Hypothesis Test 01

EN5423 | Spring 2024

w05_hypothesis_01.pdf
(Week 5)

Contents

1 CLASSIFICATION OF HYPOTHESIS 1

1.1 Classification Based on Measurement Scales 2

1.2 Divisions Based on the Method of Computing a p-value 3

2 STRUCTURE OF HYPOTHESIS TESTS 5

2.1 Choose the Appropriate Test 5

2.2 Establish the Null and Alternate Hypotheses 7

2.3 Decide on an Acceptable Type I Error Rate, *A* 8

2.4 Compute the Test Statistic and the p-value 9

1) Understanding α-Value (Significance Level) *9*

2) Interpreting P-Values *10*

3) Decision Making *10*

4) Why Reporting P-Values is Useful *10*

2.5 Make the Decision to Reject *H*₀ or Not 11

3 THE RANK-SUM TEST AS AN EXAMPLE OF HYPOTHESIS TESTING 11

$H_0: Prob(he \geq innocent) = 0.5$

$H_A: Prob(he \leq innocent) < 0.5$

$H_A: Prob(he \geq innocent) > 0.5$

$H_A: Prob(he \geq innocent) \neq 0.5$

Hypothesis Tests Intro

Why Scientists Collect Data:
• Scientists gather data to understand the systems they study better. Before collecting data, they often have hypotheses—predictions about how these systems behave.
• The main goal of collecting data is to see if these hypotheses are true based on the data.

Role of Statistical Tests:
• Statistical tests help us figure out if the patterns we see in the data (like differences between groups or relationships between variables) are real or just by chance.
• They give us a structured way to make these decisions.

Testing Hypotheses with Data Comparisons:
• In environmental research, scientists have long compared data, like water quality in different aquifers, to see if there are significant differences.
• Historically, some of these comparisons were based on expert opinion rather than formal tests. But, using hypothesis tests has big advantages:

1) Consistency: They ensure that anyone analyzing the data in the same way will get the same results. This makes the findings reliable and verifiable by others.
2) Quantitative Evidence: They provide a number (the p-value) that shows how strong the evidence against the hypothesis is. This helps in making informed decisions about whether to reject a hypothesis, while considering the risk of being wrong.

Introduction to Hypothesis Testing:
• We start by explaining the basics of hypothesis testing, including when and how to use it.
• The rank-sum test is an example we use to show how hypothesis tests work and how p-values are calculated.
• We also talk about testing for normality (checking if data follows a normal distribution), which is important for many statistical tests.
• This chapter sets the stage with key concepts and terms we will use throughout our discussions on this course.

1 Classification of Hypothesis

Picking the correct hypothesis test from the many options available can be confusing. However, tests can be organized into five types based on the kind of data you are dealing with (**Figure 1**). There are three main categories of hypothesis tests: parametric, nonparametric, and permutation tests. Each type calculates p-values differently, which are key to understanding the test results.

• **What Guides the Choice of Test?** The nature of your data and what you aim to discover in your study will help decide which category and specific test to use.

• **Understanding Variables:** In our discussions, we will often mention two types of variables:

1) Response Variable: This is what we are interested in studying. In regression analysis, it is also known as the dependent or *y* variable. It is the outcome we are trying to understand or predict.

2) Explanatory Variable: Also called the independent variable, this one helps explain changes in the response variable. For instance, when we are looking at differences between two groups, the explanatory variable tells us which group a particular data point comes from.

