Two-tailed Test

A two-tailed test is a statistical procedure used to compare the null hypothesis that a population parameter is equal to a particular value against the alternative hypothesis that the population parameter is *different* from this value. Evidence regarding the null hypothesis is obtained from a test statistic, and the test is said to be “two-tailed” because its alternative hypothesis does not specify whether the parameter is greater than or less than the value specified by the null hypothesis. Hence, both large and small values of the test statistic, this is, values on both tails of its distribution, provide evidence against the null hypothesis. This type of test is relevant for situations in which researchers wish to test a null hypothesis but they do not have a prior belief about the direction of the alternative, a situation which is likely to happen in practice. The term “two-tailed test” is usually reserved for the particular case of one-dimensional hypotheses, even though it may be used more generally.

Two-sided hypothesis testing

In hypothesis testing, the hypotheses are always statements about a population parameter which partitions the set of possible values that the parameters may take. For example, letting μ be the parameter for which the hypothesis test is performed, a null hypothesis, referred to as H_0, may be defined as

$$H_0 : \mu = \mu_0$$

and its two-sided alternative hypothesis, referred to as H_1, is defined as

$$H_1 : \mu \neq \mu_0$$

The alternative hypothesis H_1 does not make a statement about whether μ is greater than μ_0 or less than μ_0, which makes this a two-sided test. The difference between a one-sided test and a two-sided test lies solely in the specification of the alternative hypothesis. As a consequence, while a one-sided test specifies in its alternative hypothesis that the parameter is either greater than or less than the value specified in the null hypothesis (H_1 is *either* $\mu > \mu_0$ *or* $\mu < \mu_0$), in a two-sided test the direction of the alternative hypothesis is left unspecified.

Evidence for or against the null hypothesis is obtained by means of a test statistic, which is a function of the available data. Just as in the one-sided case, in a two-sided hypothesis test the decision of whether to reject the null hypothesis H_0 is based on a test statistic $W(X) = W(X_1, X_2, \ldots, X_N)$ which is a function of a (random) sample $X_1, X_2, \ldots, X_N$ of size N from the population under study. The test specifies a rejection rule that indicates in what situations H_0 should be rejected. In a two-sided test, rejection occurs for both large and small values of $W(X)$, while in a one-sided test rejection occurs either for large or small values of the test statistic (but not both) as dictated by the alternative hypothesis. Formally, a two-sided rejection rule is defined as:

$$\text{Reject } H_0 \qquad \text{if } W(X) < c_1 \text{ or } W(X) > c_2$$

Do not reject H_0 if $c_1 \leq W(X) \leq c_2$

In order to establish the values of the critical values c_1 and c_2, it is common practice to follow the Neyman-Pearson approach and first choose a significance level α. The significance level α of the test is an upper bound to the probability of mistakenly rejecting H_0 when H_0 is true (probability of type I error). Once the significance level has been fixed, the constants c_1 and c_2 are chosen so that the probability of rejecting H_0 when H_0 is true is (at most) equal to the significance level. In other words, c_1 and c_2 are chosen so that

$$\Pr_{H_0}(W(X) < c_1) + \Pr_{H_0}(W(X) > c_2) \leq \alpha$$

where $\Pr_{H_0}(z)$ indicates the probability of z computed assuming that the null hypothesis H_0 is true.

This still may leave the constants c_1 and c_2 undetermined, since there may be infinitely many ways in which the sum of these two terms can be made equal to α. Thus, the researcher must usually make a decision regarding how to divide the probability α between the two terms, this is, between the two tails of the distribution of $W(X)$ under H_0. If the researcher has no prior information regarding the direction of the alternative, then it seems appropriate to divide this total probability symmetrically between the two tails. This is, the condition $\Pr_{H_0}(W(X) < c_1) = \Pr(W(X) > c_2)$ is imposed and therefore

$$\Pr_{H_0}\left(W\left(X\right) < c_1\right) = \Pr_{H_0}\left(W\left(X\right) > c_2\right) \leq \frac{\alpha}{2}$$

If the researcher has prior information regarding the population parameter that may affect the alternative hypothesis, then this total probability may be divided asymmetrically between the two tails. However, an asymmetric allocation of α between both tails is not used very often, since in cases when information regarding the direction of the effect under study is available, researchers usually choose a one-sided alternative.

