BY JULIA E. SEAMAN
AND I. ELAINE ALLEN

Not Significant, But Important?

Know the pitfalls of p-values and formal hypothesis tests

IN MARCH 2011, the Supreme Court ruled that even if a result from a controlled clinical trial was not statistically significant, it still might be important.

In the case, brought by investors in the biopharmaceutical firm Matrixx Initiatives, judges ruled that the company failed to disclose reports that its over-the-counter medicine, Zicam, sometimes caused a loss of the sense of smell.

Matrixx Initiatives argued it had no reason to disclose these adverse events because the results did not reach statistical significance. The court rejected that argument and let the case proceed in a lower court. An article in the *Wall Street Journal* trumpeted this result and quoted statisticians supporting the Supreme Court.[1]

What does this say about hypothesis testing and statistical evidence? Doesn't this type of ruling fly in the face of the scientific method?

Since R.A. Fisher's investigation of agricultural plantings in the 1920s, statisticians have used hypothesis testing and critical significance levels to make conclusions about their data. Surprisingly, when Fisher proposed his null and alternative hypotheses, they were not welcomed by all statisticians.

Jerzy Neyman and Egon Pearson attacked this notion of statistical significance and suggested it made more sense to test competing hypotheses against one another. Rather than "rejecting" or "not rejecting" a null hypothesis, these competing hypotheses-statistical tests gauge the likelihood of a "false positive." To Fisher, it's necessary to note the exact p-value was important and not a critical cut-off value, such as 5% probability for overall decision making.[2]

Important vs. significant difference example / TABLE 1

	Massachusetts start-up		Global start-up	
	Male	Female	Male	Female
2009	8.8%	5.0%	8.4%	5.0%
p = Not significant p < 0.001				

P-values

What is a p-value? What exactly are you testing? What does probability mean in real life?

The definition of a p-value is correctly explained by Tom Siegfried: "Experimental data yielding a p-value of 0.05 means that there is only a 5% chance of obtaining an observed (or more extreme) result if no real effect exists (that is, if the no-difference hypothesis is correct)."[3]

The definition shows the p-value gives information about the probability of obtaining evidence. It doesn't quantify the strength of the evidence. So, if there is a significant difference and the null hypothesis is rejected, how do you know if the result is important? The short answer is you don't.

A nonsignificant difference may or may not be an important difference, as shown in Table 1, which contains some results from a study by the Global Entrepreneurship Monitor (GEM). The study was written after the research consortium interviewed more than 200,000 individuals globally in more than 60 countries. Each country contributed a sample of 2,000 or more.[4]

For example, some results from the study compare male and female entrepreneurship rates globally and within Massachusetts. The rates are comparable, but the statistical results are conflicting. Are these results important? Why are the tests giving different results?

The Massachusetts sample is 1,000, of which only about 100 are entrepreneurs. The global sample is close to 200,000, of which about 20,000 are entrepreneurs. In both tests, the percentage of males starting a business is almost 50% larger than females starting a business.

Given differing statistical conclusions for the same evidence, it is important not to come to different overall conclusions. One way to ensure statistical significance is to increase your sample size. Alternatively, if your sample size is not large enough, nothing will be significant because there is not enough power to discriminate between the two groups.

A difference may be significant in multiple tests but remain unimportant to the overall evaluation or project. With advances in manufacturing using computers to input the specifications, machines may have small but highly significant differences that are meaningless to the overall process because the tolerances are so small.

In Table 2, for example, statistical significance is driven by the small (or lack of) variability within a machine, unlike the earlier GEM example, which was driven by sample size. A p-value indicates the likelihood there is a relationship that is not due to chance. A p-value does not indicate the strength of the relationship or whether the differences it is examining are relevant.[5]

Table 2 shows how two cutters created the same widget part of length 5 in a factory. Three samples from each cutter were

Non-important significance example / TABLE 2

	Machine 1	Machine 2
Test lengths	5.00123	4.99821
	5.00099	4.99799
	5.00111	4.99803
Average	5.00111	4.99808
Standard deviation	0.00012	0.00012
$p < 0.001$		

accurately measured and showed significant differences. Is this result important?

The machines consistently cut widget parts that were different lengths from one another. But both cutters created parts that were much less than 0.1% different from the true length, a difference that was likely to have little overall effect. In this case, the machines were so accurate that a significant difference was achieved, but it was a trivial significance.

