

Significance Tests / Hypothesis Testing

Suppose someone suggests a *hypothesis* that a certain population is 0. Recalling the convoluted way in which statistics works, one way to do this would be to

- construct a confidence interval for the population mean and
- *reject the hypothesis* if the interval failed to include 0.
- We would *fail to reject the hypothesis* if the interval contained 0.

We fail to reject the hypothesis if

$$\bar{x} - 1.96 SEM \leq 0 \leq \bar{x} + 1.96 SEM$$

which can be rewritten

$$-1.96 \leq \frac{\bar{x} - \mu}{s / \sqrt{n}} \leq +1.96$$

On the other hand, we reject the hypothesis if

$$\frac{\bar{x} - \mu}{s / \sqrt{n}} \leq -1.96 \quad \text{or} \quad \frac{\bar{x} - \mu}{s / \sqrt{n}} \geq 1.96$$

The statistic $\frac{\bar{x} - \mu}{s / \sqrt{n}}$ is denoted by the symbol t . The test can be summarized as: Reject the hypothesis that the population mean is 0 if and only if the absolute value of t is greater than 1.96.

There is a 5% chance of obtaining a 95% CI that excludes 0 when it is in fact the population mean. For this reason, we say that this test has been performed at the 0.05 level of significance. Had a 99% CI been used, we would say that the test had been performed at the 0.01 level of significance, that is, the *significance level* (or simply the *level*) of the test is the probability of rejecting a hypothesis when it is true.

Statistical theory says that in many situations where a population value is estimated by drawing random samples, the sample and population values will be within two standard errors of each other 95% of the time. That is, 95% of the time,

$$-1.96 SE \leq \text{population value} - \text{sample value} \leq 1.96 SE \quad [*]$$

This is the case for means, differences between means, proportions, differences between proportions, and regression coefficients. After an appropriate transformation, this is the case for odds ratios and even correlation coefficients.

We have used this fact to construct 95% confidence intervals by restating the result as

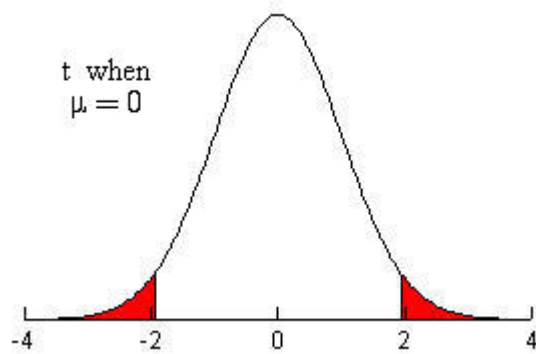
$$\text{sample value} - 1.96 SE \leq \text{population quantity} \leq \text{sample value} + 1.96 SE$$

For example, a 95% CI for the difference between two population means, $\mu_x - \mu_y$, is given by

$$(\bar{x} - \bar{y}) - 1.96 \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}} \leq \mu_x - \mu_y \leq (\bar{x} - \bar{y}) + 1.96 \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$$

When we perform significance tests, we reexpress [*] by noting that 95% of the time

$$-1.96 \leq \frac{\text{sample value} - \text{population quantity}}{SE} \leq 1.96$$



Suppose you wanted to test whether a population quantity were equal to 0. You could calculate the value of

$$t = \frac{(\text{sample value}) - 0}{SE}$$

which we get by inserting the hypothesized value of the population mean difference (0) for the population_quantity. If $t < -1.96$ or $t > 1.96$ (that is, $|t| > 1.96$), we say the data are not consistent with a

population mean difference of 0 (because t does not have the sort of value we expect to see when the population value is 0) or "we **reject the hypothesis that the population mean difference is 0**". If t were -3.7 or 2.6, we would reject the hypothesis that the population mean difference is 0 because we've observed a value of t that is unusual if the hypothesis were true.