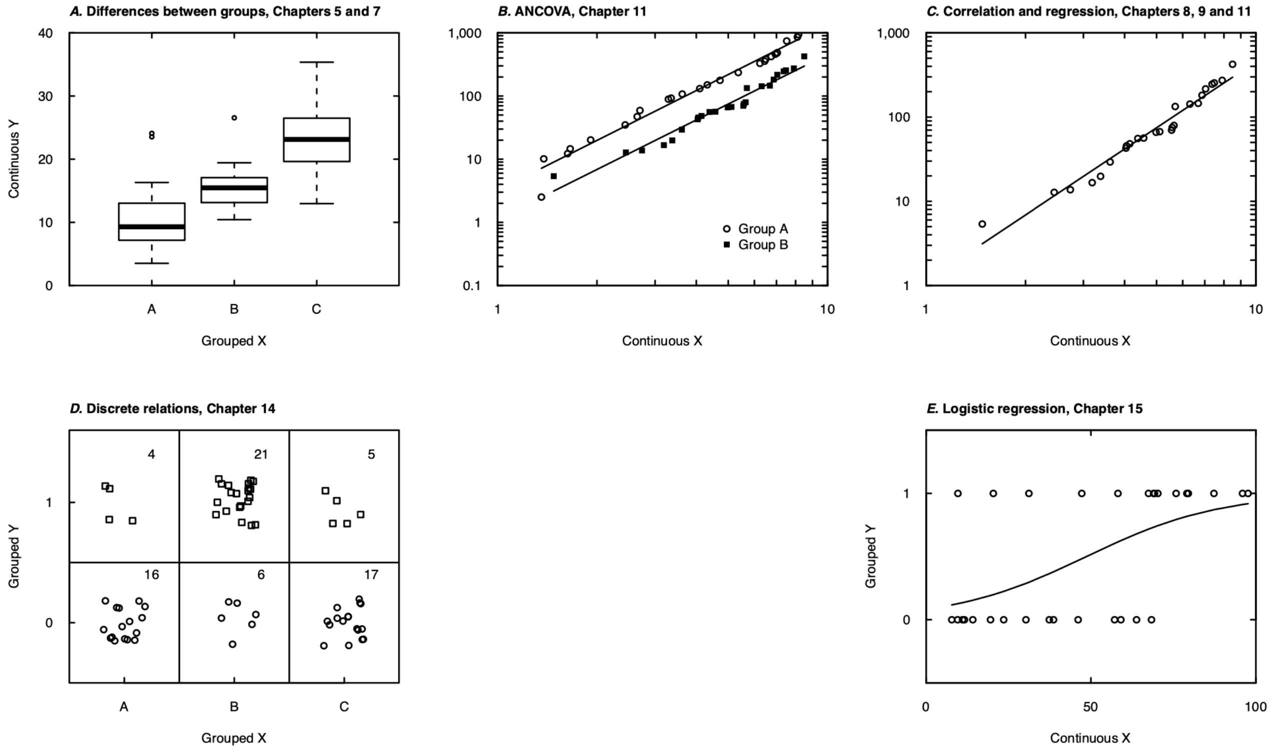


Figure 1. Five classes of hypothesis tests. (A) Differences between groups, (B) analysis of covariance (ANCOVA), (C) correlation and regression, (D) discrete relations, and (E) logistic regression.

1.1 Classification Based on Measurement Scales

• Five Types of Statistical Tests:

1. **Based on Variable Types:** The five classes of tests are differentiated by the types of variables involved—whether they are continuous or categorical.
   - **Continuous variables** are those that can take any value within a range (like streamflow or concentration).
   - **Categorical variables** are those that fall into distinct groups (like type of aquifer or land-use group).
2. **Tests for Continuous Response Variables:**
   - The first three classes (A-C in **Figure 1**) focus on continuous response variables. These are used when we are measuring something that varies over a range, like water quality or flow rate.
   - **Boxplots (A)** illustrate tests comparing the central tendency (like the average) of a continuous response variable across different groups defined by categorical variables.
   - **Graph (C)** relates to linear regression and correlation, showing how one continuous variable affects another.

3. **Combining Continuous and Categorical Variables:**
   - **Graph (B),** or analysis of covariance (ANCOVA), combines elements of both, examining how a continuous response variable is influenced by both continuous and categorical explanatory variables.
4. **Tests for Categorical Response Variables:**
   - In contrast, graphs D and E focus on cases where the response variable is categorical.
   - **Graph (D)** is about understanding associations between two categorical variables, often using contingency tables (similar to that for use of t-tests or analysis of variance; ANOVA).
   - **Graph (E)** involves modeling probabilities (like logistic regression) where the response is categorical but can be influenced by both types of variables.

Key Points:

- The type of statistical test to use depends on whether your variables are continuous, categorical, or a mix of both.
- **Continuous Variables:** Think of things that can vary smoothly, where analysis often involves comparing averages or looking at relationships.
- **Categorical Variables:** These are about putting data into distinct buckets, and the analysis might focus on counting how many fall into each category or the likelihood of being in a category based on other factors.
- Tests are chosen based on the nature of your data and what you are trying to find out, helping to make sense of complex environmental data in a structured way.

1.2 Divisions Based on the Method of Computing a p-value

• Hypothesis tests come in three main types: parametric, nonparametric, and permutation tests.

• Each type calculates *p*-values differently, crucial for understanding test results. Before using any test, it is essential to check its assumptions.

Parametric tests assume:

- Data follow a specific distribution (often normal).
- Data are independent and identically distributed (not correlated over time).
- When comparing groups, they have equal variance (adjustments are available if not).

These tests work well when their assumptions match the data. They excel in modeling and estimating, especially in complex designs, thanks to easily computed parameters like mean or variance. However, they lose power if the data do not meet these assumptions.

• Nonparametric tests do not assume a data distribution shape. They use data ranks to answer questions about frequencies and are suited for simpler analyses. **These tests require data to be independent and randomly sampled**, but they can handle a wider variety of data shapes than parametric tests.

• Permutation tests calculate *p*-values by simulating thousands of outcomes under the null hypothesis. They are versatile, not assuming a normal distribution, and are useful for comparing means or other statistics. Like nonparametric tests, they assume random data.

• Parametric analysis of variance (ANOVA) studies more complex than what we cover in this text can be designed and tested in ways not yet possible for nonparametric or permutation methods. Parametric multiple-regression equations can model more complex situations than what methods from the other two divisions currently accomplish.

Key Points:

- **Parametric tests** are powerful for complex models but require specific data distributions.
- **Nonparametric tests** are versatile, using data ranks for simpler questions, without needing a specific distribution.
- **Permutation tests** simulate outcomes to assess statistical significance, useful across various hypotheses, without distribution assumptions.

Independent and Identically Distributed (i.i.d)

The term "data are independent and identically distributed" (often abbreviated as i.i.d.) refers to two key assumptions about the data used in statistical analyses:

Independent: This means that the value of one observation does not influence or is not influenced by the value of another observation. For example, the outcome of one coin toss does not affect the outcome of the next coin toss.

Identically Distributed: This means that all observations come from the same probability distribution and thus share the same probability structure. For instance, if we are tossing a fair coin, every toss has an identical probability distribution: 50% chance of heads and 50% chance of tails.

When data are said to be “not correlated over time,” it emphasizes the independence part of the assumption, particularly in time series data or sequential measurements. It means that the value of a current observation does not depend on the values of previous observations.

These assumptions are crucial for many statistical models because they underlie the theoretical foundations that allow for the derivation of statistical properties and inference methods. If these assumptions are violated, the results of statistical tests and models may not be reliable.

