Lecture 18

More about tests and Intervals

CHOONG-WAN WOO | COCOAN lab | http://cocoanlab.github.io

Review: Key Points

Chapter 21: Inferences About Means

- Central Limit Theorem: $Mean(\bar{y}) = \mu, SD(\bar{y}) = \frac{\sigma}{\sqrt{n}}$, standard error, $SE(\bar{y}) = \frac{s}{\sqrt{n}}$
- Degrees of freedom: the number of values that are free to vary after we estimate parameters.
- Confidence interval for Means: $\bar{y} \pm t^*_{df} \times SE(\bar{y})$, where $SE(\bar{y}) = \frac{s}{\sqrt{n}}$, and $df = n - 1$ (t^* instead of z^*)
- Hypothesis tests and confidence interval are built from the same calculations, and looking at the same question from two different perspectives.
- Hypothesis tests start with a proposed *parameter value* and ask if the *data* are consistent with that value, while confidence intervals start with the *data* and finds an interval of *plausible values* for where the parameter may lie.

What's P-value again?

- P-value = P(Data | H_0), not P(H_0 | Data)
  - The conditional probability of getting the data given that the null hypothesis is true
  - *NOT* the probability that the null hypothesis is true
  - *NOT* the conditional probability that the null hypothesis is true given the data
  - P-value = 0.03
    - does *NOT* mean "there is a 3% chance that the null hypothesis is true".
    - does mean "given the null hypothesis, there's a 3% chance of observing the observed statistic value.

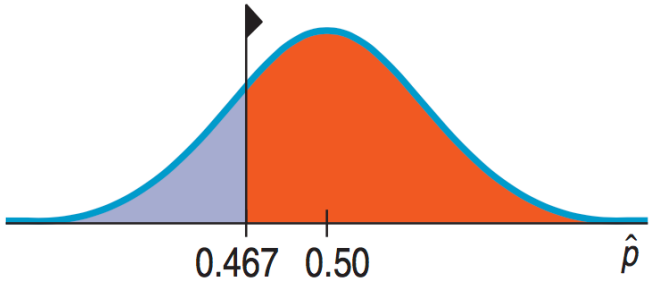
Small P-value

- First, yay!
- It means the result we just observed is unlikely to occur if the null hypothesis is true.
  - does *NOT* mean that the null hypothesis is “less true”.

- How small the P-value should be?
  - depends on a lot of things, e.g., your prior belief in the null hypothesis, your trust in your data, in the experimental method, in the survey protocol, etc.
  - P-value serve as a measure of the strength of the evidence against the null hypothesis
  - should *NEVER* serve as a hard and fast rule for decisions.
  - *YOU* have to take the responsibility for the decision on yourself.

High P-value

- No evidence for rejecting H_0
- We cannot reject the null hypothesis.
- For one-sided test, if P-value is higher than 0.5, you know that your test statistic is on the "wrong" side.



- High P-values mean
  - What we've observed is not surprising.
  - We have no reason to reject our null hypothesis.
  - Does *NOT* prove that the null hypothesis is true
  - Do *NOT* say that you "accept the null hypothesis".
  - You *should* say that "the data have failed to provide sufficient evidence to reject the null hypothesis".

Alpha levels

- We need to make *decisions*!

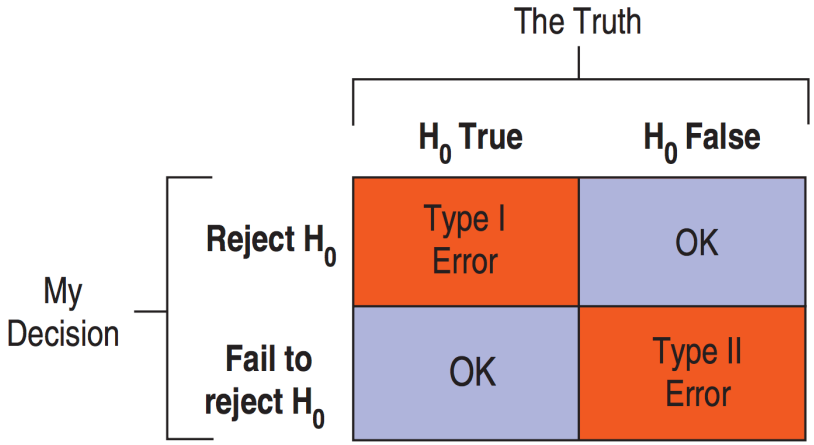
- An **alpha** level defines how small the P-value should be:
  - Threshold for a decision
  - Significance level, "*statistically significant*"
  - Greek letter, α
  - Common α levels are 0.05, 0.01, and 0.001

