9.1: Introduction to Hypothesis Testing

Basic Theory

Preliminaries

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 $\boldsymbol{X}$ taking values in a set S. In general, $\boldsymbol{X}$ 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

$$\boldsymbol{X} = (X_1, X_2, \ldots, X_n) \tag{9.1.1}$$

where X_i is the vector of measurements for the ith object. The most important special case occurs when $(X_1, X_2, \ldots, X_n)$ 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.

Hypotheses

A *statistical hypothesis* is a statement about the distribution of $\boldsymbol{X}$. Equivalently, a statistical hypothesis specifies a *set* of possible distributions of $\boldsymbol{X}$: the set of distributions for which the statement is true. A hypothesis that specifies a single distribution for $\boldsymbol{X}$ is called *simple*; a hypothesis that specifies more than one distribution for $\boldsymbol{X}$ is called *composite*.

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.

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 observed value $\boldsymbol{x}$ of the data vector $\boldsymbol{X}$. Thus, we will find an appropriate subset R of the sample space S and reject H_0 if and only if $\boldsymbol{x} \in R$. 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 $\boldsymbol{x}$ to overturn this assumption in favor of the alternative.

An hypothesis test is a statistical analogy to proof by contradiction, in a sense. Suppose for a moment that H_1 is a statement in a mathematical theory and that H_0 is its negation. One way that we can prove H_1 is to assume H_0 and work our way logically to a contradiction. In an hypothesis test, we don't "prove" anything of course, but there are similarities. We assume H_0 and then see if the data $\boldsymbol{x}$ are sufficiently at odds with that assumption that we feel justified in rejecting H_0 in favor of H_1.

Often, the critical region is defined in terms of a statistic $w(\boldsymbol{X})$, known as a *test statistic*, where w is a function from S into another set T. We find an appropriate rejection region $R_T \subseteq T$ and reject H_0 when the observed value $w(\boldsymbol{x}) \in R_T$. Thus, the rejection region in S is then $R = w^{-1}(R_T) = \{\boldsymbol{x} \in S : w(\boldsymbol{x}) \in R_T\}$. As usual, the use of a statistic often allows significant *data reduction* when the dimension of the test 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.

Types of errors:

1. A *type 1 error* is rejecting the null hypothesis H_0 when H_0 is true.
2. A *type 2 error* is failing to reject the null hypothesis H_0 when the alternative hypothesis H_1 is true.

Similarly, there are two ways to make a *correct* decision: we could reject H_0 when H_1 is true or we could fail to reject H_0 when H_0 is true. The possibilities are summarized in the following table:

Hypothesis Test

State | Decision	Fail to reject H_0	Reject H_0
H_0 True	Correct	Type 1 error
H_1 True	Type 2 error	Correct

Of course, when we observe $\boldsymbol{X} = \boldsymbol{x}$ and make our decision, either we will have made the correct decision or we will have committed an error, and usually we will never know which of these events has occurred. *Prior* to gathering the data, however, we can consider the probabilities of the various errors.

If H_0 is true (that is, the distribution of $\boldsymbol{X}$ is specified by H_0), then $\mathbb{P}(\boldsymbol{X} \in R)$ 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 $\boldsymbol{X}$ and thus there is a *set* of type 1 error probabilities.

The maximum probability of a type 1 error, over the set of distributions specified by H_0, is the *significance level* of the test or the *size* of the critical region.

The significance level is often denoted 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 $\boldsymbol{X}$ is specified by H_1), then $\mathbb{P}(\boldsymbol{X} \notin R)$ 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 $\boldsymbol{X}$, 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.

The extreme cases can give us some insight. First consider the decision rule in which we *never* reject H_0, regardless of the evidence $\boldsymbol{x}$. This corresponds to the rejection region $R = \emptyset$. A type 1 error is impossible, so the significance level is 0. On the other hand, the probability of a type 2 error is 1 for any distribution defined by H_1. At the other extreme, consider the decision rule in which we always rejects H_0 regardless of the evidence $\boldsymbol{x}$. This corresponds to the rejection region $R = S$. A type 2 error is impossible, but now the probability of a type 1 error is 1 for any distribution defined by H_0. In between these two worthless tests are meaningful tests that take the evidence $\boldsymbol{x}$ into account.

Power

If H_1 is true, so that the distribution of $\boldsymbol{X}$ is specified by H_1, then $\mathbb{P}(\boldsymbol{X} \in R)$, the probability of rejecting H_0 is the *power* of the test for that distribution.

Thus the power of the test for a distribution specified by H_1 is the probability of making the correct decision.

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

$$\mathbb{P}(\boldsymbol{X} \in R_1) \geq \mathbb{P}(\boldsymbol{X} \in R_2) \text{ for every distribution of } \boldsymbol{X} \text{ specified by } H_1 \tag{9.1.2}$$

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.

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 a uniformly most powerful 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.

