GLOBAL EDITION

Statistical Methods
for the Social Sciences

FIFTH EDITION

ALAN AGRESTI

Pearson

Statistical Methods for the Social Sciences

Fifth Edition

Global Edition

Alan Agresti

University of Florida

Pearson

Harlow, England • London • New York • Boston • San Francisco • Toronto • Sydney • Dubai • Singapore • Hong Kong
Tokyo • Seoul • Taipei • New Delhi • Cape Town • São Paulo • Mexico City • Madrid • Amsterdam • Munich • Paris • Milan

It is better to report the P-value than to indicate merely whether the result is "statistically significant." Reporting the P-value has the advantage that the reader can tell whether the result is significant at any level. The P-values of 0.049 and 0.001 are both "significant at the 0.05 level," but the second case provides much stronger evidence than the first case. Likewise, P-values of 0.049 and 0.051 provide, in practical terms, the same amount of evidence about H_0. It is a bit artificial to call one result "significant" and the other "nonsignificant." Some software places the symbol * next to a test statistic that is significant at the 0.05 level, ** next to a test statistic that is significant at the 0.01 level, and *** next to a test statistic that is significant at the 0.001 level.

TYPE I AND TYPE II ERRORS FOR DECISIONS

Because of sampling variability, decisions in tests always have some uncertainty. The decision could be erroneous. The two types of potential errors are conventionally called *Type I* and *Type II* errors.

Type I and Type II Errors

When H_0 is true, a ***Type I error*** occurs if H_0 is rejected.
When H_0 is false, a ***Type II error*** occurs if H_0 is not rejected.

The two possible decisions cross-classified with the two possibilities for whether H_0 is true generate four possible results. See Table 6.8.

TABLE 6.8: The Four Possible Results of Making a Decision in a Significance Test. Type I and Type II errors are the incorrect decisions.

		Decision	
		Reject H_0	Do Not Reject H_0
Condition of H_0	H_0 true	Type I error	Correct decision
	H_0 false	Correct decision	Type II error

REJECTION REGIONS: STATISTICALLY SIGNIFICANT TEST STATISTIC VALUES

The collection of test statistic values for which the test rejects H_0 is called the ***rejection region***. For example, the rejection region for a test of level $\alpha = 0.05$ is the set of test statistic values for which $P \leq 0.05$.

For two-sided tests about a proportion, the two-tail P-value is ≤ 0.05 whenever the test statistic $|z| \geq 1.96$. In other words, the rejection region consists of values of z resulting from the estimate falling at least 1.96 standard errors from the H_0 value.

THE α-LEVEL IS THE PROBABILITY OF TYPE I ERROR

When H_0 is true, let's find the probability of Type I error. Suppose $\alpha = 0.05$. We've just seen that for the two-sided test about a proportion, the rejection region is $|z| \geq 1.96$. So, the probability of rejecting H_0 is exactly 0.05, because the probability of the values in this rejection region under the standard normal curve is 0.05. But this is precisely the α-level.

The probability of a Type I error is the α-level for the test.

With $\alpha = 0.05$, if H_0 is true, the probability equals 0.05 of making a Type I error and rejecting H_0. We control P(Type I error) by the choice of α. The more serious the consequences of a Type I error, the smaller α should be. In practice, $\alpha = 0.05$ is most common, just as an error probability of 0.05 is most common with confidence intervals (i.e, 95% confidence). However, this may be too high when a decision has serious implications.

For example, consider a criminal legal trial of a defendant. Let H_0 represent innocence and H_a represent guilt. The jury rejects H_0 and judges the defendant to be guilty if it decides the evidence is sufficient to convict. A Type I error, rejecting a true H_0, occurs in convicting a defendant who is actually innocent. In a murder trial, suppose a convicted defendant may receive the death penalty. Then, if a defendant is actually innocent, we would hope that the probability of conviction is much smaller than 0.05.

When we make a decision, we do not know whether we have made a Type I or Type II error, just as we do not know whether a particular confidence interval truly contains the parameter value. However, we can control the probability of an incorrect decision for either type of inference.

AS *P*(TYPE I ERROR) GOES DOWN, *P*(TYPE II ERROR) GOES UP

In an ideal world, Type I or Type II errors would not occur. However, errors do happen. We've all read about defendants who were convicted but later determined to be innocent. When we make a decision, why don't we use an extremely small P(Type I error), such as $\alpha = 0.000001$? For instance, why don't we make it almost impossible to convict someone who is really innocent?

