Assess how many p-values are below 0.05 (significance level)
significant_tests = sum(p < 0.05 for p in p_values)
print(f"Number of significant tests (out of {num_repetitions}):
{significant_tests}")
```




```
Number of significant tests (out of 1000): 554
```

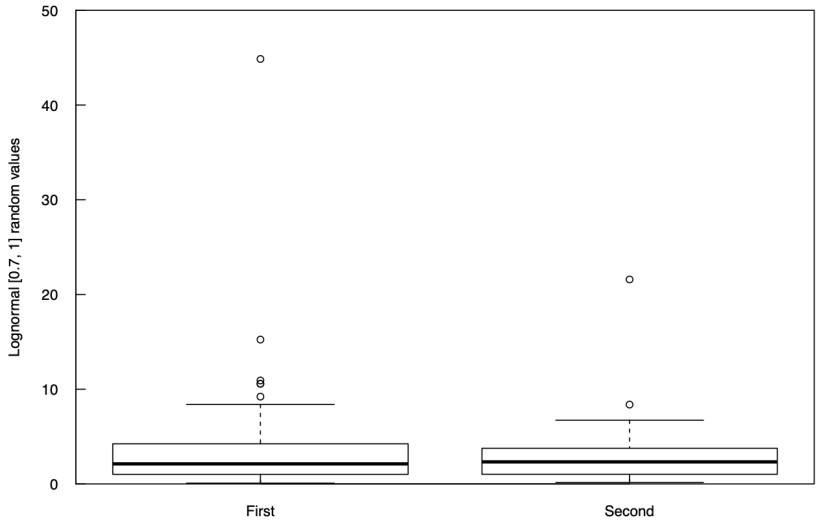


Fig. 4. Box plots of two groups of 50 samples each of randomly generated data from a single lognormal distribution (and so have the same variance). Bartlett's test declares the variances different. The non-normality of the data is a violation of the test's (strict) assumption of a normal distribution.

• The rejection of the null hypothesis of equal variance using Bartlett's test is incorrect when data for both groups are generated from the same lognormal distribution.

• Ideally, a test should incorrectly indicate a difference in variance with a 5 percent probability if it properly adheres to the statistical assumptions of the test, which include normality.

• To demonstrate Bartlett's test's sensitivity to non-normal distributions, a Monte Carlo script in Ex_in_two_groups_02.ipynb, 2,000 data sets from a lognormal (0.7, 1) distribution and runs Bartlett's test on each set.

• A result of TRUE, where the test's p-value falls below 0.05, suggests a rejection of equal variances; this outcome is theoretically expected in 5 percent (or 100 out of 2,000) of the trials.

• Bartlett's test is observed to reject the null hypothesis of equal variances in 554 out of 1000 trials, amounting to 55.4 percent of the cases—significantly higher than the expected 5 percent.

• This excessive rejection rate indicates that Bartlett's test is overly sensitive and incorrect far more often than it should be when the data are non-normally distributed.

• Tests for equal variance are crucial for understanding key data attributes like precision (defined inversely to standard deviation) and how it varies between groups.

• Better alternatives to Bartlett's test are discussed in subsequent sections, which are more suitable for assessing changing variance among two or more groups and should be used in place of Bartlett's test. These tests are applicable to multi-group comparisons as well as two-group comparisons.

2.1 Fligner-Killeen Test for Equal Variance (Nonparametric)

• Out of the 56 tests evaluated by Conover and Iman (1981), the Fligner-Killeen test was found to be the most robust for unequal variance when data are non-normally distributed. It begins by computing the absolute value of the residuals (AVR) from each group median. For j =1 to k groups and i =1 to nj observations

$$AVR_{ij} = | x_{ij} - \text{median}_j | \qquad \text{Eq. (6)}$$

• The test then ranks the AVR and weights each rank to produce a set of scores. A linear-rank test (a nonparametric test of location) is computed on the scores. The null hypothesis is that the average score is the same in all groups, indicating that the variances are the same in all groups. ***The alternative hypothesis is that at least one group's variance differs.***

• The Fligner-Killeen test correctly does not find a difference in variance between the two lognormal groups generated from the same lognormal (0.7,1) distribution.

Example 5 Fligner-Killeen Test for Equal Variance (Nonparametric)

```
# Fligner-Killeen Test for Equal Variance (Nonparametric)

# Parameters for the lognormal distribution
mean_log = 0.7
sd_log = 1

# Generate two groups from the same lognormal distribution
group1 = np.random.lognormal(mean_log, sd_log, 2000)
group2 = np.random.lognormal(mean_log, sd_log, 2000)

# Combine the groups for testing
data = np.concatenate((group1, group2))
groups = ['group1']*2000 + ['group2']*2000

# Conduct the Fligner-Killeen test of homogeneity of variances
stat, p_value = fligner(group1, group2)

print("Fligner-Killeen test of homogeneity of variances")
print(f"data: data and groups")
print(f"Fligner-Killeen: med chi-squared = {stat:.5f}, p-value = {p_value:.4f}")
Fligner-Killeen test of homogeneity of variances
data: data and groups
Fligner-Killeen: med chi-squared = 0.17314, p-value = 0.6773
```

2.2 Levene's Test for Equal Variance (Parametric)

- Levene's test determines if the average distance from the median is the same in all groups.
- It assumes that data follow a normal distribution but is less sensitive to this assumption than Bartlett's test.
- The test loses power (an increase in the p-value) when applied to data that are not shaped like a normal distribution.

- Levene's test computes the Average Absolute Residual (AVR) for each observation and performs an analysis of variance (ANOVA) on these AVRs.

- ANOVA computed for only two groups is very similar to a t-test.

- The null hypothesis of Levene's test is that the average absolute residual is the same in all groups because the variance is the same in all groups.

- The alternative hypothesis is that at least one group's variance differs.

- Levene's test is commonly found in statistics software and is recommended for use in many guidance documents, including by the U.S. Environmental Protection Agency.

- Conover and others (1981) found that Levene's test performed better than other parametric tests of heteroscedasticity evaluated in their study.

- It appropriately does not find a difference in variance between two lognormal groups generated from the same distribution (p-value = 0.5017).

Example 6 Levene's Test for Equal Variance (Parametric)

```
# Eample 6 Levene's Test for Equal Variance (Parametric)
val_expl = np.random.lognormal(mean=0.7, sigma=1, size=100)  # 
Example data generation
group = np.array([1]*50 + [2]*50)  # Example grouping

# Performing Levene's Test for Homogeneity of Variances using median 
as the center
stat, p_value = levene(val_expl[group == 1], val_expl[group == 2], 
center='median')

# Printing the results
print("Levene's Test for Homogeneity of Variance (center = median)")
print(f"Group degrees of freedom: 1, F value: {stat:.4f}, p-value: 
{p_value:.4f}")
```

```
Levene's Test for Homogeneity of Variance (center = median)
Group degrees of freedom: 1, F value: 0.4547, p-value: 0.5017
```