- Important considerations:
  - The thresholds (α levels) are arbitrary, really. But we need something for decision-making.
  - You must select the alpha level *BEFORE* you look at the data.
  - In some fields, the decision about the significance is presented with *s.
    - E.g., * $p < .05$, ** $p < .01$, *** $p < .001$
  - But still, it's better to report the P-value in addition to the *s.

Type I and II errors

- We can make wrong decisions.

- Two types of errors:
  - **Type I error:** The null hypothesis is true, but we mistakenly reject it. – *False positive*
  - **Type II error**: The null hypothesis is false, but we fail to reject it. – *False negative*



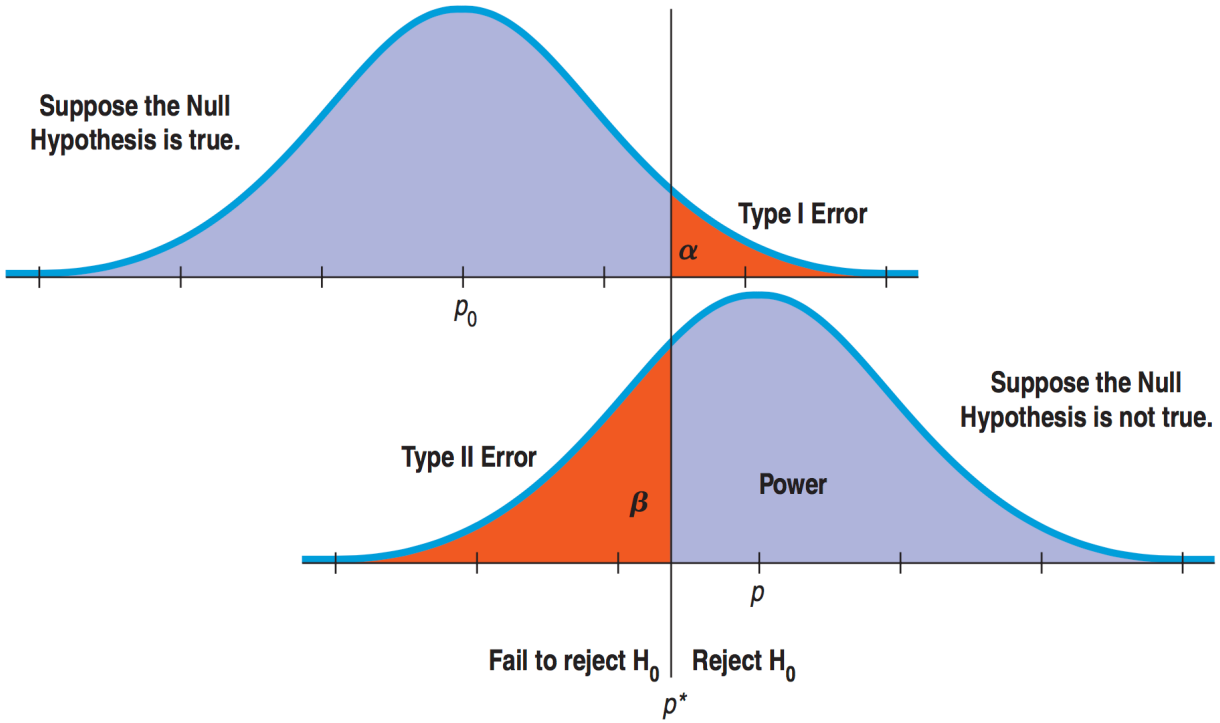
More about the errors

- *P*(Type I error) = α
  - To reject H_0, the P-value must fall below α.
  - It's actually same with setting the probability of a Type I error to α.
  - A type I error can happen only when H_0 is true.

- *P*(Type II error) = β
  - H_0 is false but we fail to reject it
  - β, the probability of this type of mistake
  - β is harder to assess than α because we don't know the true value of the parameter.