A common example of a situation where data are not independent and identically distributed (not i.i.d.) is in financial time series, such as stock prices or exchange rates. In these cases, two main issues can arise:

Not Independent: The value of a financial asset at one time point is often influenced by its value at previous time points. This means there is a correlation over time, which violates the independence assumption. For example, if a stock's price has been rising steadily over the past few days, it might continue to rise in the short term due to momentum.

Not Identically Distributed: The volatility (variance of returns) of financial assets can change over time. There might be periods of high volatility (like during a financial crisis) and periods of low volatility (in stable economic conditions). This variability means the probability distribution of the asset's returns changes over time, violating the identically distributed assumption.

In such not i.i.d. scenarios, standard statistical methods that assume independence and identical distribution may not be appropriate, and more sophisticated models designed to handle time series data, like ARIMA models for forecasting or GARCH models for volatility, are used instead. These models account for the correlation between observations and the changing volatility over time.

2 Structure of Hypothesis Tests

All hypothesis tests follow the same six steps, which are discussed in the following sections:

1. Choose the appropriate test and review its assumptions.

2. Establish the null and alternative hypotheses, H_0 and H_A.

3. Decide on an acceptable error rate, α.

4. Compute the test statistic from the data.

5. Compute the p-value.

6. Reject the null hypothesis if $p \leq \alpha$; do not reject if $p > \alpha$.

2.1 Choose the Appropriate Test

• Choose test methods based on what your data looks like and what you want to find out. The first thing to consider is the type of data you have, as shown in **Figure 1**.

• The second thing is what you aim to achieve with the test. You can use tests to compare average values between two or more groups, look at how spread out the data groups are, or see how two or more variables relate to each other.

• For instance, if you want to compare the average values of two different groups, you could use a two-sample t-test, a rank-sum test, or a two-sample permutation test (refer to **Table 1**).

• It is also important to decide if the average value is better represented by the mean (like the balance point) or the median (the middle value). Later, I will guide you through different test goals, offering various options for each.

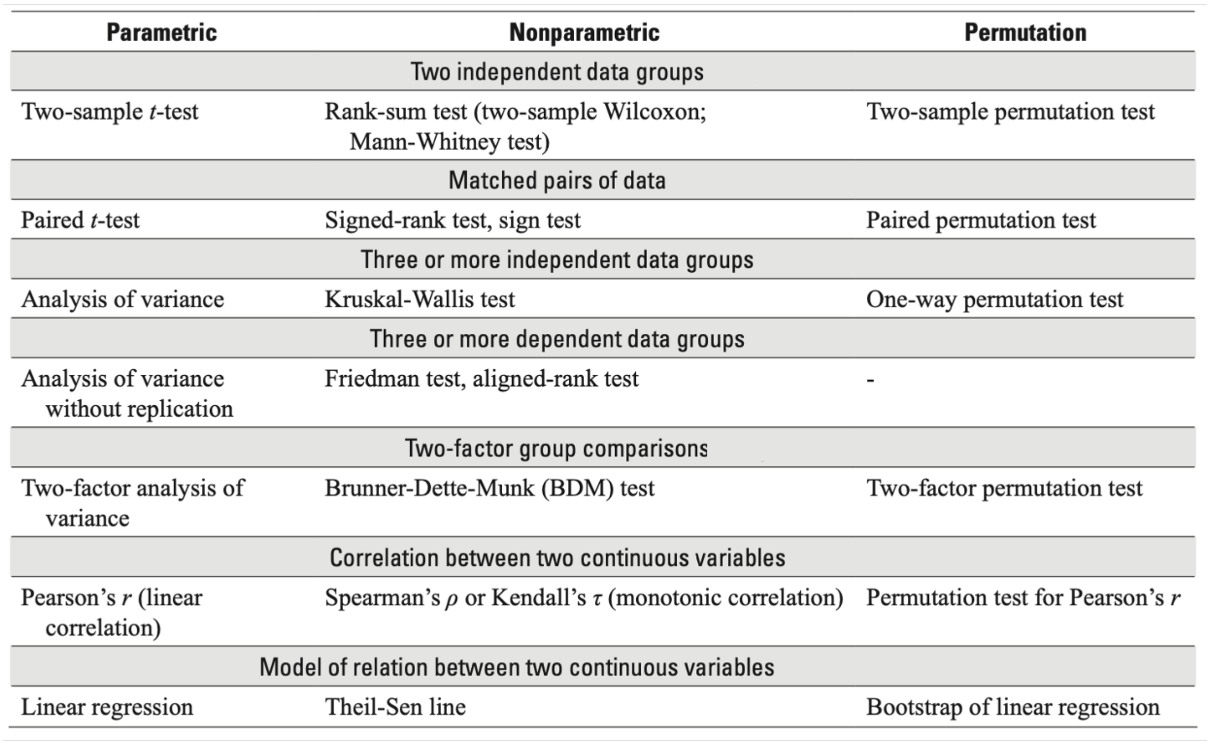
• Choosing between parametric, nonparametric, or permutation tests hinges on your research goals and data characteristics.

- **Parametric tests** are suitable when you believe your data follow a normal distribution. These tests are centered around the mean, ideal for studies interested in aggregates or totals, like total pollution levels in a river or total exposure to a contaminant over time. The mean serves well when summing up data is key.

- **Nonparametric tests** do not require your data to follow any specific distribution. They focus on median values or frequency statistics, making them useful for identifying common or typical differences across groups. **If you are looking at whether one group generally has higher values than another**, nonparametric tests are relevant because they examine differences in ranks or frequencies.

- When dealing with skewed data or outliers, **parametric tests may not effectively detect differences**, potentially leading to errors or a loss of test power. In such cases, **permutation tests** are preferable for assessing mean differences. They offer a robust alternative by directly testing the observed data through simulation, avoiding the pitfalls of assuming a particular data distribution.

