Putting *P* Values in Perspective

Kristin L. Sainani, PhD

P values are ubiquitous in the medical research literature, but they are often misunderstood and given more importance than they deserve. *P* values "less than .05" usually garner a lot of excitement, but in many cases, this is unwarranted. This column reviews what a *P* value is, what it offers, and what its limitations are. The reader will be advised as to what parameters are necessary to adequately interpret the *P* value.

WHAT CAN A *P* VALUE DO FOR YOU?

Take a simple hypothetical example. A team of researchers randomly assign 200 women with severe osteoporosis to 1 of 2 groups: an exercise intervention group (n = 100) or a control group (n = 100). Now suppose 30 of 100 women in the control group, as compared with only 20 of 100 women in the exercise group, sustain a fracture during follow-up. Is this conclusive evidence that the exercise intervention reduces fractures?

The average reader has been conditioned with the typical reflex response: the answer lies with the *P* value; however, stop for a moment and consider the information already at hand. Fracture risk decreased from 30% to 20% in the exercise group versus the control group (a 10% reduction in absolute risk or 33% reduction in relative risk). This **effect size** is key information in determining whether a treatment might meaningfully benefit patients (and it does not take a statistics degree to calculate or understand!).

Effect size. A measure of the magnitude of an observed effect—for example, how big the difference between groups is.

So what does the *P* value add? If a 20-women study found the same effect size—3 fractures in the control group (3 of 10 or 30%) versus 2 in the exercise group (2 of 10 or 20%)—one would likely dismiss the findings because intuition suggests that a 1-woman difference could easily arise by chance. Conversely, if a 20,000-women study found the same effect size—3000 fractures (30%) versus 2000 (20%)—one would put stock in these findings because intuition suggests that an imbalance this big is highly unlikely to arise by chance alone. Unfortunately, intuition is inadequate in all but these extreme cases. With the 200-women study, some guidance is needed to determine whether a difference of 20 of 100 versus 30 of 100 could arise by chance. The *P* value, calculated as part of a **hypothesis test,** provides this guidance.

Hypothesis test. A formal, statistical framework for deciding whether an observed effect is likely to be real or due to chance.

The *P* Value Is the Probability of the Data Given the Null Hypothesis

A hypothesis test starts with a **null hypothesis,** or a hypothesis of no effect. In this example, the null hypothesis assumes that the exercise intervention is ineffective (no better than control in preventing fractures). If this is true, each group should sustain about the same number of fractures (eg, 25 and 25). Of course, a perfectly even split is not expected. Some difference will likely arise just

Null hypothesis. The hypothesis of no effect—for example, the hypothesis that 2 groups do not differ. *P* values are calculated based on the assumption that the null hypothesis is true.

K.L.S. Division of Epidemiology, Department of Health Research and Policy, Stanford University, HRP Redwood Building, Stanford, CA 94305. Address correspondence to: K.L.S.; e-mail: kcobb@stanford.edu
Disclosure: nothing to disclose

Disclosure Key can be found on the Table of Contents and at www.pmrjournal.org

Submitted for publication July 13, 2009; accepted July 13.

PM&R
1934-1482/09/$36.00
Printed in U.S.A.

Vol. 1, 873-877, September 2009
DOI: 10.1016/j.pmrj.2009.07.003

I. IN DEPTH: HOW IS THE *P* VALUE CALCULATED MATHEMATICALLY?

The *P* value is calculated with a formal statistical test, which contains three elements: the observed effect size, the sample size, and the variability of the outcome. For this example, the appropriate test is a *Z* test of the difference between two proportions:

Effect size (difference in risk)

$$Z=\frac{10\%}{\sqrt{2\times\frac{(.25)(.75)}{100}}}$$

A measure of **variability.** The variance of a binary (yes/no) outcome is proportional to p*(1-p), where p is the overall probability of the outcome, here 25%.

Sample size (per group)

The *Z* value can be directly translated to a *P* value. For example, here $Z = 1.63$, which corresponds to a *P* value of 10% (as determined by referencing a standard normal table). Larger *Z* values give smaller (more significant) *P* values.

From the formula, it is clear that larger effect sizes and larger sample sizes result in larger *Z* values and thus smaller (more significant) *P* values, whereas increased variability in the outcome results in larger (less significant) *P* values.

There are many other statistical tests besides the *Z* test. For example, if the outcome is a continuous variable, such as blood pressure, a *t* test may be used to compare the differences in means (rather than proportions) between two groups. Like the *Z* test, the *t* test comprises the observed effect size (difference in means), a measure of variability (standard deviation of the outcome), and the sample size.