- Tradeoff between Type I and II errors
  - Reducing Type I error can increase Type II error, and vice versa.

Alpha and beta in a picture (1)



- Tradeoffs between alpha and beta

Why most popular alpha = 0.05?

FISHER AND $\alpha = 0.05$

Why did Sir Ronald Fisher suggest 0.05 as a criterion for testing hypotheses? It turns out that he had in mind small initial studies. Small studies have relatively little power. Fisher was concerned that they might make too many Type II errors—failing to discover an important effect—if too strict a criterion were used. Once a test failed to reject a null hypothesis, it was unlikely that researchers would return to that hypothesis to try again.

On the other hand, the increased risk of Type I errors arising from a generous criterion didn't concern him as much for exploratory studies because these are ordinarily followed by a replication or a larger study. The probability of a Type I error is α—in this case, 0.05. The probability that two independent studies would both make Type I errors is $0.05 \times 0.05 = 0.0025$, so Fisher was confident that Type I errors in initial studies were not a major concern.

The widespread use of the relatively generous 0.05 criterion even in large studies is most likely not what Fisher had in mind.

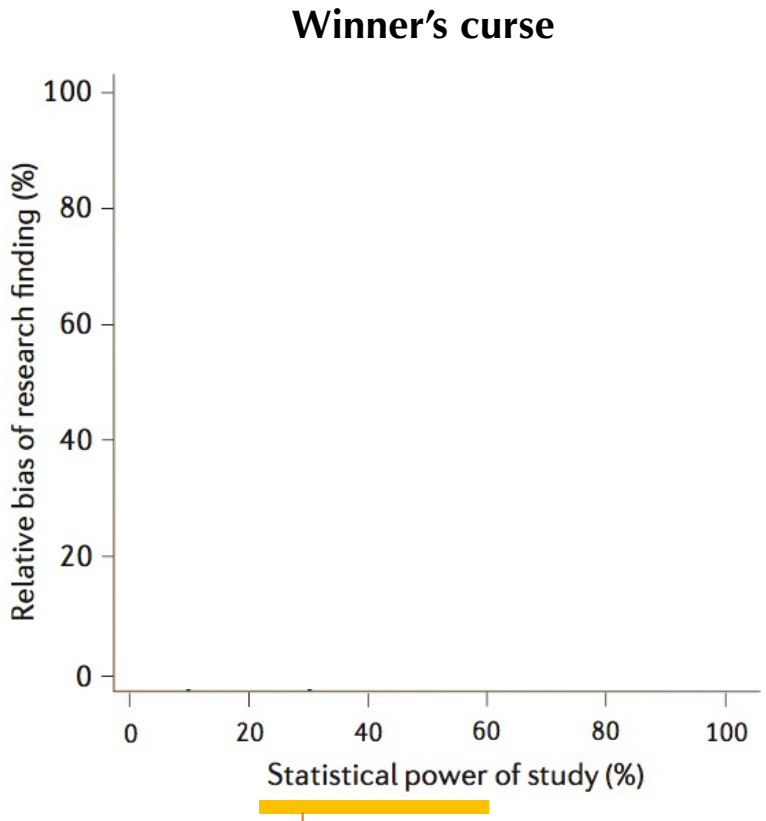
Power

- Remember, we can never prove a null hypothesis is true. We can only fail to reject it.
- When we fail to reject a null hypothesis, we can ask whether we looked hard enough.
- We hope our test is **STRONG** enough to reject it.

- Power of a test: the probability that it correctly rejects a false null hypothesis
  - With high powered test, we can be confident that we've looked hard enough even if our test failed to reject a null hypothesis.
  - Type II error (β) is the probability that a test fails to reject a false null hypothesis.
  - The **power (or statistical power)** is the probability that it **correctly** rejects a false null hypothesis
    - 1 – P(Type II error) = 1 - β

  - Whenever a study fails to reject its null hypothesis, the test's power comes into question.
  - But only when a study fails to reject the null hypothesis? NO!