Table 1. Guide to the classification of some hypothesis tests with continuous response variables.



Parametric	Nonparametric	Permutation
	Two independent data groups	
Two-sample *t*-test	Rank-sum test (two-sample Wilcoxon; Mann-Whitney test)	Two-sample permutation test
	Matched pairs of data	
Paired *t*-test	Signed-rank test, sign test	Paired permutation test
	Three or more independent data groups	
Analysis of variance	Kruskal-Wallis test	One-way permutation test
	Three or more dependent data groups	
Analysis of variance without replication	Friedman test, aligned-rank test	-
	Two-factor group comparisons	
Two-factor analysis of variance	Brunner-Dette-Munk (BDM) test	Two-factor permutation test
	Correlation between two continuous variables	
Pearson's r (linear correlation)	Spearman's ρ or Kendall's τ (monotonic correlation)	Permutation test for Pearson's r
	Model of relation between two continuous variables	
Linear regression	Theil-Sen line	Bootstrap of linear regression

• **Further Considerations**:

- The **Central Limit Theorem (CLT)** argument supports parametric tests by suggesting that sample means approximate a normal distribution in large samples, regardless of the original data distribution. This is contingent on the data's skewness and the sample size. However, in real-world applications like water resources, data often deviate from symmetry, necessitating larger samples for the CLT to apply effectively. In contexts with asymmetric distributions or insufficient sample sizes, relying on the **CLT can lead to significant power loss in detecting true differences.**

- **Parametric tests** are criticized for their susceptibility to differences in variance and the influence of outliers on means, which can obscure true differences. The Welch-Satterthwaite correction is a modification that helps but at the expense of test power.

- **Nonparametric tests**, while not directly testing means, do not inherently require equal variances across groups. They excel in scenarios with data skewness or outliers by focusing on medians or ranks, offering higher power in detecting differences under such conditions.

- **Robustness** in statistical testing refers to a test's ability to maintain type I error rates, not necessarily its power to detect true effects under non-ideal conditions, such as skewed data distributions.

- The choice between testing methods also depends on the degree of data distribution normality and the specific hypotheses being tested. For skewed distributions or when dealing with outliers, nonparametric or permutation tests often provide greater detection power.

- **Test selection** should prioritize methods with greater power for the anticipated data characteristics. When facing asymmetric data or outliers, consider permutation tests for their higher power in detecting true mean differences. Comparisons between parametric and nonparametric methods can be informed by their relative efficiencies, especially in large samples.

Key Points: Your test choice must align with your research objectives and the nature of your data. While parametric tests may offer simplicity and are robust under certain conditions, nonparametric and permutation tests provide flexibility and power in analyzing data with non-standard distributions or anomalies. Always ensure your methodological choice is informed by both the statistical properties of your data and the specific questions you aim to answer.

2.2 Establish the Null and Alternate Hypotheses

• Before gathering or analyzing data, it is crucial to define the null and alternative hypotheses. These hypotheses clearly outline the study's goals and help maintain focus, avoiding bias influenced by data observations or desired outcomes.

• The **null hypothesis (*H_0*)** assumes the default state of affairs before any data collection. It typically suggests ***no change or difference***, asserting that any observed variations are due to chance.

• For instance, an environmental engineer might hypothesize that water wells upstream and downstream from a waste site have equal contaminant levels, regardless of their hope for the outcome.

• The **alternative hypothesis (*H_A* or *H_1*)** represents what we might conclude if the evidence suggests the null hypothesis is unlikely. This can either directly oppose the null hypothesis or ***specify a particular direction of difference***.

• Alternative hypotheses can lead to one-sided or two-sided tests, depending on whether any deviation (one-sided) or a specific direction of deviation (two-sided) from the null hypothesis would be considered significant.

- **Two-sided tests** are used when differences in any direction from the null hypothesis could validate the alternative hypothesis. For example, finding that the actual 100-year flood level is significantly different from what was designed, either higher or lower, would support the alternative hypothesis.

- **One-sided tests** are applied when only variations in one specific direction from the null hypothesis would support the alternative hypothesis. For example, if the concern is only whether the 100-year flood level exceeds the design level, as this would require infrastructure upgrades, then the test is one-sided. Data showing the flood level is below the design would support the null hypothesis.

• Choosing between a one-sided and two-sided test depends on the initial study objectives. It is inappropriate to decide the type of test after observing the data; ***this decision should be made based on the specific interests and hypotheses at the start of the study***. It is not appropriate to look at the data, find that group A is considerably larger in value than group B, and perform a one-sided test that group A is larger. This would be ignoring the real possibility that had group B been larger there would have been interest in that situation as well.

• Examples in water resource management include 1) testing for reduced flood levels after dam construction, 2) lower nutrient levels following wastewater treatment improvements, or 3) increased contaminant levels near suspected pollution sources.