Conditional probability

It is also critical to understand that the statistical hypothesis test is always a conditional test—conditional on the null hypothesis (usually of "no difference") being true. It is not, as usually stated, that there is a real statistically significant difference between two sets of data, but rather the conditional probability of observing data is as extreme (or more extreme) than what was seen in the sample.

In other words, the significance test calculates the resulting p-value, assuming the null hypothesis is true. The p-value result describes the sample gathered for the test and tells the investigator how unusual the sample is.

A statistically significant test result is meaningless without the proper design and interpretation. Before any analysis, the investigator must be careful the hypothesis in question is actually being tested.

Additionally, some thought should be given ahead of time to possible confounders, quality control and statistical corrections. If a hypothesis is not rejected, appropriate questions for the investigator are, "Was the sample size large enough? Was the outcome measured with sufficient precision as to detect a difference (or discriminate) between the null and alternative hypotheses if a true difference existed?"

When a hypothesis is rejected, the appropriate question is, "Was my sample size so large—or my measurement error so small—that I would have rejected any null hypothesis?" Table 3 converts these questions into a short list of how to report hypothesis-based test results.

Frequentist vs. Bayesian

Even following the rubric in Table 3, many statisticians still believe classic statistical tests are inferior because they rely on strict comparison to the null hypothesis. Is there a better way? What might be in the future for null hypotheses and statistical tests?

To avoid the pitfalls of p-values and formal hypothesis tests, many researchers are pointing to Bayesian methods instead of the classical frequentist methods examined here and based on Fisher's work. Bayesian techniques rely on the data for pointing in the direction of any conclusions based on the data.

Stated another way, the Bayesian approach calculates the probability of the hypothesis given the data, while the frequentist approach computes the probability of the data given the (null) hypothesis. More flexible in terms of including data and information but complicated in its calculation of prior and posterior probabilities, Bayesian methods of analysis will be the topic of a future discussion. QP

REFERENCES

1. Carl Bialik, "Making a Stat Less Significant," *The Wall Street Journal,* April 2, 2011.
2. Tobias Johansson, "Hail the Impossible: P-values, Evidence and Likelihood," *Scandinavian Journal of Psychology,* Vol. 52, 2011, pp. 113-125.

Reporting statistical results and p-values / TABLE 3

State the magnitude of a meaningful (but not necessarily statistically significant) difference.

1. Report the magnitude of the statistical difference between control (null) and test (alternative) results based on the data.
2. Always give information about the hypothesis *a priori* and calculate the actual power of the test *a posteriori.*
3. Always report the actual p-value instead of stating significance based on $*p < 0.05$, $**p < 0.01$, and $***p < 0.001$, and include some p-values larger 0.05, especially for important comparisons.
4. Report conclusions based on meaningful and statistical differences, and if they do not agree, report reasons for the difference (such as sample size, effect size and variability).

3. Tom Siegfried, "Odds Are, It's Wrong: Science Fails to Face the Shortcomings of Statistics," *Science News,* Vol. 27, March 2010.
4. Global Entrepreneurship Monitor (GEM), *U.S. GEM Report,* 2009, www3.babson.edu/eship/research-publications/gem.cfm.
5. Michael Januszyk and G.C. Gurtner, "Statistics in Medicine," *Plastic and Reconstructive Surgery,* Vol. 127, No. 1, 2011, pp. 437-444.

BIBLIOGRAPHY

Gorard, Stephen, "All Evidence is Equal: The Flaw in Statistical Reasoning," *Oxford Review of Education*, Vol. 36, No. 1, 2010, pp. 63-77.

Lee, J. Jack, "Demystify Statistical Significance—Time to Move on From the P Value to Bayesian Analysis," *Journal of the National Cancer Institute,* Vol. 103, No. 1, 2010, pp. 2-3.

JULIA E. SEAMAN is a doctoral student in pharmacogenomics at the University of California, San Francisco, and a statistical consultant for the Babson Survey Research Group at Babson College. She earned a bachelor's degree in chemistry and mathematics from Pomona College in Claremont, CA.

ELAINE ALLEN is research director of the Arthur M. Blank Center for Entrepreneurship, director of the Babson Survey Research Group, and professor of statistics and entrepreneurship at Babson College in Wellesley, MA. She earned a doctorate in statistics from Cornell University in Ithaca, NY. Allen is a member of ASQ.