Primer on Statistical Significance and *P* Values

In the world of medical journals, few phrases evoke more authority than "the differences observed were statistically significant." Unfortunately, readers frequently accord too much importance to this statement and are often distracted from more pressing issues. This Primer reviews the meaning of the term *statistical significance* and includes some important caveats for critical readers to consider whenever it is used.

Assessing the Role of Chance

Consider a study of a new weight loss program: Group A receives the intervention and loses an average of 10 pounds, while group B serves as a control and loses an average of 3 pounds. The main effect of the weight loss program is therefore estimated to be a 7-pound weight loss (on average). But we would rarely expect that any two groups would have exactly the same amount of weight change. So could it just be chance that group A lost more weight?

There are two basic statistical methods used to assess the role of chance: confidence intervals (the subject of next issue's Primer) and hypothesis testing. As shown in the Figure below, both use the same fundamental inputs.

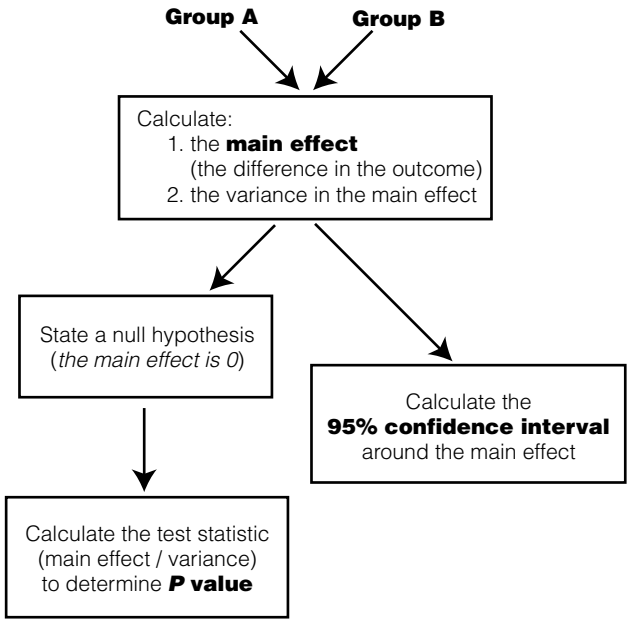

FIGURE 1. Statistical approach to comparing two groups.

Hypothesis testing goes on to consider a condition—the null hypothesis—that no difference exists. In this case, the null hypothesis is that the weight change in the two groups is the same. The test addresses the question, "If the true state of affairs is no difference (i.e., the null hypothesis is true), what is the probability of observing this difference (i.e., 7 lbs) or one more extreme (i.e., 8 lbs, 9 lbs, etc.)"? This probability is called the *P* value and, for most of us, translates roughly to "the probability that the observed result is due to chance."

If the *P* value is less than 5%, researchers typically assert that the findings are "statistically significant." In the case of the weight loss program, if the chance of observing a difference of 7 pounds or more (when, in fact, none exists) is less than 5 in 100, then the weight loss program is presumed to have a real effect.

TABLE 1
Relationship between Common Language and Hypothesis Testing

COMMON LANGUAGE	STATISTICAL STATEMENT	CONVENTIONAL TEST THRESHOLD
"Statistically significant" "Unlikely due to chance"	The null hypothesis was rejected.	$P < 0.05$
"Not significant" "Due to chance"	The null hypothesis could not be rejected.	$P > 0.05$

Table 1 shows how our common language relates to the statistical language of hypothesis testing.

Factors That Influence *P* Values

Statistical significance (meaning a low *P* value) depends on three factors: the main effect itself and the two factors that make up the variance. Here is how each relates to the *P* value:

- *The magnitude of the main effect.* A 7-lb difference will have a lower *P* value (i.e., more likely to be statistically significant) than a 1-lb difference.
- *The number of observations.* A 7-lb difference observed in a study with 500 patients in each group will have a lower *P* value than a 7-lb difference observed in a study with 25 patients in each group.
- *The spread in the data (commonly measured as a standard deviation).* If everybody in group A loses about 10 pounds and everybody in group B loses about 3 pounds, the *P* value will be lower than if there is a wide variation in individual weight changes (even if the group averages remain at 10 and 3 pounds). Note: More observations do not reduce spread in data.

Caveats about the Importance of *P* Values

Unfortunately, *P* values and statistical significance are often accorded too much weight. Critical readers should bear three facts in mind:

1. *The* P *< 0.05 threshold is wholly arbitrary.*

There is nothing magical about a 5% chance—it's simply a convenient convention and could just as easily be 10% or 1%. The arbitrariness of the 0.05 threshold is most obvious when *P* values are near the cut-off. To call one finding significant when the *P* value is 0.04 and another not significant when it is 0.06 vastly overstates the difference between the two findings.

Critical readers should also realize that dichotomizing *P* values into simply "significant" and "insignificant" loses information in the same way that dichotomizing any clinical laboratory value into "normal" and "abnormal" does. Although serum sodium levels of 115 and 132 are both below normal, the former is of much greater concern than the latter. Similarly, although both are significant, a *P* value of 0.001 is much more "significant" than a *P* value of 0.04.

2. *Statistical significance does not translate into clinical importance.*

Although it is tempting to equate statistical significance with clinical importance, critical readers should avoid this temptation. To be clinically important requires a substantial change in an outcome that matters. Statistically significant changes, however, can be observed with trivial outcomes. And because significance is powerfully influenced by the number of observations, statistically significant changes can be observed with trivial changes in important outcomes. As shown in Table 2, large studies can be significant without being clinically important and small studies may be important without being significant.

3. *Chance is rarely the most pressing issue.*

Finally, because *P* values are quantifiable and seemingly objective, it's easy to overemphasize the importance of statistical significance. For most studies, the biggest threat to an author's conclusion is not random error (chance), but systematic error (bias). Thus, readers must focus on the more difficult, qualitative questions: Are these the right patients? Are these the right outcomes? Are there measurement biases? Are observed associations confounded by other factors?

TABLE 2

Big Studies Make Small Differences "Significant"*

SIZE, *n* (IN EACH GROUP)	WEIGHT LOSS		MAIN EFFECT	*P* VALUE	APPROPRIATE CONCLUSION
	GROUP A (INTERVENTION)	GROUP B (CONTROL)			
10	20 lb	3 lb	17 lb	0.07	Not significant, but promising
1000	5 lb	3 lb	2 lb	0.03	Significant, but clinically unimportant

**The standard deviation of the weight change is assumed to be 20 lb.*

A compendium of **ecp** *primers from past issues can be viewed and/or requested at http://www.acponline.org/journals/ecp/primers.htm.*