2.3 Decide on an Acceptable Type I Error Rate, α

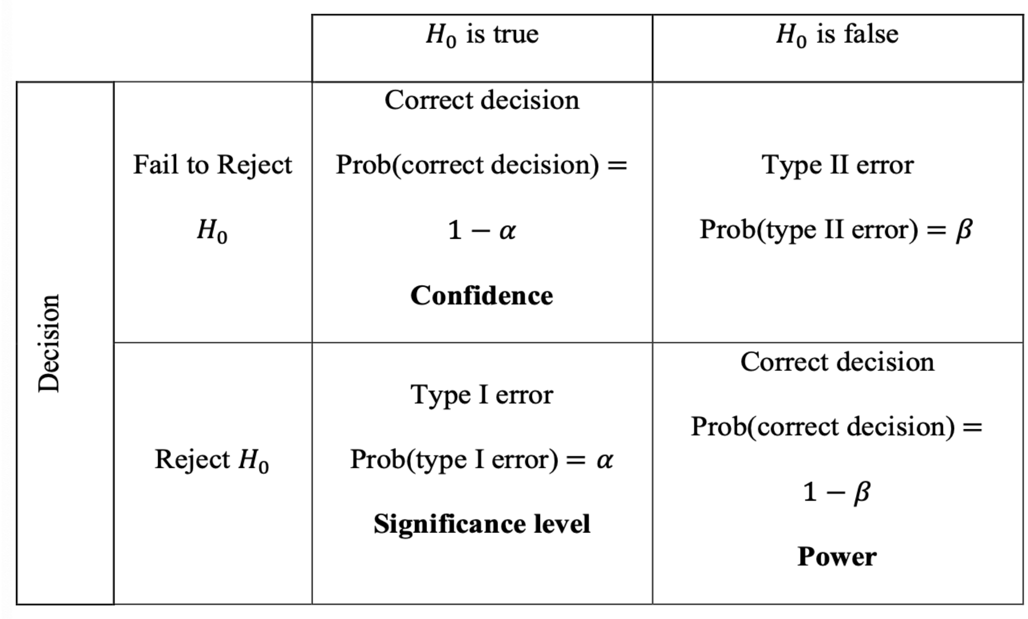
• The α-value, or significance level, is the chance of mistakenly rejecting the null hypothesis (saying there is a change or difference when there is not one). Commonly set at 5% (0.05) or 1% (0.01), this value represents the acceptable risk of a Type I error (a false positive) to the researcher.

• However, different situations may call for different α-values. For example, if rejecting the null hypothesis could lead to costly actions, like an expensive cleanup, a lower α (e.g., 0.01) might be chosen to reduce the risk of unnecessary expenditure. Conversely, in a preliminary study sorting sites into categories for further investigation, a higher α (e.g., 0.10 or 0.20) might be used to ensure no potentially significant sites are overlooked.

• Minimizing α to avoid Type I errors might seem ideal, but setting α too low can increase the risk of Type II errors (false negatives), where a real effect is missed. The balance between these types of errors is crucial. The probability of a Type II error is represented by β, and the test's power, or its ability to detect a true effect, is 1 - β. To reduce the chances of both Type I and Type II errors, researchers can:

1. Increase the sample size, which enhances the study's ability to detect true differences or effects.
2. Choose the most powerful test for their data type, ensuring the test is sensitive enough to detect real differences.

In water quality studies, for example, ***failing to identify contamination due to a test with low power can have serious implications***. This underscores the importance of carefully considering sample size and test selection, especially when dealing with data characteristics like outliers or skewness that could affect test accuracy.



		H_0 is true	H_0 is false
Decision	Fail to Reject H_0	Correct decision Prob(correct decision) = $1 - \alpha$ **Confidence**	Type II error Prob(type II error) = β
	Reject H_0	Type I error Prob(type I error) = α **Significance level**	Correct decision Prob(correct decision) = $1 - \beta$ **Power**

Figure 2. Four possible results of hypothesis testing.

2.4 Compute the Test Statistic and the p-value

• Test statistics sum up the data to help decide if the null hypothesis (H_0) stands. If the test statistic does not greatly differ from expected values under H_0, we keep H_0. But if it is unlikely under H_0, we reject H_0. The *p*-value tells us how rare the test statistic is if H_0 is true.

• The *p*-value is the chance of seeing the calculated statistic, or a more extreme one, if H_0 is true. It shows how strong the evidence against H_0 is in the data. **A smaller *p*-value means stronger evidence against H_0**. Unlike α-levels, which are pre-set risks of a Type I error (false positive) acceptable to the researcher, *p*-values depend on the actual data and show the evidence's strength.

• For example, in testing if wells downstream have higher pollution than upstream, deciding to act on $\alpha = 0.01$, means you are okay with a 1% risk of unnecessary action. Reporting *p*-values lets others with different risk tolerances decide on their actions. A *p*-value of 0.02 might sway someone okay with a 5% risk to act, whereas with $\alpha = 0.01$, they might not.

Easier concepts of *α*-values, *p*-values, and decision-making in hypothesis testing.

Imagine we are conducting a study to determine if wells located downstream of a potential pollution source have higher levels of contaminants than wells upstream. The objective is to decide whether remediation efforts are necessary based on the contamination levels.

1) Understanding α-Value (Significance Level)

- The ***α*-value** (or significance level) is predetermined before the analysis. It represents the threshold of risk the researcher is willing to accept for making a Type I error, which in this context is falsely concluding that downstream wells are more contaminated when they are not.

- Choosing ***α* = 0.01** means the researcher has set a very stringent criterion for evidence against the null hypothesis (H0: there is no difference in contamination levels between upstream and downstream wells). In practical terms, there's a willingness to accept a 1% chance of mistakenly initiating costly remediation efforts.

2) Interpreting P-Values

- The **p-value** is the probability, given the data, of observing a test statistic as extreme as, or more extreme than, the one observed if the null hypothesis is true.
- A ***p*-value of 0.02** suggests that there is a 2% chance of observing the detected difference in contamination levels (or a more significant difference) if, in reality, there is no difference. This is evidence against the null hypothesis but is not strong enough to meet the stringent α = 0.01 threshold.