Effect size (difference in means)

$$t = \frac{mean_1 - mean_2}{\sqrt{2\times\frac{\sigma}{n}}}$$

A measure of **variability.** This is the standard deviation of a continuous outcome (such as blood pressure.

Sample size (per group)

Similar to a *Z* test, the *t* value can be directly translated to a *P* value (using a *t* distribution table).

The actual P value is calculated through statistical software. It is important for the reader to evaluate whether the appropriate statistical test has been run, but one can assume that the P *value itself has been calculated correctly.*

by chance fluctuation—but how much? Table 1 offers some answers. For example, there is a 33% chance of getting a disparity at least as big as 22 fractures in one group and 28 in the other. This probability is called the *P* value. Figure 1 graphically illustrates the *P* values for various possible outcomes. The *P* value is calculated by adding the probability of the observed outcome plus those of all the more disparate outcomes (e.g., the 33% probability is calculated by adding the probability of getting exactly 22 fractures in one group and 28 in the other plus the probability of getting exactly 21 and 29 plus the probability of getting exactly 20 and 30, and so on).

When the null hypothesis is true, *P* values of ≤.05 arise by chance about 1 in 20 times; *P* values of ≤.01 arise by chance about 1 in 100 times; and *P* values of ≤.001 arise by chance about 1 in 1000 times. Besides effect size, the *P* value also depends on the sample size and inherent variability in the outcome (see I. In-Depth sidebar), both of which are fixed in this example.

In this hypothetical study, there were 20 fractures observed in the exercise group and 30 in the control group, a 10% difference in absolute risk; this outcome has a *P* value of 10% (Table 1, Figure 1). A probability of 10% is not large, but it is not terribly small either—so the data are deemed to be consistent with the null hypothesis. On the other hand, if a more improbable outcome is obtained—such as 17 fractures in the exercise group and 33 in the control group (a 16% absolute risk difference, 1% probability under the null hypothesis)—this might lead to the rejection of the null hypothesis. Typically, a threshold of $P < .05$ is considered to be "improbable enough," but this cutoff is arbitrary and given undue importance in the medical literature [1]. A 1 in 20 false-positive threshold may be appropriate in some cases but not in others. Moreover, nothing magical happens at .05—and there is little difference in the evidence provided by a *P* value of .051 versus a *P* value of .049.

WHAT CAN'T A *P* VALUE DO FOR YOU?

It is tempting to answer a research question, such as "Does this exercise intervention reduce fractures?" by citing a *P* value. But the *P* value answers only one narrow question: is the observed effect likely the work of chance alone? Attributing additional meaning to the *P* value can lead to serious errors [2,3], described below.

Table 1. *The probabilities of various possible outcomes of the study if the null hypothesis is true and the groups have equal fracture risk (for simplicity, this table only shows where the total number of fractures is 50, but other totals are possible)*

No. fractures in the exercise group (n = 100)	No. fractures in the control group (n = 100)	Absolute risk difference (exercise − control)	*P* value: The probability of the outcome and all more disparate outcomes*
25	25	0%	1.00
24	26	2%	.74
23	27	4%	.51
22	28	6%	.33
21	29	8%	.19
20	30	10%	.10
19	31	12%	.05
18	32	14%	.02
17	33	16%	.01
25	25	0%	1.00
26	24	−2%	.74
27	23	−4%	.51
28	22	−6%	.33
29	21	−8%	.19
30	20	−10%	.10
31	19	−12%	.05
32	18	−14%	.02
33	17	−16%	.01

**P* values given are 2-sided; for example, the *P* value associated with a split of 24-26 includes the probability that the control group sustains 26 or more fractures and the probability that the exercise group sustains 26 or more fractures.

Statistical Significance Does Not Guarantee Clinical Significance

Researchers commonly equate small *P* values with large effect sizes, but *P* values and effect sizes are not equivalent [4]. As mentioned above, in addition to effect size, *P* values depend on sample size and variability. In fact, if the sample size is sufficiently large, trivial effects may achieve statistical significance. For example, if a study with 100,000 women in each of 2 groups found a 1% difference in fracture rates, this would be highly significant ($P < .0001$). (A 1000-woman difference is highly unlikely to arise just due to chance). But clearly the benefit of this treatment (for fracture reduction) is too small to warrant recommending it. Thus, it is important to pay attention to effect sizes, as well as **confidence intervals,** which give a probable range of values for the effect size. The 95% confidence interval here is 0.6% to 1.4%, which indicates that the benefit of exercise is most likely no bigger than 1.4%.

Confidence interval. A plausible range of values for the true effect size. For example, with a 95% confidence interval, one can be 95% certain that the true effect size lies in the interval. Narrower confidence intervals reflect greater precision, and wider confidence intervals reflect greater uncertainty.