When we make α smaller in a significance test, we need a smaller P-value to reject H_0. It then becomes harder to reject H_0. But this means that it will also be harder even if H_0 is false. The stronger the evidence required to convict someone, the more likely we will fail to convict defendants who are actually guilty. In other words, the smaller we make P(Type I error), the larger P(Type II error) becomes, that is, failing to reject H_0 even though it is false.

If we tolerate only an extremely small P(Type I error), such as $\alpha = 0.000001$, the test may be unlikely to reject H_0 even if it is false—for instance, unlikely to convict someone even if they are guilty. This reasoning reflects the fundamental relation:

- The smaller P(Type I error) is, the larger P(Type II error) is.

For instance, in an example in Section 6.6, when P(Type I error) = 0.05 we'll find that P(Type II error) = 0.02, but when P(Type I error) decreases to 0.01, P(Type II error) increases to 0.08. Except in Section 6.6, we shall not find P(Type II error), as it is beyond our scope. In practice, making a decision requires setting only α, the P(Type I error).

Section 6.6 shows that P(Type II error) depends on just how far the true parameter value falls from H_0. If the parameter is nearly equal to the value in H_0, P(Type II error) is relatively high. If it falls far from H_0, P(Type II error) is relatively low. The farther the parameter falls from the H_0 value, the less likely the sample is to result in a Type II error.

For a fixed P(Type I error), P(Type II error) depends also on the sample size n. The larger the sample size, the more likely we are to reject a false H_0. To keep both P(Type I error) and P(Type II error) at low levels, it may be necessary to use a very

large sample size. The P(Type II error) may be quite large when the sample size is small, unless the parameter falls quite far from the H_0 value.

EQUIVALENCE BETWEEN CONFIDENCE INTERVALS AND TEST DECISIONS

We now elaborate on the equivalence for means[5] between decisions from two-sided tests and conclusions from confidence intervals, first alluded to in Example 6.3 (page 159). Consider the significance test of

$$H_0: \mu = \mu_0 \qquad \text{versus} \qquad H_a: \mu \neq \mu_0.$$

When $P < 0.05$, H_0 is rejected at the $\alpha = 0.05$ level. When n is large (so the t distribution is essentially the same as the standard normal), this happens when the test statistic $t = (\bar{y} - \mu_0)/se$ is greater in absolute value than 1.96, that is, when $\bar{y}$ falls more than $1.96(se)$ from μ_0. But if this happens, then the 95% confidence interval for μ, namely, $\bar{y} \pm 1.96(se)$, does not contain the null hypothesis value μ_0. See Figure 6.7. These two inference procedures are consistent.

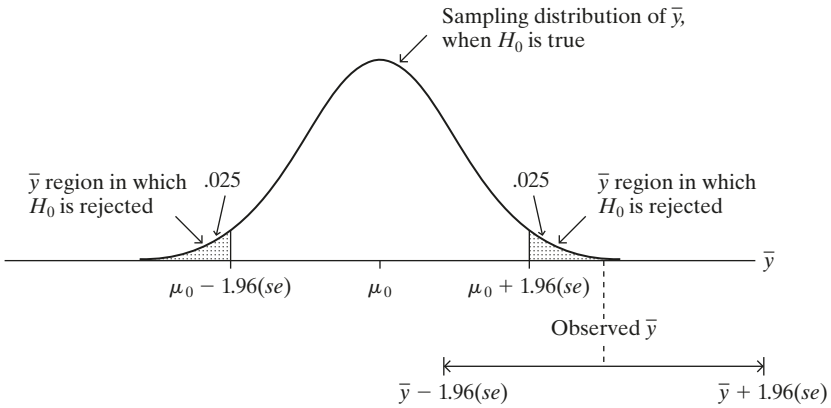

FIGURE 6.7: Relationship between Confidence Interval and Significance Test. For large n, the 95% confidence interval does not contain the H_0 value μ_0 when the sample mean falls more than 1.96 standard errors from μ_0, in which case the test statistic $|t| > 1.96$ and the P-value < 0.05.

Significance Test Decisions and Confidence Intervals

> In testing $H_0: \mu = \mu_0$ against $H_a: \mu \neq \mu_0$, when we reject H_0 at the 0.05 α-level, the 95% confidence interval for μ does not contain μ_0. The 95% confidence interval consists of those μ_0 values for which we do not reject H_0 at the 0.05 α-level.

In Example 6.2 about mean political ideology (page 157), the P-value for testing $H_0: \mu = 4.0$ against $H_a: \mu \neq 4.0$ was $P = 0.20$. At the $\alpha = 0.05$ level, we do not reject $H_0: \mu = 4.0$. It is believable that $\mu = 4.0$. Example 6.3 (page 159) showed that a 95% confidence interval for μ is (3.95, 4.23), which contains $\mu_0 = 4.0$.

