Reporting for Hypothesis Tests

By Peter Fortini

Q What is a *p*-value the probability of?

A The *p*-value is a commonly used brief method to report the results of a statistical hypothesis test, as in "The effect is significant (p = 0.013)." Quoting the *p*-value provides more information than the simple statement that the null hypothesis of the test is rejected or not rejected.

The standard practice for calculating and using basic statistics (E2586) added a section briefly describing the concepts of hypothesis testing and the terms associated with it, including the *p*-value, in 2018.[1] The definition given for *p*-value as a term is: The probability of observing a test statistic at least as extreme as what was actually obtained, under the assumption of the null hypothesis.

In 2016, a statement on *p*-values was generated by a special committee of the American Statistical Association. The occasion for the statement and some of its content will be discussed below, but the following answer to our question is in the introduction to the statement:[2] "Informally, a *p*-value is the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or greater than its observed value."

Hypothesis testing is an important part of statistical analysis. It is used to claim that one variable, *X*, has an effect on another, *Y*. This is accomplished by setting up as the null hypothesis that *X* has no effect on *Y*. Then a test statistic is designed to be sensitive to the effect. A level of significance is chosen. If the test statistic is in a critical region consisting of values more likely in case *X* to have an effect on *Y*, then the null hypothesis is said to be "rejected." Other uses of hypothesis testing include tests for outliers and tests that assumptions are met, such as that a sample comes from a normal distribution, carried out as part of the analysis of data prior to testing or estimating variable effects.

The *p*-value is related to the significance level of a test. To evaluate a hypothesis test, you first choose a test statistic with a known distribution under the null hypothesis. Test statistics are designed to measure the departure of data from the null hypothesis in the direction we are interested in. Test statistics are then adjusted so that their distributions will be known and thus tabulated. Examples are the *z* statistic:

$$z = (\bar{x} - \mu_0)/(\frac{\sigma}{\sqrt{n}})$$

and the Student's *t* statistic:

$$t = (\bar{x} - \mu_0)/(\frac{s}{\sqrt{n}})$$

for testing a hypothesis about a mean, H_0: $\mu = \mu_0$. The normalization by σ/√n or by s/√n enables us to compare the statistic using a single table, regardless of the standard deviation (σ or s) and the number of observations.

ORIGIN OF THE P-VALUE AND STANDARD SIGNIFICANCE LEVELS FOR HYPOTHESIS TESTING

Some historical background is in order. The p-value predates the concept of significance testing or hypothesis testing. History also shows the origin of the levels of significance that are most often used in hypothesis testing.

The application that introduced the *p*-value was that of evaluating the quality of a graduation curve (Pearson curve) fit to a set of data.[3] The criterion was called "goodness of fit." In fitting a curve to a data distribution, a distribution equation form was selected. Parameters of the curve were estimated. The data were grouped, and the number of observations (O_i) in each group compared to the expected number (E_i).

Continued on page 46

Continued from page 44

The chi-square statistic:

$$x^2=\sum\frac{(O_i-E_i)^2}{E_i}$$

assessed goodness of fit of the data to the graduated curve. For large numbers of observations, the statistic has a chi-squared distribution. A table of the chi-squared distribution gives the probability the calculated value would be equaled or exceeded by chance as a function of the number of groups. (This was before the reduction in degrees of freedom due to fitted parameters was understood.) This was the *p*-value. A low value would indicate a poor fit of the distribution to the data. A high value, near 1, was equally suspect as it would indicate overfitting to the sample data.

For a single distribution, such as the standard normal for the *z*-test, the complete distribution function, probability as a function of the value of the variable, is still provided in textbooks and books of mathematical and statistical tables. The chi-squared distribution, and later the Student's *t* distribution for testing means, differences, and regression coefficients, all require tabulation for a series of degrees of freedom. This makes for a lengthy table, cumbersome to use. In addition, when Fisher wrote his classic text[4], he was prevented by copyright from reproducing the fairly compact table of the chi-squared distribution given by Elderton. Therefore, a new format for the table was provided. Instead of giving cumulative probability for a range of values of the statistic at each degree of freedom, it gives selected percentiles of the distribution at each degree of freedom. Percentiles 0.05, 0.01, and later, 0.001, were the ones featured in the tabulation. These thereby became the standards to declare test results significant, highly significant, and extremely significant.

The process used when performing tests of hypotheses, into the 1970s, was to calculate the test statistic for the data at hand, then compare the calculated value to the tabulated percentile of the distribution in a book of tables to determine statistical significance.

Enter the digital computer. With statistical analyses being performed with computers and software, it is now easier to write a subroutine to compute the cumulative distribution than it is to carry a table of percentiles in memory, and the length of the table is no longer an issue. Thus, the *p*-value came again to be a preferred form of reporting the results of the test.

p-VALUES BECOME CONTROVERSIAL

In recent years, the *p*-value has become controversial in science. In medical research and in other areas, it is found that attempts to replicate published studies frequently fail to find the same effect. Ionannidis pointed out in an influential, provocatively titled article how many medical research findings relying on statistical significance tests fail to be confirmed in follow-up studies if any were published.[5] The American Statistical Association responded to the concerns by developing their statement. A highly readable summary of the issues is in a report by the National Association of Scholars.[6]

It is not so much the *p*-value itself at fault here, but the manner in which hypothesis testing is misapplied in scientific literature. The need to understand the limitations of techniques, things that you must bear in mind when carrying out statistical analysis, has always been known. It is the ease of carrying out statistical analyses without understanding these limitations that causes trouble. During drafting and approval of the section on hypothesis testing of E2586, the writers were well aware of the concerns about misuse and misinterpretation and added notes of clarification and cautions to that standard.

Some of these concerns are as follows:

— The *p*-value cannot be interpreted in any sense as a probability that the null hypothesis is true. While it is a valid measure to indicate how incompatible a data set are with the hypothesis, that measure is not interpretable as a probability. The probability of the null hypothesis, given data, can be evaluated in a Bayesian framework given prior probabilities of the null and alternative hypotheses. The probability of a null hypothesis that is rejected by a statistical test will be lower than if not rejected. However, the probability of the null hypothesis is not given by the *p*-value.

— Statistical significance as indicated by the *p*-value does not measure the practical importance of an effect. Also, a null hypothesis test rejection is not definitive evidence for any particular alternative.

— When many statistical comparisons are carried out on a data set, and primarily those that are statistically significant are reported and used to base conclusions, the nominal significance level for testing no longer applies and conclusions must be considered speculative.

— Selective reporting, with data being reported only when the statistical test is significant, is a serious issue that can distort the scientific process and compromise the validity of conclusions. The solution to this issue is discipline in formulating hypotheses to test and expectations of results to be obtained before data are collected, and full reporting regardless of the results of the test. ■

REFERENCES

1. The Standard Practice for Calculating and Using Basic Statistics (E2586).
2. R. L. Wasserstein and N. A. Lazar, "The ASA's Statement on p-values: Context, Process, and Purpose," *The American Statistician,* Vol. 70, No. 2, May 2016, pp. 129-133.
3. W. P. Elderton, *Frequency Curves and Correlation,* 4th ed., Cambridge University Press, 1953 (1st ed. 1906).
4. R. A. Fisher, *Statistical Methods for Research Workers,* 14th ed., Hafner, 1970 (1st ed. 1925).
5. J. P. A. Ioannidis, "Why Most Published Research Findings Are False," *PLoS Medicine,* Vol. 2, No. 8, August 2005, E124.
6. D. Randall and C. Welser, *The Irreproducibility Crisis of Modern Science,* National Association of Scholars, April 2018.