Winner's curse

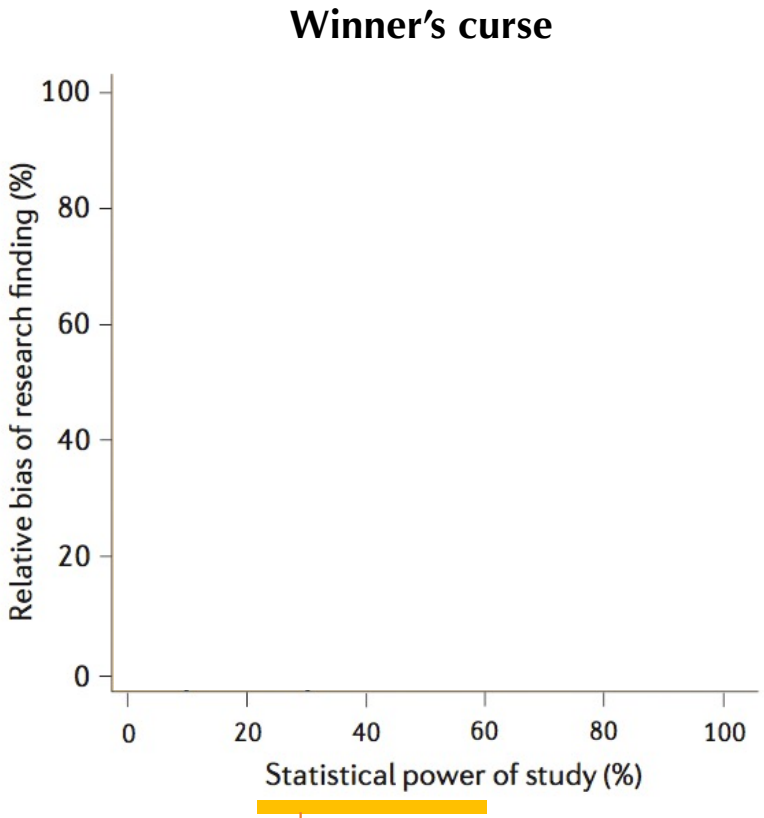


↑ Sample size
↑ Effect size
↓ Threshold
↑ Power

The probability that it will correctly reject the null hyp
othesis when the null hypothesis is false

From Button et al., 2013, *Nat Rev Neurosci*, Zollner & Pritchard, 2007, *Am J Hum Genet*

Winner's curse

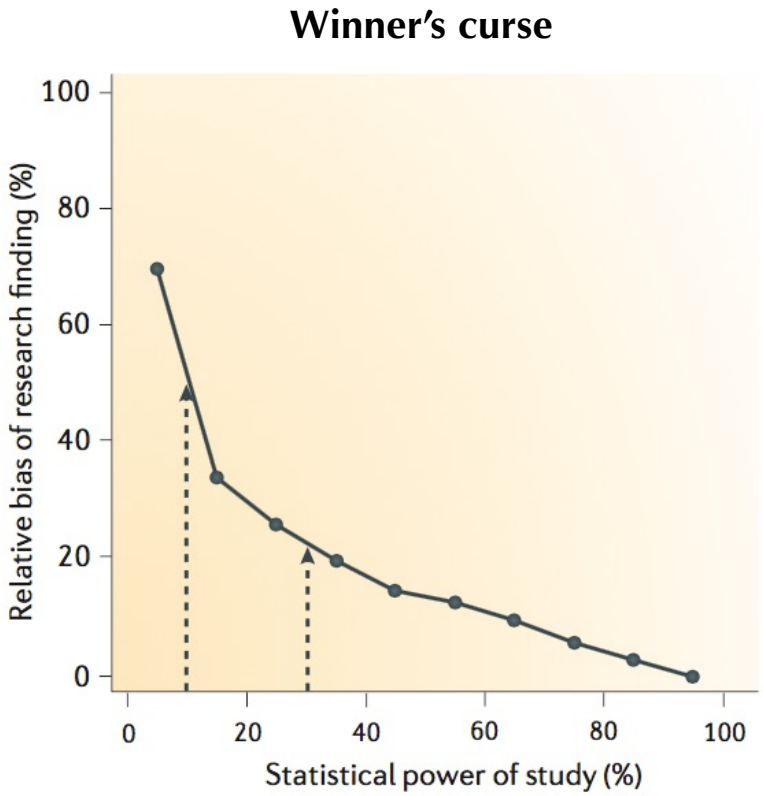


Stronger results than real (e.g., voodoo correlation)