Rejecting H_0 at a particular α-level is equivalent to the confidence interval for μ with the same error probability not containing μ_0. For example, if a 99% confidence interval does not contain 0, then we would reject $H_0: \mu = 0$ in favor of $H_a: \mu \neq 0$ at the $\alpha = 0.01$ level with the test. The α-level is P(Type I error) for the test and the probability that the confidence interval method does not contain the parameter.

[5] This equivalence also holds for proportions when we use the two-sided test of Section 6.3 and the confidence interval method presented in Exercise 5.77.

MAKING DECISIONS VERSUS REPORTING THE *P*-VALUE

The approach to hypothesis testing that incorporates a formal decision with a fixed P(Type I error) was developed by the statisticians Jerzy Neyman and Egon Pearson in the late 1920s and early 1930s. In summary, this approach formulates null and alternative hypotheses, selects an α-level for the P(Type I error), determines the rejection region of test statistic values that provide enough evidence to reject H_0, and then makes a decision about whether to reject H_0 according to what is actually observed for the test statistic value. With this approach, it's not even necessary to find a P-value. The choice of α-level determines the rejection region, which together with the test statistic determines the decision.

The alternative approach of finding a P-value and using it to summarize evidence against a hypothesis is due to the great British statistician R. A. Fisher. He advocated merely reporting the P-value rather than using it to make a formal decision about H_0. Over time, this approach has gained favor, especially since software can now report precise P-values for a wide variety of significance tests.

This chapter has presented an amalgamation of the two approaches (the decision-based approach using an α-level and the P-value approach), so you can interpret a P-value yet also know how to use it to make a decision when that is needed. These days, most research articles merely report the P-value rather than a decision about whether to reject H_0. From the P-value, readers can view the strength of evidence against H_0 and make their own decision, if they want to.

6.5 Limitations of Significance Tests

A significance test makes an inference about whether a parameter differs from the H_0 value and about its direction from that value. In practice, we also want to know whether the parameter is sufficiently different from the H_0 value to be practically important. In this section, we'll learn that a test does not tell us as much as a confidence interval about practical importance.

STATISTICAL SIGNIFICANCE VERSUS PRACTICAL SIGNIFICANCE

It is important to distinguish between *statistical significance* and *practical significance*. A small P-value, such as $P = 0.001$, is highly statistically significant. It provides strong evidence against H_0. It does not, however, imply an *important* finding in any practical sense. The small P-value merely means that if H_0 were true, the observed data would be very unusual. It does not mean that the true parameter value is far from H_0 in practical terms.

Example 6.7

Mean Political Ideology for All Americans The political ideology $\bar{y} = 4.089$ reported in Example 6.2 (page 157) refers to a sample of Hispanic Americans. We now consider the entire 2014 GSS sample who responded to the political ideology question. For a scoring of 1.0 through 7.0 for the ideology categories with 4.0 = moderate, the $n = 2575$ observations have $\bar{y} = 4.108$ and standard deviation $s = 4.125$. On the average, political ideology was the same for the entire sample as it was for Hispanics alone.[6]

As in Example 6.2, we test H_0: $\mu = 4.0$ against H_a: $\mu \neq 4.0$ to analyze whether the population mean differs from the moderate ideology score of 4.0. Now,

[6] And it seems stable over time, equaling 4.13 in 1980, 4.16 in 1990, and 4.10 in 2000.

$se = s/\sqrt{n} = 1.425/\sqrt{2575} = 0.028$, and

$$t = \frac{\bar{y} - \mu_0}{se} = \frac{4.108 - 4.0}{0.028} = 3.85.$$

The two-sided P-value is $P = 0.0001$. There is *very* strong evidence that the true mean exceeds 4.0, that is, that the true mean falls on the conservative side of moderate. But, on a scale of 1.0 to 7.0, 4.108 is close to the moderate score of 4.0. Although the difference of 0.108 between the sample mean of 4.108 and the H_0 mean of 4.0 is highly significant statistically, the magnitude of this difference is very small in practical terms. The mean response on political ideology for all Americans is essentially a moderate one. ■

In Example 6.2, the sample mean of $\bar{y} = 4.1$ for $n = 369$ Hispanic Americans had a P-value of $P = 0.20$, not much evidence against H_0. But now with $\bar{y} = 4.1$ based on $n = 2575$, we have instead found $P = 0.0001$. This is highly *statistically significant*, but not *practically significant*. For practical purposes, a mean of 4.1 on a scale of 1.0 to 7.0 for political ideology does not differ from 4.00.

