

Virtual Laboratories > 9. Hypothesis Testing > **1** 2 3 4 5 6 7

1. Introduction

The Basic Statistical Model

As usual, our starting point is a **random experiment** with an underlying **sample space** and a **probability measure** \mathbb{P} . In the basic statistical model, we have an observable **random variable** X taking values in a set S . In general, X can have quite a complicated structure. For example, if the experiment is to sample n objects from a population and record various measurements of interest, then

$$X = (X_1, X_2, \dots, X_n)$$

where X_i is the vector of measurements for the i^{th} object. The most important special case occurs when (X_1, X_2, \dots, X_n) are independent and identically distributed. In this case, we have a **random sample** of size n from the common distribution.

The purpose of this section is to define and discuss the basic concepts of statistical **hypothesis testing**. Collectively, these concepts are sometimes referred to as the **Neyman-Pearson** framework, in honor of **Jerzy Neyman** and **Egon Pearson**, who first formalized them.

General Hypothesis Tests

A **statistical hypothesis** is a statement about the distribution of the data variable X . Equivalently, a statistical hypothesis specifies a *set* of possible distributions of X (namely, the set of distributions for which the statement is true). In **hypothesis testing**, the goal is to see if there is sufficient statistical evidence to reject a presumed **null hypothesis** in favor of a conjectured **alternative hypothesis**. The null hypothesis is usually denoted H_0 while the alternative hypothesis is usually denoted H_1 . A hypothesis that specifies a single distribution for X is called **simple**; a hypothesis that specifies more than one distribution for X is called **composite**.

An hypothesis test is a *statistical decision*; the conclusion will either be to **reject** the null hypothesis in favor of the alternative, or to **fail to reject** the null hypothesis. The decision that we make must, of course, be based on the data vector X . Thus, we will find a subset R of the sample space S and reject H_0 if and only if $X \in R$. The set R is known as the **rejection region** or the **critical region**. Note the asymmetry between the null and alternative hypotheses. This asymmetry is due to the fact that we *assume* the null hypothesis, in a sense, and then see if there is sufficient evidence in X to overturn this assumption in favor of the alternative.

Often, the critical region is defined in terms of a statistic $W(X)$, known as a **test statistic**. As usual, the use of a statistic allows **data reduction** when the dimension of the statistic is much smaller than the dimension of the data vector.

Errors

The ultimate decision may be correct or may be in error. There are two types of errors, depending on which of the hypotheses is actually true:

1. A **type 1 error** is rejecting the null hypothesis when it is true.
2. A **type 2 error** is failing to reject the null hypothesis when it is false.

Similarly, there are two ways to make a *correct* decision: we could reject the null hypothesis when it is false or we could fail to reject the null hypothesis when it is true. The possibilities are summarized in the following table:

State/Decision	Fail to reject H_0	Reject H_0
H_0 True	Correct	Type 1 error
False	Type 2 error	Correct

If H_0 is true (that is, the distribution of X is specified by H_0), then $\mathbb{P}(X \in R)$ is the probability of a type 1 error for this distribution. If H_0 is composite, then H_0 specifies a variety of different distributions for X and thus there is a *set* of type 1 error probabilities. The *maximum* probability of a type 1 error is known as the **significance level** of the test or the **size** of the critical region, which we will denote by α . Usually, the rejection region is constructed so that the significance level is a prescribed, small value (typically 0.1, 0.05, 0.01).

If H_1 is true (that is, the distribution of X is specified by H_1), then $\mathbb{P}(X \notin R)$ is the probability of a type 2 error for this distribution. Again, if H_1 is composite then H_1 specifies a variety of different distributions for X , and thus there will be a set of type 2 error probabilities. Generally, there is a tradeoff between the type 1 and type 2 error probabilities. If we reduce the probability of a type 1 error, by making the rejection region R smaller, we necessarily increase the probability of a type 2 error because the complementary region $S \setminus R$ is larger.

Power

If H_1 is true (that is, the distribution of X is specified by H_1), then $\mathbb{P}(X \in R)$, the probability of rejecting H_0 (and thus making a correct decision), is known as the **power** of the test for the distribution.

Suppose that we have two tests, corresponding to rejection regions R_1 and R_2 , respectively, each having significance level α . The test with region R_1 is **uniformly more powerful** than the test with region R_2 if

$$\mathbb{P}(X \in R_1) \geq \mathbb{P}(X \in R_2) \text{ for any distribution of } X \text{ specified by } H_1$$

Naturally, in this case, we would prefer the first test. Often, however, two tests will not be uniformly ordered; one test will be more powerful for some distributions specified by H_1 while the other test will be more powerful for other distributions specified by H_1 . Finally, if a test has significance level α and is uniformly more powerful than any other test with significance level α , then the test is said to be a **uniformly most powerful test** at level α . Clearly, such a test is the best we can do.