The two-sided rejection rule is easier to construct when the distribution of $W\left(X\right)$ under the null hypothesis is symmetric, since in this case the critical values c_1 and c_2 are equal in absolute value. In this case there is only one unknown constant that needs to be established based on the underlying distribution of the test statistic.

Comparison with one-sided test

The difference in the specification of the alternative hypothesis between a one-tailed test and a two-tailed test has important conceptual consequences. As illustrated in the example below, using a two-sided test is generally conservative in the sense that it is more difficult to reject the null hypothesis with this test than with the correct one-sided test for a given significance level. This occurs because a more extreme value of the test statistic will be necessary to reject the null hypothesis at the same α significance level with a two-sided test than with a one-sided test, due to the fact that in the former the total probability of rejecting H_0 when it is true (type I error) is split between both tails of the distribution of $W(X)$.

For example, when the null hypothesis H_0 is tested using both an α-level one-sided test to the right and an α-level two-sided test, and the distribution of the test statistic is continuous, the critical value c^* of the one-sided test is defined by $\Pr_{H_0}\left(W(X) > c^*\right) = \alpha$, and the critical values c_l^{**} and c_u^{**} of the two-sided test are defined by $\Pr_{H_0}\left(W(X) < c_l^{**}\right) + \Pr_{H_0}\left(W(X) > c_u^{**}\right) = \alpha$. It is easy to see that in this case $c^* < c_u^{**}$ and there exist values of $W(X)$ such that $c^* < W(X) < c_u^{**}$. When this happens, H_0 will be rejected with the one-sided test but will not be rejected with the two-sided test.

This point is further illustrated in Figure 1, where the top panel shows the significance level of a one-sided hypothesis test for the particular case of a normal distribution of the test statistic under H_0, and the bottom panel shows a two-sided test with the same significance level and the same test statistic, where the significance level has been split symmetrically across both tails. As can be seen in the figure, for all values of $W(X)$ between 1.64 and 1.96, the null hypothesis is rejected with a one-sided test but is *not* rejected with a two-sided test. The two-sided test requires a larger value of $W(X)$ to reject H_0 than the one-sided test shown in the figure, because the probability of type I error on the upper tail is forced to be smaller in the two-sided test ($\alpha / 2$) than in the one-sided test (α). This illustrates how a two-sided test may require a more surprising value of $W(X)$ to reject the null hypothesis than a one-sided test, which makes the two-sided test more conservative.

[Two-tailed_Test_Figure_1 about here]

A numerical example

Imagine a situation in which a researcher is interested in establishing whether two competing math text books have the effect of increasing the mathematical skills of elementary school students. In particular, the researcher is interested in whether assigning the practice exercises of the books as homework has en effect on math test scores. For this purpose, N students are randomly assigned to two different groups, referred to as group A and group B, of size N_A and N_B, respectively. Students in group A are assigned the exercises in book A as homework over the course of a month, and students in group B are assigned the exercises in book B as homework during the same period of time. Students solve the exercises individually and are not allowed to interact with one another. The researcher is interested in establishing whether children who are assigned the exercises in one book perform better in a math exam at the end of the experiment than children assigned the exercises in the other book, but based on the available information, the researcher has no prior belief as to which book is more effective than the other. In this case, a two-sided hypothesis test is appropriate, since the alternative hypothesis should be left unspecified.

Students are given a math exam at the beginning and at the end of the experiment, and the change in test scores is recorded for each student. Assuming that the difference in test scores is approximately normally distributed with means μ_A and μ_B in groups A and B, respectively, and equal variance, the mean differential effect of the two types of exercises can be analyzed using a two-sided difference-in-means test to determine whether $\mu_A - \mu_B$ is different from zero. Formally, the null and alternative hypotheses are formulated as follows:

$$H_0 : \mu_A - \mu_B = 0$$

$$H_1 : \mu_A - \mu_B \neq 0$$

The researcher chooses to test H_0 using the test-statistic

$$W = \frac{\overline{X}_A - \overline{X}_B}{\sqrt{s^2\left(\frac{1}{N_A} + \frac{1}{N_B}\right)}}$$