3) Decision Making

- If you're the decision-maker who set $\alpha = 0.01$, the p-value of 0.02 does not meet your criteria for action; it's not low enough to conclude significant contamination difference, given your low tolerance for error.
- However, another stakeholder with a higher tolerance for error (willing to accept a 5% risk, or $\alpha = 0.05$) might view the p-value of 0.02 as sufficient evidence to warrant action. This stakeholder perceives the risk of false alarm (unnecessary cleanup) as less critical than the risk of overlooking a real contamination issue.

4) Why Reporting P-Values is Useful

- Reporting p-values alongside test results provides a nuanced understanding of the evidence against the null hypothesis. It allows individuals or stakeholders with different risk tolerances to make informed decisions based on the same data but through the lens of their own acceptable error levels.
- This flexibility is particularly valuable in scenarios where the consequences of Type I errors (false positives) and Type II errors (false negatives) have different implications for action, costs, and further investigation.

In summary, the *α*-value sets a threshold for action based on the researcher's risk tolerance, while the *p*-value offers a measure of the evidence against the null hypothesis. Decision-making then involves comparing the p-value with the α-value, taking into account the context and consequences of potential errors.

P-values for nonparametric tests can be found in three ways:

1. **Exact test:** Provides precise *p*-values by comparing the test statistic against all possible outcomes for the sample sizes. Previously requiring extensive tables, now software handles this until sample sizes get too big. Exact tests give the most accurate results for small samples.
2. **Large-sample approximation (LSA):** For large datasets, it approximates *p*-values to save time, assuming the test statistic's distribution mirrors a common one like chi-square or normal. This method is used for efficiency and when data ties make exact tests unfeasible, though it doesn't assume the data itself follows the distribution.

3. **Permutation test:** Useful for very large datasets, this method samples possible outcomes to estimate the *p*-value. It is seen as more accurate than LSAs for approximating *p*-values without assuming a specific distribution for the test statistic.

These methods reflect different strategies to calculate *p*-values, each with its context for use, especially in terms of sample size and data characteristics.

2.5 Make the Decision to Reject *H*₀ or Not

• When the *p*-value is less than or equal to the decision criteria (the *α*-level), H_0 is rejected.

• When the *p*-value is greater than *α*, H_0 is not rejected. The null hypothesis is never accepted or proven to be true, it is assumed to be true until proven otherwise and is not rejected when there is insufficient evidence to do so.

• In short, reject H_0 when the *p*-value ≤ *α*. ***However, it is good practice to report the p-values*** for those that may want to make decisions with a different significance level and type I error rate.

3 The Rank-sum Test as an Example of Hypothesis Testing (***HW5 #1***)

Core Exercise:

1) Suppose that aquifers X and Y are sampled to determine whether the concentrations of a contaminant in the aquifers are similar or different. This is a test for differences in location or central value and will be covered in detail later class.

Two samples, x_i, are taken from aquifer *X* (n = 2), and five samples, y_i, from aquifer *Y* (m = 5) for a total of seven samples (N = n + m = 7). Also suppose that there is a prior reason (that likely motivated the sampling) to believe that *X* values tend to be lower than *Y* values: aquifer *X* is deeper and likely to be uncontaminated. The null hypothesis (H_0) and alternative hypothesis (H_A) of this one-sided test are as follows:

H_0: x_i and y_i are samples from the same distribution, or

$H_0: Prob(x_i \geq y_i) = 0.5, i = 1,2,\ldots,n; j = 1,2,\ldots,m$

H_A: x_i is from a distribution that is generally lower than that of y_i, or

$H_A: Prob(x_i \geq y_i) < 0.5$

2) Remember that in one-sided tests, such as the one being discussed, data indicating differences opposite in direction to the alternative hypothesis (H_A)—for example, x_i frequently being larger than y_i when H_A suggests x_i should be smaller—are considered supporting evidence for the null hypothesis (H_0). In one-sided tests, we are only interested in departures from H_0 in one specific direction as defined by H_A.

3) Having established the null (H_0) and alternative hypotheses (H_A), it is crucial to set an acceptable **Type I error** probability, *α*. This process is akin to the principle used in a court of law, where innocence is presumed (here, assuming that concentrations between groups are equivalent) unless evidence strongly suggests otherwise—specifically, that Aquifer *Y* has

higher concentrations. This evidence must demonstrate that the observed differences are unlikely to have arisen by chance alone, a determination guided by the chosen significance level, α, which effectively sets our threshold for 'reasonable doubt'.

Evidence and Decision Threshold: By setting α, researchers establish a threshold for what constitutes sufficient evidence to reject H_0 in favor of H_A. The lower the α, the ***stronger the evidence must be to conclude that differences observed are not due to chance***, paralleling the legal standard of "beyond a reasonable doubt".

4) When the primary interest is comparing mean concentrations between two groups, the *t*-test is a common choice. However, a fundamental assumption of the *t*-test is that the data in each group are normally distributed. In cases where sample sizes are very small—such as two and five in this scenario—conducting a reliable test for normality becomes impractical due to the limited power to detect departures from normality.

Given the objective to determine whether concentrations in one group are higher than in the other, and considering the small sample sizes, ***a nonparametric approach becomes more suitable***. The ***rank-sum test***, also known as the ***Mann-Whitney U*** test, is an appropriate alternative. This nonparametric test ***does not assume normality and is designed to compare the central tendencies of two independent samples based on the ranks of their combined data***.

5) To perform the rank-sum test, all observations from both groups are ranked together from the lowest to the highest concentration. Each value is assigned a rank from 1 (lowest concentration) to 7 (highest concentration) in this case, since there are seven values in total.

The ranks for the observations in the smaller group (let's call this Group *X*, with a sample size of two) are then summed to obtain a statistic known as *W*. This rank sum, *W*, serves as the basis for the ***exact test*** to assess whether there is a statistically significant difference in concentrations between the two groups.