***P*-value**

In most cases, we have a general procedure that allows us to construct a test (that is, a rejection region R_α) for any given significance level $\alpha \in (0, 1)$. Typically, R_α decreases (in the subset sense) as α decreases. In this context, the ***P*-value** of the data variable X , denoted $P(X)$ is defined to be the smallest α for which $X \in R_\alpha$; that is, the smallest significance level for which H_0 is rejected, given X . Knowing $P(X)$ allows us to test H_0 at any significance level, for the given data: If $P(X) \leq \alpha$ then we would reject H_0 at significance level α ; if $P(X) > \alpha$ then we fail to reject H_0 at significance level α . Note that $P(X)$ is a *statistic*.

Tests of an Unknown Parameter

Hypothesis testing is a very general concept, but an important special class occurs when the distribution of the data variable X depends on a parameter θ , taking values in a parameter space Θ . The parameter may be vector-valued, so that $\theta = (\theta_1, \theta_2, \dots, \theta_k)$ and $\Theta \subseteq \mathbb{R}^k$ for some k . The hypotheses generally take the form

$$H_0: \theta \in \Theta_0 \text{ versus } H_1: \theta \notin \Theta_0$$

where Θ_0 is a prescribed subset of the parameter space Θ . In this setting, the probabilities of making an error or a correct decision depend on the true value of θ . If R is the rejection region, then the **power function** is given by

$$Q(\theta) = \mathbb{P}_\theta(X \in R), \quad \theta \in \Theta$$

1. Show that

- $Q(\theta)$ is the probability of a type 1 error when $\theta \in \Theta_0$.
- $\max \{Q(\theta) : \theta \in \Theta_0\}$ is the significance level of the test.

2. Show that

- $1 - Q(\theta)$ is the probability of a type 2 error when $\theta \notin \Theta_0$.
- $Q(\theta)$ is the power of the test when $\theta \notin \Theta_0$.

Suppose that we have two tests, corresponding to rejection regions R_1 and R_2 , respectively, each having significance level α . The test with rejection region R_1 is uniformly more powerful than the test with rejection

region R_2 if

$$Q_1(\theta) \geq Q_2(\theta), \quad \theta \notin \Theta_0$$

Most hypothesis tests of an unknown real parameter θ fall into three special cases:

1. $H_0: \theta = \theta_0$ versus $H_1: \theta \neq \theta_0$
2. $H_0: \theta \geq \theta_0$ versus $H_1: \theta < \theta_0$
3. $H_0: \theta \leq \theta_0$ versus $H_1: \theta > \theta_0$

where θ_0 is a specified value. Case 1 is known as the **two-sided test**; case 2 is known as the **left-tailed test**, and case 3 is known as the **right-tailed test** (named after the conjectured alternative). There may be other unknown parameters besides θ (known as **nuisance parameters**).

Equivalence Between Hypothesis Test and Confidence Sets

There is an equivalence between hypothesis tests and **confidence sets** for a parameter θ .

3. Suppose that $C(X)$ is a $1 - \alpha$ level confidence set for θ . Show that the test below has significance level α for the hypothesis $H_0: \theta = \theta_0$ versus $H_1: \theta \neq \theta_0$:

$$\text{Reject } H_0 \text{ if and only if } \theta_0 \notin C(X)$$

equivalently, we *fail* to reject H_0 at significance level α if and only if θ_0 is in the corresponding $1 - \alpha$ level confidence set.

4. In particular, show that this equivalence applies to interval estimates of a real parameter θ and the common tests for θ . In each case below, the confidence interval has confidence level $1 - \alpha$ and the test has significance level α

- a. Suppose that $(L(X), U(X))$ is a two-sided confidence interval for θ . Reject $H_0: \theta = \theta_0$ versus $H_1: \theta \neq \theta_0$ if and only if $\theta_0 \leq L(X)$ or $\theta_0 \geq U(X)$
- b. Suppose that $L(X)$ is a confidence lower bound for θ . Reject $H_0: \theta \geq \theta_0$ versus $H_1: \theta < \theta_0$ if and only if $\theta_0 \leq L(X)$
- c. Suppose that $U(X)$ is a confidence upper bound for θ . Reject $H_0: \theta \leq \theta_0$ versus $H_1: \theta > \theta_0$ if and only if $\theta_0 \geq U(X)$

Pivot Variables and Test Statistics

Recall that confidence sets of an unknown parameter θ are often constructed through a **pivot variable**, that is, a random variable $W(X, \theta)$ that depends on the data vector X and the parameter θ . but whose distribution does not depend on θ . In this case, a natural test statistic is $W(X, \theta_0)$.

[Virtual Laboratories](#) > [9. Hypothesis Testing](#) > **1** [2](#) [3](#) [4](#) [5](#) [6](#) [7](#)

[Contents](#) | [Applets](#) | [Data Sets](#) | [Biographies](#) | [External Resources](#) | [Key words](#) | [Feedback](#) | [©](#)