where $s^2 = \left(\frac{1}{N_A + N_B - 2}\right)\left[\sum_{i=1}^{N_A}\left(X_{iA} - \overline{X}_A\right)^2 + \sum_{i=1}^{N_B}\left(X_{iB} - \overline{X}_B\right)^2\right]$ where $\overline{X}_A$ and $\overline{X}_B$ are the sample means of the change in test scores in groups A and B, respectively, and X_{iA} and X_{iB} are the changes in test scores for student i in each of the groups. W is the t-statistic for the difference in means when variances are unknown but equal and has a t distribution under H_0 with $N_A + N_B - 2$ degrees of freedom. However, since the number of degrees of freedom in this example is large ($N_A + N_B - 2 = 133$), the distribution of W can be approximated by a normal.

Assuming there are 70 students in group A and 65 students in group B, and that

$\overline{X}_A = 0.1286$, $\overline{X}_B = 0.0461$, $\sum_{i=1}^{N_A}\left(X_{iA} - \overline{X}_A\right)^2 = 7.8429$, $\sum_{i=1}^{N_B}\left(X_{iB} - \overline{X}_B\right)^2 = 2.8615$, the value of the test statistic is $W = 1.6866$. In order to decide whether to reject the null hypothesis that both types of exercises have the same effect at increasing the math skills of students as measured by the improvement in their test scores, the significance level of the test must be established. If the significance level is set is set at 5% and this mass is equally distributed on both tails, the rejection rule is

$$\begin{array}{ll} \text{Reject } H_0 & \text{if } \left|W(X)\right| > 1.96 \\ \text{Do not reject } H_0 & \text{if } \left|W(X)\right| \leq 1.96 \end{array}$$

since 1.96 and -1.96 are, respectively, the 2.5% and 97.5% quantiles of the normal distribution. Given that $\left|W(X)\right| = 1.6866 < 1.96$, H_0 cannot be rejected and the researcher cannot reject the hypothesis that exercises in book A are equally effective at improving math skills than exercises in book B.

In this example, had the researcher performed a one-sided test with the alternative hypothesis that $\mu_A - \mu_B > 0$, the rejection rule would have been

$$\begin{array}{ll} \text{Reject } H_0 & \text{if } W(X) > 1.64 \\ \text{Do not reject } H_0 & \text{if } W(X) \leq 1.64 \end{array}$$

and the null hypothesis would have been rejected in favor of the alternative hypothesis that the type of exercises in book A are more effective at increasing the mathematical skills of students than the type of exercises in book B. Thus, using a two-sided test the null hypothesis cannot be rejected, even when a one-sided test (to the right) would have rejected the null hypothesis.

Rocío Titiunik

See also: Hypothesis Testing; p-value; Significance Level; Significance (Statistical Significance); Statistic; Type I Error; Type II Error; Test.

Further readings

Box, G. E. P., Hunter, J. S., & Hunter, W. G. (2005). *Statistics for experimenters. Design, innovation and discovery*. Hoboken, NJ: Wiley-Interscience.

Casella, G., & Berger, R. L. (2002). *Statistical inference*. Pacific Grove, CA: Duxbury Press.