By using the rank-sum test, ***we avoid the pitfalls of assuming normality with small samples*** and can still test the hypothesis that concentrations in one group are higher than in the other in a robust manner.

6) Next, *W* would be computed and compared to a table of test statistic quantiles to determine the *p*-value. Where do these tables come from? We will derive the table for sample sizes of two and five as an example.

What are the possible values *W* may take, given that the null hypothesis is true? The collection of all the possible outcomes of *W* defines its distribution, and therefore composes the table of rank-sum test statistic quantiles. Shown below are all the possible combinations of ranks of the two *x* values.

1,2	1,3	1,4	1,5	1,6	1,7
	2,3	2,4	2,5	2,6	2,7
		3,4	3,5	3,6	3,7
			4,5	4,6	4,7
				5,6	5,7
					6,7

If H_0 is true, each of the 21 possible outcomes must be equally likely. That is, it is just as likely for the two *x*s to be ranks 1 and 2, or 3 and 5, or 1 and 7, and so on. Each of the outcomes results in a value of W, the sum of the two ranks. The 21 W values corresponding to the above outcomes are:

3	4	5	6	7	8
	5	6	7	8	9
		7	8	9	10
			9	10	11
				11	12
					13

The expected value of W is the mean (and in this case, also the median) of the above values, or 8. Given that each outcome is equally likely when H_0 is true, the probability of each possible W value is listed in **Table 2** where probability is expressed as a fraction. For example, of 21 W values, one is equal to three, therefore, the probability that $W = 3$ is 1/21.

Table 2. Probabilities and one-sided *p*-values for the rank-sum test with $n = 2$ and $m = 5$.

W	Rescaled W from Python	Prob(W)	Prob($\leq W$)
3	0	1/21	0.048
4	1	1/21	0.095
5	2	2/21	0.190
6	3	2/21	0.286
7	4	3/21	0.429
8	5	3/21	0.571
9	6	3/21	0.714
10	7	2/21	0.810
11	8	2/21	0.905
12	9	1/21	0.952
13	10	1/21	1.00

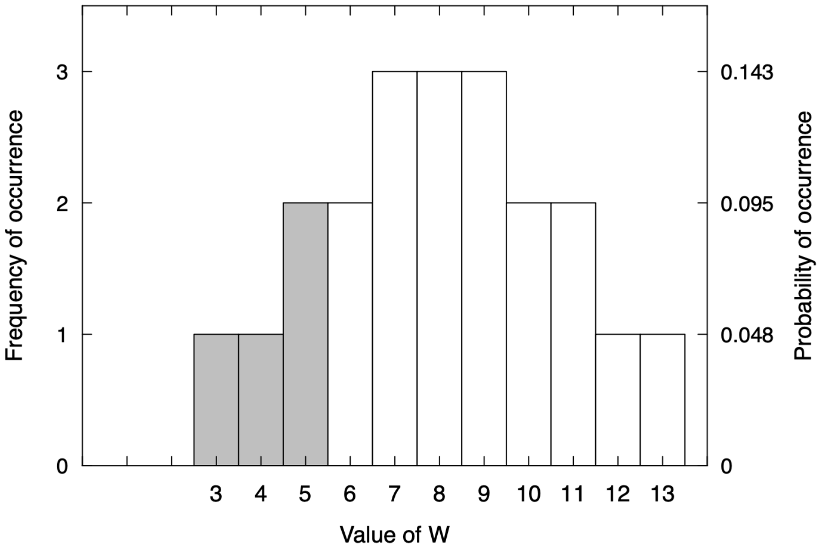


Figure 3. Probabilities of occurrence for a rank-sum test with sample sizes of 2 and 5. The *p*-value for a one-sided test equals the area shaded.

What if the data collected produced two x values having ranks 1 and 4? Then W would be 5 (or 2 using Python, as in **Table 2**), lower than the expected value $E[W] = 8$. If H_A were true rather than H_0, W would tend toward low values. What is the probability that W would be as low as 5, or lower, if H_0 were true? It is the sum of the probabilities for W =3, 4, and 5, or 4/21 = 0.190 (see **Figure. 4**). ***This number is the p-value for the test statistic of 5***.

It says that the chance of a departure from $E[W]$ of at least this magnitude occurring when H_0 is true is 0.190, which is not uncommon (about 1 chance in 5). **Thus, the evidence against H_0 is not too convincing**. If the ranks of the two *x*s had been 1 and 2, then W = 3 (0 using Python) and the p-value would be 1/21= 0.048. This result is much less likely than the previous case but is still about 5 percent. ***In fact, owing to such a small sample size the test can never result in a highly compelling case for rejecting H_0.***

7) The one-sided rank-sum test performed in Python using Mann-Whitney U test in *scipy*, on randomly generated data looks like the following:

Python example

```
import numpy as np
from scipy.stats import mannwhitneyu

# Set the random seed to ensure reproducibility
np.random.seed(100)

# Generate random normal variables
x = np.random.normal(loc=40, scale=5, size=2)
y = np.random.normal(loc=50, scale=5, size=5)

# Perform the one-sided Mann-Whitney U test (equivalent to the
Wilcoxon rank-sum test)
# Note: In scipy’s mannwhitneyu function, use of the alternative
parameter
# 'less' indicates a one-sided test where the hypothesis is that x
has a tendency
# to have smaller values than y.
u_statistic, p_value = mannwhitneyu(x, y, alternative='less')

print(f"U-statistic: {u_statistic}, P-value: {p_value}")
>> U-statistic: 0.0, P-value: 0.047619047619047616
alternative hypothesis: true location shift is less than 0
```