Lack of Statistical Significance Is Not Proof of No Effect

Just as large samples may yield small *P* values even when effects are trivial, important effects may miss statistical significance if the sample size is too small. Small studies often have low **statistical power,** and thus a high chance of incurring **false negatives.**

Statistical power. The probability of correctly rejecting the null hypothesis—and concluding that there is an effect—if an effect truly exists.

For example, a study with 50% power has a 50% (1 in 2) chance of missing a real effect; even a study with 80% power (typically considered sufficient) has a 1 in 5 chance of a false negative. With the hypothetical exercise study, it might be tempting to conclude that the exercise intervention is ineffective based on the nonsignificant result ($P = .10$). But consider the 95% confidence interval for the difference in risk between the control and exercise groups: −2% to +22% (see II. In-Depth sidebar to learn how this is calculated). Although the authors cannot exclude a value of 0 (no benefit), most of the plausible values lie in the beneficial (positive) range. The best conclusion is that exercise might be beneficial, but larger studies are needed.

False negative, also called a type II error. A missed effect; failing to reject the null hypothesis when an effect exists. Statistical power is the probability of avoiding a type II error.

False Positives Are More Common Than You Think

Hypothesis testing comes with an inherent **false-positive** rate. Using a significance threshold of .05, the false-positive rate for a

False positive, also called a type I error. A chance finding; rejecting the null hypothesis and concluding that there is an effect when none exists.