If $-1.96 \leq t \leq 1.96$ (that is, $|t| \leq 1.96$), we say the data are consistent with a population mean difference of 0 (because t has the sort of value we expect to see when the population value is 0) or "we **fail to reject the hypothesis that the population mean difference is 0**". For example, if t were 0.76, we would fail reject the hypothesis that the population mean difference is 0 because we've observed a value of t that is unremarkable if the hypothesis were true.

This is called "fixed level testing" (at the 0.05 level).

- **First:** We state the hypothesis to be tested. The hypothesis being tested is called the *null hypothesis* and is denoted H_0 . Often, the null hypothesis states a specific value for a population parameter.
- **Second:** We choose the level of significance at which the test will be performed. This is called the *size* or *level* of the test. It is the probability of rejecting the null hypothesis when it is true. The level of the test determines the values of the test statistic (such as t) that would cause us to reject the hypothesis.
- **Third:** We then, **and only then**, collect the data and reject the hypothesis or not depending on the observed value of the test statistic.

For example, if $H_0: \mu_x = \mu_y$ (which can be rewritten $H_0: \mu_x - \mu_y = 0$), the test statistic is

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}}$$

If $|t| > 1.96$, reject $H_0: \mu_x = \mu_y$ at the 0.05 level of significance.

When we were constructing confidence intervals, it mattered whether the data were drawn from normally distributed populations, whether the population standard deviations were equal, and whether the sample sizes were large or small. The answers to these questions helped us determine the proper multiplier for the standard error. The same considerations apply to significance tests. The answers determine the critical value of t for a result to be declared statistically significant.

When populations are normally distributed with unequal standard deviations and the sample size is small, the multiplier used to construct CIs is based on the t distribution with noninteger degrees of freedom. The same noninteger degrees of freedom appear when performing significance tests. Many ways to calculate the degrees of freedom have been proposed. The statistical program package SPSS, for example, uses the Satterthwaite formula

$$\frac{(d_x + d_y)^2}{\frac{d_x^2}{n_x - 1} + \frac{d_y^2}{n_y - 1}}, \text{ where } d_i = \frac{s_i^2}{n_i} .$$

Terminology

- **Null hypothesis**--the hypothesis under test, denoted H_0 . The null hypothesis is usually stated as the absence of a difference or an effect. It functions as what debaters call a "straw man", something set up solely to be knocked down. It is usually the investigator's intention to demonstrate an effect is present. The null hypothesis says there is no effect. The null hypothesis is rejected if the significance test shows the data are inconsistent with the null hypothesis. Null hypothesis are *never* accepted. We either reject them or fail to reject them. The distinction between "acceptance" and "failure to reject" is best understood in terms of confidence intervals. Failing to reject a hypothesis means a confidence interval contains a value of "no difference". However, the data may also be consistent with differences of practical importance. Hence, failing to reject H_0 does not mean that we have shown that there is no difference (accept H_0).
- **Alternative Hypothesis**--the alternative to the null hypothesis. It is denoted H' , H_1 , or H_A . It is usually the complement of the null hypothesis. Many authors talk about rejecting the null hypothesis in favor of the alternative. If, for example, the null hypothesis says two

population means are equal, the alternative says the means are unequal. The amount of emphasis on the alternative hypothesis will depend on whom you read. It is central to the Neyman-Pearson school of frequentist statistics. Yet, R.A. Fisher, who first introduced the notion of significance tests in a formal systematic way, never considered alternative hypotheses. He focused entirely on the null.