The probability that it will correctly reject the null hyp othesis when the null hypothesis is false

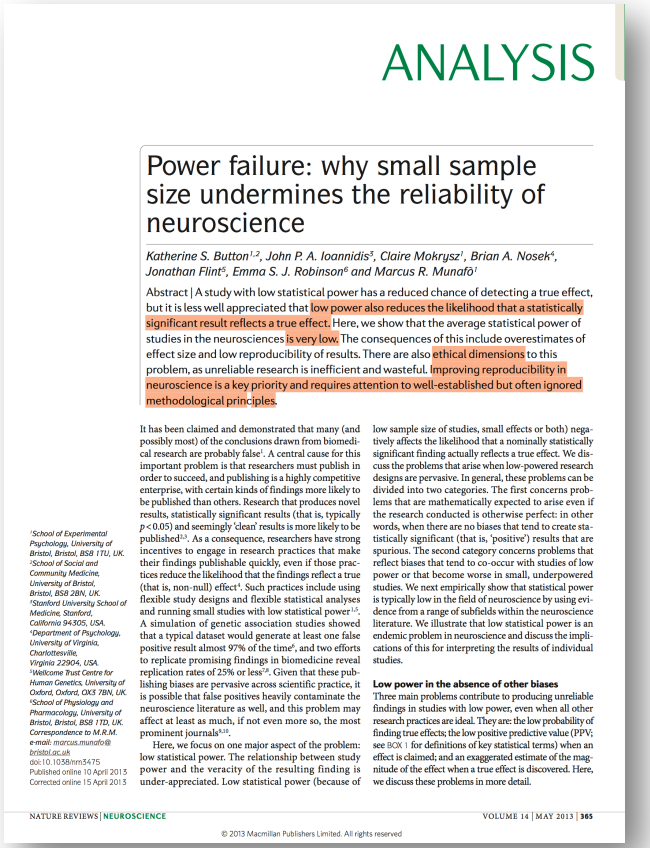
From Button et al., 2013, *Nat Rev Neurosci*, Zollner & Pritchard, 2007, *Am J Hum Genet*

Winner's curse



- Low powered studies can increase biases in literature, resulting in replication failure.

Power failure: references



ANALYSIS

Power failure: why small sample size undermines the reliability of neuroscience

Katherine S. Button[1,2], John P. A. Ioannidis[3], Claire Mokrysz[1], Brian A. Nosek[4], Jonathan Flint[5], Emma S. J. Robinson[6] and Marcus R. Munafò[1]

Abstract | A study with low statistical power has a reduced chance of detecting a true effect, but it is less well appreciated that low power also reduces the likelihood that a statistically significant result reflects a true effect. Here, we show that the average statistical power of studies in the neurosciences is very low. The consequences of this include overestimates of effect size and low reproducibility of results. There are also ethical dimensions to this problem, as unreliable research is inefficient and wasteful. Improving reproducibility in neuroscience is a key priority and requires attention to well-established but often ignored methodological principles.

It has been claimed and demonstrated that many (and possibly most) of the conclusions drawn from biomedical research are probably false[1]. A central cause for this important problem is that researchers must publish in order to succeed, and publishing is a highly competitive enterprise, with certain kinds of findings more likely to be published than others. Research that produces novel results, statistically significant results (that is, typically $p < 0.05$) and seemingly 'clean' results is more likely to be published[2,3]. As a consequence, researchers have strong incentives to engage in research practices that make their findings publishable quickly, even if those practices reduce the likelihood that the findings reflect a true (that is, non-null) effect[4]. Such practices include using flexible study designs and flexible statistical analyses and running small studies with low statistical power[1,5]. A simulation of genetic association studies showed that a typical dataset would generate at least one false positive result almost 97% of the time[6], and two efforts to replicate promising findings in biomedicine reveal replication rates of 25% or less[7,8]. Given that these publishing biases are pervasive across scientific practice, it is possible that false positives heavily contaminate the neuroscience literature as well, and this problem may affect at least as much, if not even more so, the most prominent journals[9,10].