A way of summarizing practical significance is to measure the ***effect size*** by the number of standard deviations (*not* standard errors) that $\bar{y}$ falls from μ_0. In this example, the estimated effect size is $(4.108 - 4.0)/1.425 = 0.08$. This is a tiny effect. Whether a particular effect size is small, medium, or large depends on the substantive context, but an effect size of about 0.2 or less in absolute value is usually not practically important.

SIGNIFICANCE TESTS ARE LESS USEFUL THAN CONFIDENCE INTERVALS

We've seen that, with large sample sizes, P-values can be small even when the point estimate falls near the H_0 value. The size of P merely summarizes the extent of evidence about H_0, not how far the parameter falls from H_0. Always inspect the difference between the estimate and the H_0 value to gauge the practical implications of a test result.

Null hypotheses containing single values are rarely true. That is, rarely is the parameter *exactly* equal to the value listed in H_0. With sufficiently large samples, so that a Type II error is unlikely, these hypotheses will normally be rejected. What is more relevant is whether the parameter is sufficiently different from the H_0 value to be of practical importance.

Although significance tests can be useful, most statisticians believe they are overemphasized in social science research. It is preferable to construct confidence intervals for parameters instead of performing only significance tests. A test merely indicates whether the particular value in H_0 is plausible. It does not tell us which other potential values are plausible. The confidence interval, by contrast, displays the entire set of believable values. It shows the extent to which reality may differ from the parameter value in H_0 by showing whether the values in the interval are far from the H_0 value. Thus, it helps us to determine whether rejection of H_0 has practical importance.

To illustrate, for the complete political ideology data in Example 6.7, a 95% confidence interval for μ is

$$\bar{y} \pm 1.96(se) = 4.108 \pm 1.96(0.028), \text{ or } (4.05, 4.16).$$

This indicates that the difference between the population mean and the moderate score of 4.0 is very small. Although the P-value of $P = 0.0001$ provides very strong evidence against H_0: $\mu = 4.0$, in practical terms the confidence interval shows that

H_0 is not wrong by much. By contrast, if $\bar{y}$ had been 6.108 (instead of 4.108), the 95% confidence interval would equal (6.05, 6.16). This indicates a substantial practical difference from 4.0, the mean response being near the conservative score rather than the moderate score.

When a P-value is not small but the confidence interval is quite wide, this forces us to realize that the parameter might well fall far from H_0 even though we cannot reject it. This also supports why it does not make sense to "accept H_0," as we discussed on page 167.

The remainder of the text presents significance tests for a variety of situations. It is important to become familiar with these tests, if for no other reason than their frequent use in social science research. However, we'll also introduce confidence intervals that describe how far reality is from the H_0 value.

SIGNIFICANCE TESTS AND P-VALUES CAN BE MISLEADING

We've seen it is improper to "accept H_0." We've also seen that statistical significance does not imply practical significance. Here are other ways that results of significance tests can be misleading:

- **It is misleading to report results only if they are statistically significant.** Some research journals have the policy of publishing results of a study only if the P-value ≤ 0.05. Here's a danger of this policy: Suppose there truly is no effect, but 20 researchers independently conduct studies. We would expect about $20(0.05) = 1$ of them to obtain significance at the 0.05 level merely by chance. (When H_0 is true, about 5% of the time we get a P-value below 0.05 anyway.) If that researcher then submits results to a journal but the other 19 researchers do not, the article published will be a Type I error. It will report an effect when there really is not one.
- **Some tests may be statistically significant just by chance.** You should never scan software output for results that are statistically significant and report only those. If you run 100 tests, even if all the null hypotheses are correct, you would expect to get P-values ≤ 0.05 about $100(0.05) = 5$ times. Be skeptical of reports of significance that might merely reflect ordinary random variability.
- **It is incorrect to interpret the P-value as the probability that H_0 is true.** The P-value is P(test statistic takes value like observed or even more extreme), presuming that H_0 is true. It is not $P(H_0$ true). Classical statistical methods calculate probabilities about variables and statistics (such as test statistics) that vary randomly from sample to sample, not about parameters. Statistics have sampling distributions, parameters do not. In reality, H_0 is not a matter of probability. It is either true or not true. We just don't know which is the case.
- **True effects are often smaller than reported estimates.** Even if a statistically significant result is a real effect, the true effect may be smaller than reported. For example, often several researchers perform similar studies, but the results that receive attention are the most extreme ones. The researcher who decides to publicize the result may be the one who got the most impressive sample result, perhaps way out in the tail of the sampling distribution of all the possible results. See Figure 6.8.

Example 6.8

Are Many Medical "Discoveries" Actually Type I Errors? In medical research studies, suppose that an actual population effect exists only 10% of the time. Suppose also that when an effect truly exists, there is a 50% chance of making a Type II error