The P-*value* of the observed value $\boldsymbol{x}$ of $\boldsymbol{X}$, denoted $P(\boldsymbol{x})$, is defined to be the smallest α for which $\boldsymbol{x} \in R_\alpha$; that is, the smallest significance level for which H_0 is rejected, given $\boldsymbol{X} = \boldsymbol{x}$.

Knowing $P(\boldsymbol{x})$ allows us to test H_0 at any significance level for the given data $\boldsymbol{x}$: If $P(\boldsymbol{x}) \leq \alpha$ then we would reject H_0 at significance level α; if $P(\boldsymbol{x}) > \alpha$ then we fail to reject H_0 at significance level α. Note that $P(\boldsymbol{X})$ is a *statistic*. Informally, $P(\boldsymbol{x})$ can often be thought of as the probability of an outcome "as or more extreme" than the observed value $\boldsymbol{x}$, where *extreme* is interpreted relative to the null hypothesis H_0.

Analogy with Justice Systems

There is a helpful analogy between statistical hypothesis testing and the criminal justice system in the US and various other countries. Consider a person charged with a crime. The presumed *null hypothesis* is that the person is innocent of the crime; the conjectured *alternative hypothesis* is that the person is guilty of the crime. The test of the hypotheses is a trial with evidence presented by both sides playing the role of the data. After considering the evidence, the jury delivers the decision as either *not guilty* or *guilty*. Note that *innocent* is not a possible verdict of the jury, because it is not the point of the trial to *prove* the person innocent. Rather, the point of the trial is to see whether there is sufficient evidence to overturn the null hypothesis that the person is innocent in favor of the alternative hypothesis of that the person is guilty. A *type 1 error* is convicting a person who is innocent; a *type 2 error* is acquitting a person who is guilty. Generally, a type 1 error is considered the more serious of the two possible errors, so in an attempt to hold the chance of a type 1 error to a very low level, the standard for conviction in serious criminal cases is *beyond a reasonable doubt*.

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 $\boldsymbol{X}$ depends on a parameter θ taking values in a parameter space Θ. The parameter may be vector-valued, so that $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$ and $\Theta \subseteq \mathbb{R}^k$ for some $k \in \mathbb{N}_+$. The hypotheses generally take the form

$$H_0 : \theta \in \Theta_0 \text{ versus } H_1 : \theta \notin \Theta_0 \tag{9.1.3}$$

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* Q is given by

$$Q(\theta) = \mathbb{P}_\theta(\boldsymbol{X} \in R), \quad \theta \in \Theta \tag{9.1.4}$$

The power function gives a lot of information about the test.

The power function satisfies the following properties:

1. $Q(\theta)$ is the probability of a type 1 error when $\theta \in \Theta_0$.
2. $\max\{Q(\theta) : \theta \in \Theta_0\}$ is the significance level of the test.
3. $1 - Q(\theta)$ is the probability of a type 2 error when $\theta \notin \Theta_0$.
4. $Q(\theta)$ is the power of the test when $\theta \notin \Theta_0$.

If we have two tests, we can compare them by means of their power functions.

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 $Q_1(\theta) \geq Q_2(\theta)$ for all $\theta \notin \Theta_0$.

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

Suppose that θ is a real parameter and $\theta_0 \in \Theta$ a specified value. The tests below are respectively the *two-sided test*, the *left-tailed test*, and the *right-tailed test*.

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 $H_1 : \theta > \theta_0$

Thus the tests are named after the conjectured alternative. Of course, 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.

Suppose that $C(\boldsymbol{x})$ is a $1-\alpha$ level confidence set for θ. The following test has significance level α for the hypothesis $H_0: \theta = \theta_0$ versus $H_1: \theta \neq \theta_0$: Reject H_0 if and only if $\theta_0 \notin C(\boldsymbol{x})$

Proof

By definition, $\mathbb{P}[\theta \in C(\boldsymbol{X})] = 1 - \alpha$. Hence if H_0 is true so that $\theta = \theta_0$, then the probability of a type 1 error is $P[\theta \notin C(\boldsymbol{X})] = \alpha$.

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. In particular, this equivalence applies to interval estimates of a real parameter θ and the common tests for θ given above.

In each case below, the confidence interval has confidence level $1-\alpha$ and the test has significance level α.

1. Suppose that $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ 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 $\theta_0 < L(\boldsymbol{X})$ or $\theta_0 > U(\boldsymbol{X})$.
2. Suppose that $L(\boldsymbol{X})$ is a confidence lower bound for θ. Reject $H_0: \theta \leq \theta_0$ versus $H_1: \theta > \theta_0$ if and only if $\theta_0 < L(\boldsymbol{X})$.
3. Suppose that $U(\boldsymbol{X})$ is a confidence upper bound for θ. Reject $H_0: \theta \geq \theta_0$ versus $H_1: \theta < \theta_0$ if and only if $\theta_0 > U(\boldsymbol{X})$.

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 $W(\boldsymbol{X}, \theta)$ that depends on the data vector $\boldsymbol{X}$ and the parameter θ, but whose distribution does not depend on θ and is known. In this case, a natural test statistic for the basic tests given above is $W(\boldsymbol{X}, \theta_0)$.