Here, we focus on one major aspect of the problem: low statistical power. The relationship between study power and the veracity of the resulting finding is under-appreciated. Low statistical power (because of low sample size of studies, small effects or both) negatively affects the likelihood that a nominally statistically significant finding actually reflects a true effect. We discuss the problems that arise when low-powered research designs are pervasive. In general, these problems can be divided into two categories. The first concerns problems that are mathematically expected to arise even if the research conducted is otherwise perfect: in other words, when there are no biases that tend to create statistically significant (that is, 'positive') results that are spurious. The second category concerns problems that reflect biases that tend to co-occur with studies of low power or that become worse in small, underpowered studies. We next empirically show that statistical power is typically low in the field of neuroscience by using evidence from a range of subfields within the neuroscience literature. We illustrate that low statistical power is an endemic problem in neuroscience and discuss the implications of this for interpreting the results of individual studies.

Low power in the absence of other biases

Three main problems contribute to producing unreliable findings in studies with low power, even when all other research practices are ideal. They are: the low probability of finding true effects; the low positive predictive value (PPV; see BOX 1 for definitions of key statistical terms) when an effect is claimed; and an exaggerated estimate of the magnitude of the effect when a true effect is discovered. Here, we discuss these problems in more detail.

[1]School of Experimental Psychology, University of Bristol, Bristol, BS8 1TU, UK. [2]School of Social and Community Medicine, University of Bristol, Bristol, BS8 2BN, UK. [3]Stanford University School of Medicine, Stanford, California 94305, USA. [4]Department of Psychology, University of Virginia, Charlottesville, Virginia 22904, USA. [5]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, OX3 7BN, UK. [6]School of Physiology and Pharmacology, University of Bristol, Bristol, BS8 1TD, UK. Correspondence to M.R.M. e-mail: marcus.munafo@bristol.ac.uk

doi:10.1038/nrn3475
Published online 10 April 2013
Corrected online 15 April 2013

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

Published research findings are sometimes refuted by subsequent evidence, with ensuing confusion and disappointment. Refutation and controversy is seen across the range of research designs, from clinical trials and traditional epidemiological studies [1–3] to the most modern molecular research [4,5]. There is increasing concern that in modern research, false findings may be the majority or even the vast majority of published research claims [6–8]. However, this should not be surprising. It can be proven that most claimed research findings are false. Here I will examine the key factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a p-value less than 0.05. Research is not most appropriately represented and summarized by p-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on p-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. "Negative" research is also very useful. "Negative" is actually a misnomer, and the misinterpretation is widespread. However, here we will target relationships that investigators claim exist, rather than null findings.

As has been shown previously, the probability that a research finding is indeed true depends on the prior probability of it being true (before doing the study), the statistical power of the study, and the level of statistical significance [10,11]. Consider a 2×2 table in which research findings are compared against the gold standard of true relationships in a scientific field. In a research field both true and false hypotheses can be made about the presence of relationships. Let R be the ratio of the number of "true relationships" to "no relationships" among those tested in the field. R is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R+1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2×2 table, one gets $\text{PPV} = (1 - \beta)R/(R - \beta R + \alpha)$. A research finding is thus

The Essay section contains opinion pieces on topics of broad interest to a general medical audience.

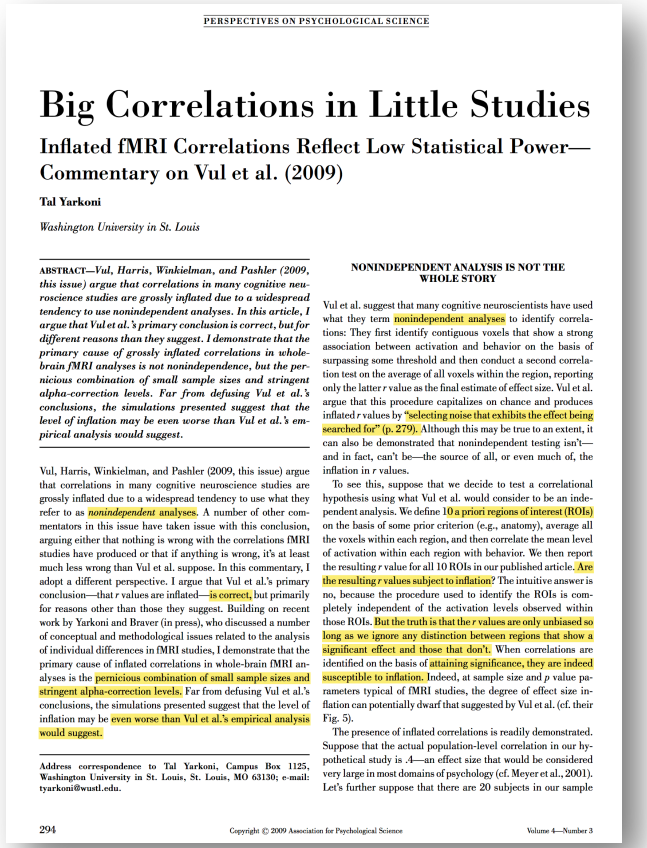
Citation: Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2(8): e124.

