

A. Review

A **hypothesis** H_0 is an assertion about a (population) parameter. If the parameter is θ , then H_0 is of the form:

$$H_0 \Leftrightarrow \theta \in \Theta_0,$$

where Θ_0 is a subset of the set of all values that the parameter may have.

A **test** of H_0 is a statement T_0 about a sample \vec{X} of the form:

$$T_0 \Leftrightarrow \vec{X} \in A.$$

Here, A is a subset of sample space; it is called the acceptance region. When T_0 is true, we say that the test result is “accept H_0 .” The complement of A is called the rejection region. When \vec{X} lies in this region, we say the test result is “reject H_0 .”

Comment. The test is dependent on the sample, \vec{X} . Of course, the test only yields a decision when some specific data \vec{x} are inserted. But we are often interested in how the test behaves under repeated applications. For example, the *power function* of T_0 is

$$\beta(\theta) := P_\theta(\vec{X} \notin A),$$

the probability of rejecting H_0 expressed as a function of θ . (We will study β below.) Note that in the definition of β , the test itself is treated as a θ -indexed family of discrete random variables. If θ is fixed, then each of the two possible values of T_0 (“accept” or “reject”) has a probability, and these two numbers sum to 1.

Error types:

Type I: reject a true H_0 . (H_0 is true, T_0 is false.)

Type II: accept a false H_0 . (H_0 is false, T_0 is true.)

B. A Classical Example

Sir R. A. Fisher’s *Lady Tasting Tea* is a famous didactic example. A lady at a party asserts that she is capable of distinguishing the order in which milk and tea are poured into a cup by taste alone. The statisticians at the party devise an experiment. The lady is given 8 cups, four of each kind. The cups are presented in random order, and the lady is asked to taste them and divide them into two groups of four according to kind. (Note that there are $\binom{8}{4} = 70$ ways to divide the cups, and only one correct way.)

Problem. Describe a test of the null hypothesis, “The lady has no ability to distinguish.” What parameter is being tested? What is the sample space? What would be a reasonable rejection region? What is the probability of a Type I error?

Solution. The unknown parameter is the presence or absence of a special talent. H_0 is taken to be the hypothesis that the lady has no ability to detect the order of the milk and tea. The 70-element set of possible classifications of cups is the sample space. The

rejection region consists of the correct classification of the eight cups. When H_0 true, each of the 70 points in sample space has equal probability. Thus, the chances of a Type I error (rejecting the null when in fact it is true) is $1/70$. The probability of a Type II error is dependent on additional assumptions. (For example, if the talent does not come in shades or degrees, but is absolute if present, then Type II errors cannot occur.)*

The test in the solution will produce a Type I error (i.e., attribute to the lady a skill she does not have) with probability $1/70 = 0.014\dots$. This is far less than 5% that is often taken as the allowable Type I error rate. In many cases, statisticians set a rate for Type I errors before designing the test, and then seek to beat that rate. What explains this approach? The null hypothesis is often associated with a “no effect” or “no difference” interpretation, in which case to make a Type I error is to assert an effect that really isn’t there. This might be more embarrassing to a researcher than a Type II error, accepting the null when the alternative is actually the case, or *not* recognizing an effect that *does* exist.

C. New Terminology: significance level, size, power function.

If an upper bound for the Type I error probability is imposed, it is called the *level* (or *significance level*) of the test. The test that we subjected the lady to was a level .05 test. (It was also a level .02 test.) The *size* of a test is the maximum probability of a Type I error. The size of the tea test was $1/70 = 0.014\dots$. The letter α is often used for the size or the level.

The *power function* $\beta(\theta)$ of a test is the probability that a sample is in the rejection region.

Comment. I don’t know the origin of the terminology, but here is my speculation. An instrument such as a telescope or a microscope is said to have greater power if it is capable of detecting more distant, faint or tiny objects. Such sensitivity often comes with increased danger of detecting something that is not really there. As we have said, the null hypothesis is traditionally the assertion that no effect is present, and rejecting the null amounts to “detecting something.” A statistical test of high power is more likely to detect a faint signal (reject the null when appropriate), but may also more likely to see stuff that’s not there (reject the null when not appropriate).

Some quick exercises.

i) Show that

$$\beta(\theta) = \begin{cases} \text{probability of a Type I error,} & \text{if } \theta \in \Theta_0; \\ 1 - (\text{probability of a Type II error}), & \text{if } \theta \notin \Theta_0. \end{cases}$$

* This example illustrates why statisticians sometimes wish to avoid speaking of “accepting the null hypothesis.” If the lady missed a cup, it would not be common sense to conclude that she had no ability at all. So, I take this opportunity to emphasize again that our mathematical framework—where we either “accept” or “reject”—is adopted for the purpose of mathematical analysis, and not as a model for practical statistical consulting work. We are using the words “accept” and “reject” to refer to the two possible test results. We could have called these: “do not reject” versus “reject,” or anything else, e.g., “don’t publish” versus “publish.”

- ii) What's a better test: one with smaller significance level or one with larger significance level?
- iii) Verify that the terminology as I introduced it above has the same meaning as in 8.3.5 and 8.3.6: a test has size α if $\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha$, and it has a significance level better than α if $\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha$.

D. Example. Testing the Normal. (See pages 384-5.)

Suppose we wish to test

$$H_0 \Leftrightarrow \theta < \theta_0 \Leftrightarrow \theta_0 - \theta > 0,$$

using a sample X_1, \dots, X_n from a normal(θ, σ^2) population, where σ^2 is known. We choose a test with a rejection region $\bar{X} > K$, for fixed constant K . An appropriate value of K will be found later, after we have stated the properties we want the test to have. We will use $\beta(\theta) = P_\theta(\bar{X} > K)$. For practical computational reasons, it is better to express $\beta(\theta)$ as a Z -score (Z standard normal). Noting that \bar{X} has a normal($\theta, \sigma^2/n$) distribution, we have:

$$\begin{aligned} \beta(\theta) &= P_\theta(\bar{X} > K) \\ &= P\left(\frac{\bar{X} - \theta}{\sigma/\sqrt{n}} > \frac{K - \theta}{\sigma/\sqrt{n}}\right) \\ &= P\left(Z > \frac{K - \theta}{\sigma/\sqrt{n}}\right) \\ &= P\left(Z > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) \text{ with } c = \frac{K - \theta_0}{\sigma/\sqrt{n}}. \end{aligned}$$

If H_0 is true, then $\theta_0 - \theta > 0$. Thus, the significance level of the test (which depends only on the probability of error when H_0 is true) can be adjusted by selecting c . Indeed, if we wish the significance level to be at most α , we choose a test with size α , but size is

$$\sup_{\theta < \theta_0} \beta(\theta) = P(Z > c).$$

So, pick c large enough that $P(Z > c) < \alpha$. Having achieved the desired level, we still want to minimize the probability of Type II errors. We can achieve this by adjusting n . (To be continued.)