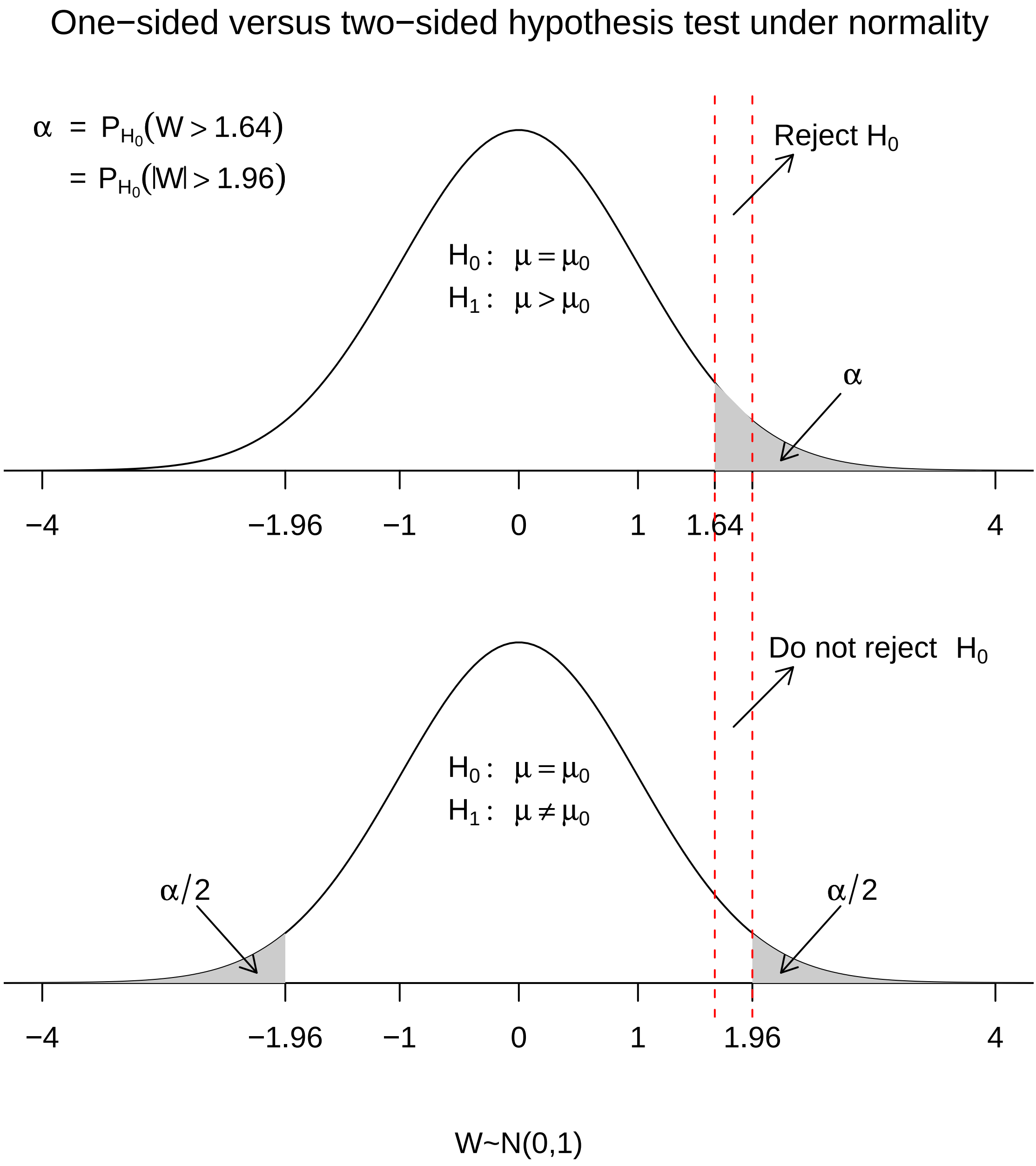
Hogg, R. V. & Craig, A. T. (1995). *Introduction to mathematical statistics*. Upper Saddle River, NJ: Prentice Hall.

Lehmann, E. L. (1986). *Testing statistical hypotheses*. New York, NY: Springer.

Lehmann, E. L. (1998). *Nonparametrics. Statistical methods based on ranks*. Upper Saddle River, NJ: Prentice Hall.

Mittelhammer, R. C. (1995). *Mathematical statistics for economics and business*. New York, NY: Springer.

Stone, C. J. (1996). *A course in probability and statistics*. Belmont, CA: Duxbury Press.


One−sided versus two−sided hypothesis test under normality
$\alpha = P_{H_0}(W > 1.64)$
$= P_{H_0}(|W| > 1.96)$
Reject H_0
$H_0: \ \mu = \mu_0$
$H_1: \ \mu > \mu_0$
α
−4
−1.96
−1
0
1
1.64
4
Do not reject H_0
$H_0: \ \mu = \mu_0$
$H_1: \ \mu \neq \mu_0$
$\alpha/2$
$\alpha/2$
−4
−1.96
−1
0
1
1.96
4
W~N(0,1)