Abbreviation: PPV, positive predictive value

John P.A. Ioannidis is in the Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece, and Institute for Clinical Research and Health Policy Studies, Department of Medicine, Tufts-New England Medical Center, Tufts University School of Medicine, Boston, Massachusetts, United States of America. E-mail: jioannid@cc.uoi.gr

Competing Interests: The author has declared that no competing interests exist.

DOI: 10.1371/journal.pmed.0020124



Big Correlations in Little Studies

Inflated fMRI Correlations Reflect Low Statistical Power—Commentary on Vul et al. (2009)

Tal Yarkoni

Washington University in St. Louis

ABSTRACT—*Vul, Harris, Winkielman, and Pashler (2009, this issue) argue that correlations in many cognitive neuroscience studies are grossly inflated due to a widespread tendency to use nonindependent analyses. In this article, I argue that Vul et al.'s primary conclusion is correct, but for different reasons than they suggest. I demonstrate that the primary cause of grossly inflated correlations in whole-brain fMRI analyses is not nonindependence, but the pernicious combination of small sample sizes and stringent alpha-correction levels. Far from defusing Vul et al.'s conclusions, the simulations presented suggest that the level of inflation may be even worse than Vul et al.'s empirical analysis would suggest.*

Vul, Harris, Winkielman, and Pashler (2009, this issue) argue that correlations in many cognitive neuroscience studies are grossly inflated due to a widespread tendency to use what they refer to as *nonindependent* analyses. A number of other commentators in this issue have taken issue with this conclusion, arguing either that nothing is wrong with the correlations fMRI studies have produced or that if anything is wrong, it's at least much less wrong than Vul et al. suppose. In this commentary, I adopt a different perspective. I argue that Vul et al.'s primary conclusion—that r values are inflated—is correct, but primarily for reasons other than those they suggest. Building on recent work by Yarkoni and Braver (in press), who discussed a number of conceptual and methodological issues related to the analysis of individual differences in fMRI studies, I demonstrate that the primary cause of inflated correlations in whole-brain fMRI analyses is the pernicious combination of small sample sizes and stringent alpha-correction levels. Far from defusing Vul et al.'s conclusions, the simulations presented suggest that the level of inflation may be even worse than Vul et al.'s empirical analysis would suggest.

Address correspondence to Tal Yarkoni, Campus Box 1125, Washington University in St. Louis, St. Louis, MO 63130; e-mail: tyarkoni@wustl.edu.

NONINDEPENDENT ANALYSIS IS NOT THE WHOLE STORY

Vul et al. suggest that many cognitive neuroscientists have used what they term nonindependent analyses to identify correlations: They first identify contiguous voxels that show a strong association between activation and behavior on the basis of surpassing some threshold and then conduct a second correlation test on the average of all voxels within the region, reporting only the latter r value as the final estimate of effect size. Vul et al. argue that this procedure capitalizes on chance and produces inflated r values by "selecting noise that exhibits the effect being searched for" (p. 279). Although this may be true to an extent, it can also be demonstrated that nonindependent testing isn't—and in fact, can't be—the source of all, or even much of, the inflation in r values.

To see this, suppose that we decide to test a correlational hypothesis using what Vul et al. would consider to be an independent analysis. We define 10 a priori regions of interest (ROIs) on the basis of some prior criterion (e.g., anatomy), average all the voxels within each region, and then correlate the mean level of activation within each region with behavior. We then report the resulting r value for all 10 ROIs in our published article. Are the resulting r values subject to inflation? The intuitive answer is no, because the procedure used to identify the ROIs is completely independent of the activation levels observed within those ROIs. But the truth is that the r values are only unbiased so long as we ignore any distinction between regions that show a significant effect and those that don't. When correlations are identified on the basis of attaining significance, they are indeed susceptible to inflation. Indeed, at sample size and p value parameters typical of fMRI studies, the degree of effect size inflation can potentially dwarf that suggested by Vul et al. (cf. their Fig. 5).