8) The above example has considered only the one-sided *p*-value, which is appropriate when there is some prior notion that *X* tends to be smaller than *Y* (or the reverse). Quite often, the situation is that there is no prior notion of which should be lower. In this case a two-sided test must be done. The two-sided test has the same null hypothesis as was stated above, but H_A is now that x_i and y_i are from different distributions, or

$$H_A\text{: Prob }(x_i \geq y_i) \neq 0.5$$

Suppose that W for the two-sided test were found to be 5. The p-value equals the probability that W will differ from $E[W]$ by this much or more, in either direction (**see Figure. 4**). It is

$$\text{Prob } (W \leq 5) + \text{Prob } (W \geq 11)$$

Where did the 11 come from? It is just as far from $E[W] = 8$ as is 5. The two-sided p-value therefore equals 8/21=0.381, twice the one-sided p-value. Symbolically we could state

$$\text{Prob } (|W - E[W]| \geq 3) = 8/21$$

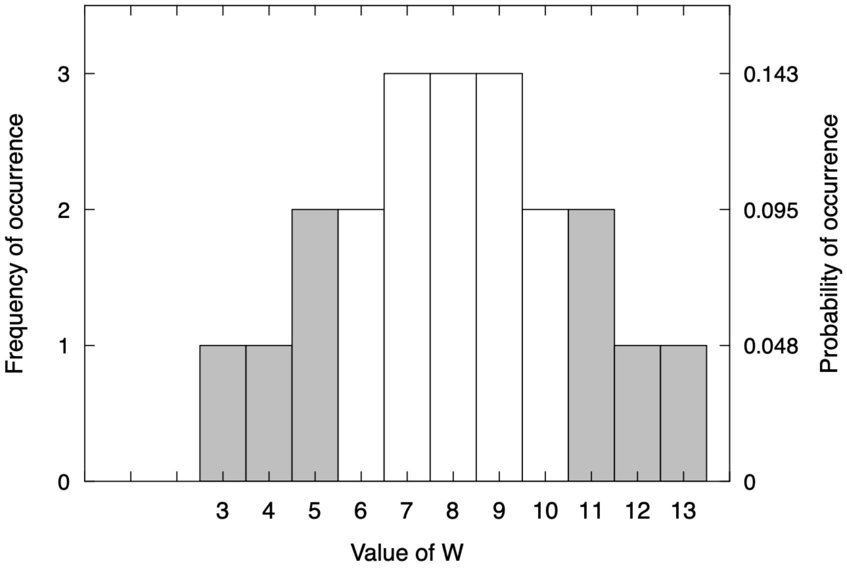


Figure 4. Probabilities of occurrence for a rank-sum test with sample sizes of 2 and 5. The p-value for a two-sided test equals the area shaded.

The two-sided rank-sum test performed in Python using Mann-Whitney U test in *scipy*, on the same randomly generated data as the previous example but with a different alternative (H_A) yields:

Python example

```
u_statistic, p_value = mannwhitneyu(x, y, alternative='two-sided')

print(f"U-statistic: {u_statistic}, P-value: {p_value}")
>> U-statistic: 0.0, P-value: 0.09523809523809523
alternative hypothesis: true location shift is not equal to 0
```

This example used a symmetric distribution, but the test can also be used with asymmetric distributions, in which case the probabilities in the two tails would differ. Fortunately, modern statistical software can handle symmetric or asymmetric distributions for us and reports two-sided p-values as the default, with a user option to select a one-sided alternative. For the alternative group A > group B, the option in R is alternative = ‘greater’ for wilcox.test. For the reverse, alternative = ‘less’ is the option.

9) To summarize, p-values describe the probability of calculating a test statistic as extreme or more extreme as the one observed, if H_0 were true. The lower the p-value the stronger the case against the null hypothesis.

Now, let us look at an α-level approach. Return to the original problem, the case of a one-sided test. Assume α is set equal to 0.1. This corresponds to a critical value for W, call it W*, such that Prob(W ≤W*) = α. Whenever W ≤W*, H_0 is rejected with no more than a 0.1 frequency of error if H_0 were always true.

However, because W can only take on discrete integer values, as seen above, a W* which exactly satisfies the equation is not usually available; instead the largest possible W* such that Prob(W ≤W*) ≤ α is used. Searching **Table 2** for possible W values and their probabilities, W* = 4 because Prob(W ≤4) = 0.095 ≤0.1. If α = 0.09 had been selected then W* would be 3.

For a two-sided test a pair of critical values, W_U^* and W_L^*, are needed, where:

$$\text{Prob}\ (W \leq W_L^*) + \text{Prob}\ (W \geq W_u^*) \leq \alpha \text{ and}$$

$$W_U^* - E[W] = E[W] - W_L^*$$

These upper and lower critical values of W are symmetrical around E[W] such that the probability of W falling on or outside of these critical levels is as close as possible to α, without exceeding it, under the assumption that H_0 is true. In the case at hand, if α = 0.1, then $W_L^* = 3$ and $W_U^* = 13$ because:

$$\text{Prob}\ (W \leq 3) + \text{Prob}\ (W \geq 13) = 0.048 + 0.048 = 0.095 \leq 0.1$$

10) Note that for a two-sided test, the critical values are farther from the expected value than in a one-sided test at the same α-level. It is important to recognize that p-values are also influenced by sample size. For a given magnitude of difference between the x and y data, and a given amount of variability in the data, p-values will tend to be smaller when the sample size is large. In the extreme case where vast amounts of data are available, it is a virtual certainty that p-values will be small even if the differences between x and y are what might be called of no practical significance.