- **Critical Region (Rejection Region)**--the set of values of the test statistic that cause the null hypothesis to be reject. If the test statistic falls in the rejection region--that is, if the statistic is a value that is in the rejection region--the null hypothesis is rejected. In the picture above, the critical region is the area filled in with red.
- **Critical Values**--the values that mark the boundaries of the critical region. For example, if a critical region is $\{t \leq -1.96, t \geq 1.96\}$, the critical values are ± 1.96 as in the picture above.
- **Power** is the probability of rejecting the null hypothesis. It is not a single value. It varies according to the underlying truth. For example, the probability of rejecting the hypothesis of equal population means depends on the actual difference in population means. The probability of the rejecting the null hypothesis increases with the difference between population means.
- The **level** (or **size**) of a test is the probability of rejecting the null hypothesis when it is true. It is denoted by the Greek letter α (*alpha*). Rejecting the null hypothesis, H_0 , when it is true is called a **Type I Error**. Therefore, if the null hypothesis is true α , the level of the test, is the probability of a type I error. α is also the power of the test when the null hypothesis, H_0 , is true. In the picture above, α is the proportion of the distribution colored in red. The choice of α determines the critical values. The tails of the distribution of t are colored in until the proportion filled in is α , which determines the critical values.
- A **Type II Error** occurs when we fail to reject the null hypothesis when it is false. The probability of a type II error depends on the way the null hypothesis is false. For example, for a fixed sample size, the probability of failing to reject a null hypothesis of equal population means decreases as the difference between population means increases. The probability of a type II error is denoted by the Greek letter β (*beta*). By definition, power = $1 - \beta$ when the null hypothesis is false.

The difference between type I & type II errors is illustrated by the following legal analogy. Under United States law, defendants are presumed innocent until proven guilty. The purpose of a trial is to see whether a null hypothesis of innocence is rejected by the weight of the data (evidence). A type I error (rejecting the null hypothesis when it is true) is "convicting the innocent." A type II error (failing to reject the null hypothesis when it is false) is "letting the guilty go free."

A common mistake is to confuse a type I or II error with its probability. α is not a type I error. It is the *probability* of a type I error. Similarly, β is not a type II error. It is the *probability* of a type II error.

There's a trade-off between α and β . Both are probabilities of making an error. With a fixed sample size, the only way to reduce the probability of making one type of error is to increase

the other. For the problem of comparing population means, consider the rejection region whose critical values are $\pm \infty$. This excludes every possible difference in sample means. H_0 will never be rejected. Since the null hypothesis will never be rejected, the probability of rejecting the null hypothesis when it is true is 0. So, $\alpha=0$. However, since the null hypothesis will never be rejected, the probability of failing to reject the null hypothesis when it is false is 1, that is, $\beta=1$.

Now consider the opposite extreme--a rejection region whose critical values are 0,0. The rejection region includes every possible difference in sample means. This test always rejects H_0 . Since the null hypothesis is always rejected, the probability of rejecting H_0 when it is true is 1, that is, $\alpha=1$. On the other hand, since the null hypothesis is always rejected, the probability of failing to reject it when it is false is 0, that is, $\beta=0$.

To recap, the test with a critical region bounded by $\pm \infty$ has $\alpha=0$ and $\beta=1$, while the test with a critical region bounded by 0,0 has $\alpha=1$ and $\beta=0$. Now consider tests with intermediate critical regions bounded by $\pm k$. As k increases from 0 to ∞ , α decreases from 1 to 0 while β increases from 0 to 1.

Every statistics textbook contains discussions of α , β , type I error, type II error, and power. Analysts should be familiar with all of them. However, α is the only one that is encountered regularly in reports and published papers. That's because standard statistical practice is to carry out significance tests at the 0.05 level. As we've just seen, choosing a particular value for α determines the value of β .

The one place where β figures prominently in statistical practice is in determining sample size. When a study is being planned, it is possible to choose the sample size to set the power to any desired value for some particular alternative to the null hypothesis. To illustrate this, suppose we are testing the hypothesis that two population means are equal at the 0.05 level of significance by selecting equal sample sizes from the two populations. Suppose the common population standard deviation is 12. Then, if the population mean difference is 10, a sample of 24 subjects per group gives an 81% chance of rejecting the null hypothesis of no difference (power=0.81, $\beta=0.19$). A sample of 32 subjects per group gives an 91% chance of rejecting the null hypothesis of no difference (power=0.91, $\beta=0.09$). This is discussed in detail in the section on [sample size determination](#).

[back to [LHSP](#)]

Copyright © 2000 Gerard E. Dallal
Last modified: 03/19/2007 17:55:55.