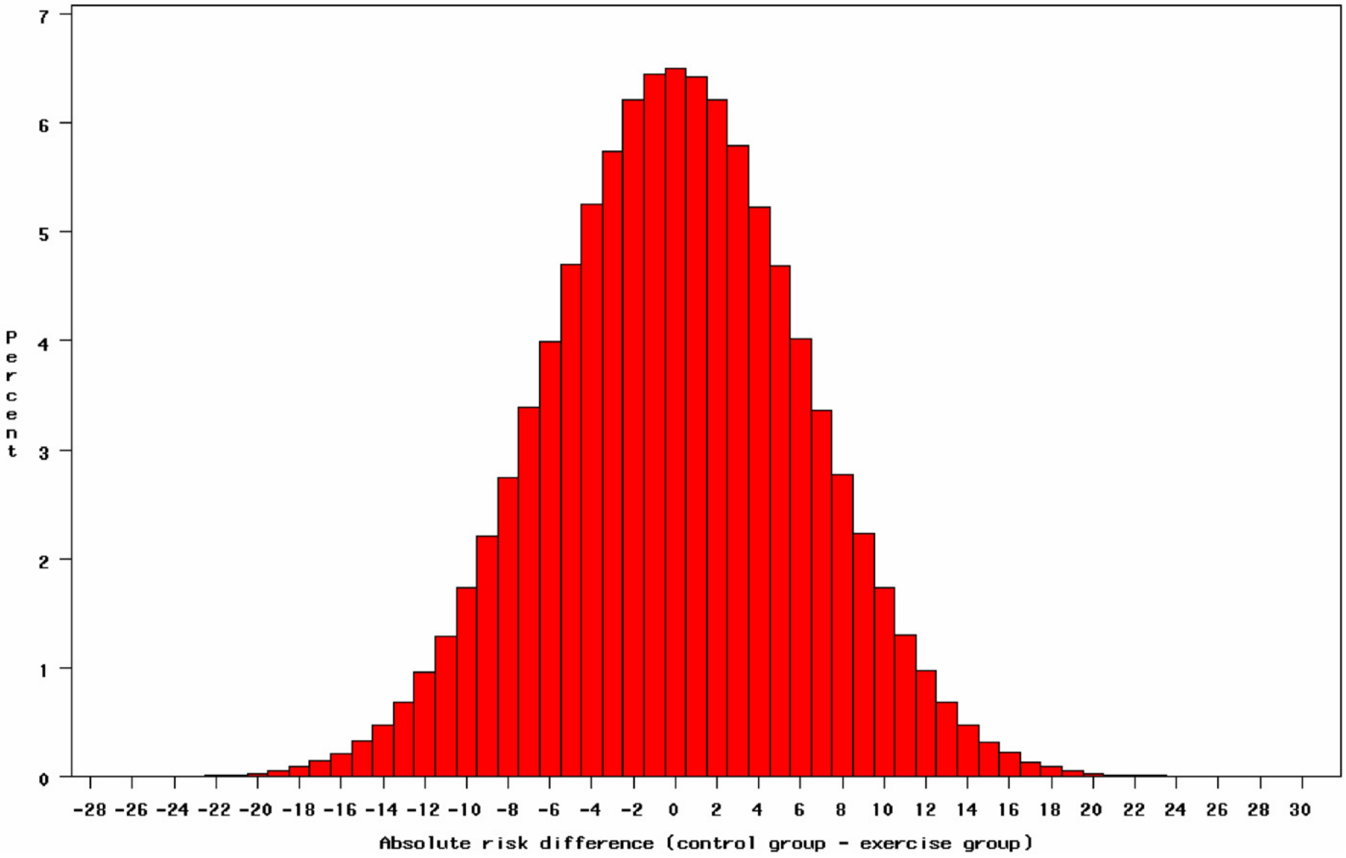

Figure 1. This histogram shows the probability of different outcomes if the null hypothesis is true. Outcomes are presented as absolute risk differences (% of women who fracture in the control group − % of women who fracture in the exercise group). The difference in the proportion of women fracturing in each group will most likely fall around 0%. However, because of random fluctuation, differences as big or bigger than 10% (in either tail) have a 10% chance of occurring. This 10% probability value is called a *P* value. (How to read this histogram: In this histogram, the *y* axis shows the probability of each outcome, where the outcomes are broken into 1 − percentage point intervals (1%, 2%, 3%, etc.). Because there are more than 40 different possible outcomes here, the probability of any particular outcome is low. For example, the probability of a difference of exactly 0% (the most likely outcome) is only about 6.5%. The 10% *P* value is calculated by adding up the probabilities of all the outcomes 10% and greater (11, 12, 13, 14, 15, etc.) and the probabilities of all the outcomes −10% and less (−11%, −12%, −13%, −14%, −15%, etc.). Another way to think of the *P* value is as the area under the curve in the tails.)

single test is 1 in 20. But when multiple tests are run, each with a 1 in 20 false-positive rate, the chance of incurring at least one false positive increases. For example, if you run 20 tests in which no effects exist, the chance of getting at least one *P* value under .05 is as high as 64%. Thus, when a few *P* values squeak under .05 in a sea of statistical tests, these are likely more consistent with chance findings than real effects—particularly if *P* values are moderate in size (between .05 and .01 [5]), derived from exploratory rather than previously planned analyses, and not adjusted for multiple testing. The problem of multiple testing will be the subject of a future column.

Statistical Significance Is Not Proof of Causation

Finally, the *P* value reveals nothing about the study design or methods that were used to generate the data. Statistical findings must be interpreted in the context of how the data were collected. If a statistically significant benefit for the exercise group was found, one still could not conclude that exercise and fracture were causally related; that would require additional information about the quality of the trial, and its consistency with the results of other studies.

CONCLUSION

The *P* value is seductive because it reduces complex and messy data into a simple, neat number. But *P* values can be misleading and should only be considered in the context of effect size, sample size, the number of tests run, and study design and implementation. Where possible, it may be preferable to report effect sizes with confidence intervals in lieu of or in addition to *P* values [4].

II. IN DEPTH: HOW IS A CONFIDENCE INTERVAL CALCULATED MATHEMATICALLY?

The formula for a confidence interval contains the same elements as a hypothesis test (effect size, variability, and sample size) plus a factor that confers the level of confidence. For example, with the hypothetical exercise trial, the 95% confidence interval for the difference in risk of fractures between the groups is:

Effect size (difference in proportions)

Measure of **variability.**

$$10\% \pm 1.96 \times \sqrt{2 \times \frac{(.25)(.75)}{100}} = (-2\%, +22\%)$$

Factor for **95% confidence**

Sample size (per group)

Because they are mathematically similar to a hypothesis test, confidence intervals can be used to determine statistical significance. If the 95% confidence interval contains the null value (here 0%), then the null value cannot be ruled out with 95% certainty, and the corresponding *P* value will be greater than .05; if the 95% confidence interval excludes the null value, however, then the *P* value will be less than .05. If a 99% confidence interval excludes the null value, then the *P* value will be less than .01, and so on. But confidence intervals convey more information than *P* values—they give a probable range of values for the effect size as well as a sense of the precision of the estimate.

REFERENCES

1. Sterne JA, Smith GD. Sifting through the evidence—what's wrong with significance tests? BMJ 2001;322:226-231.
2. Cox DR. Statistical significance tests. Br J Clin Pharmacol 1982;14:325-331.
3. Cohen J. Things I have learned (so far). Am Psychol 1990;45:1304-1312.
4. Lang JM, Rothman KJ, Cann CI. That confounded p-value. Epidemiology 1998;9:7-8.
5. Ioannidis JPA. Effect of formal statistical significance on the credibility of observational associations. Am J Epidemiol 2008;168:374-383.