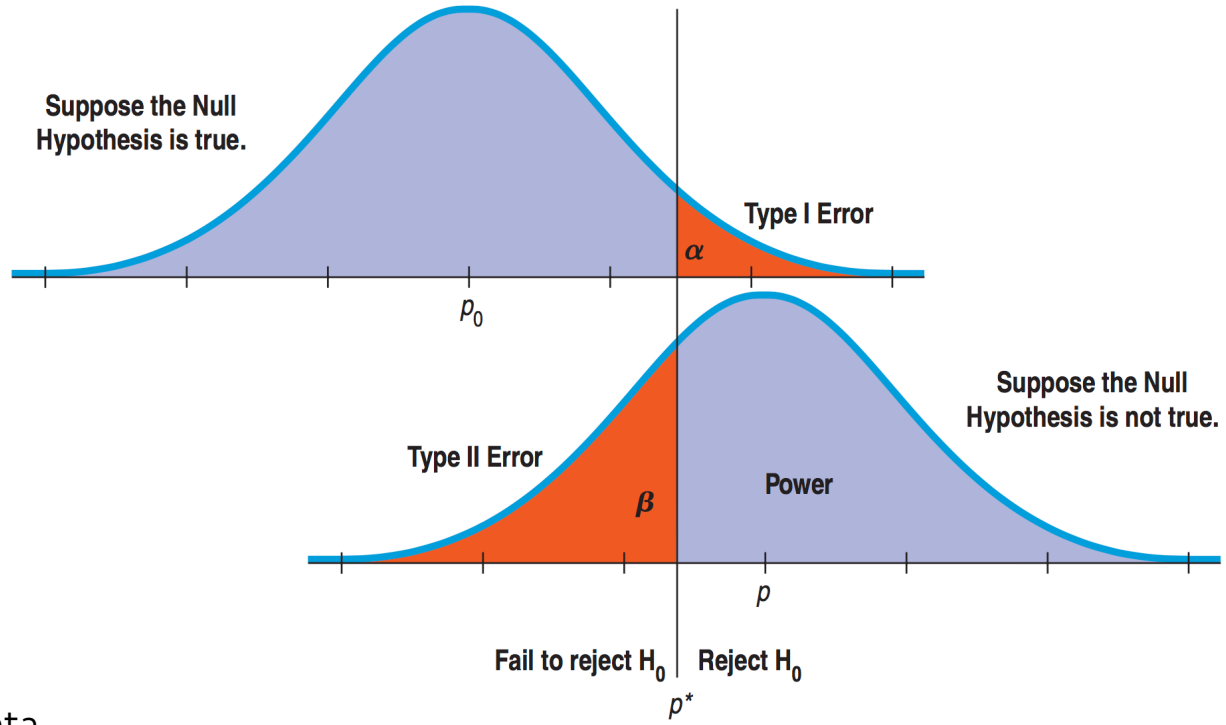
The presence of inflated correlations is readily demonstrated. Suppose that the actual population-level correlation in our hypothetical study is .4—an effect size that would be considered very large in most domains of psychology (cf. Meyer et al., 2001). Let's further suppose that there are 20 subjects in our sample

Effect size

- **Effect size:** the distance between the null hypothesis value, p_0 or $\mu_0$and the truth, p or μ.
- Power ∝ effect size
  - Large effects are easier to see and results in larger power.
  - Small effects are naturally more difficult to detect (increase Type II errors).
  - The standard deviation is also important (it provides a ruler).
  - Effect magnitude / standard deviation is basically **"signal-to-noise ratio"**.

- **Different types** of effect sizes: correlation-based, distance-based, etc.
  - Correlation, r and R^2
  - Cohen's $d = \frac{\bar{x}_1 - \bar{x}_2}{s}$
    - similar to z scores, $z = \frac{x_i - \bar{x}}{s}$, but Cohen's d is for comparisons between groups

Alpha and beta in a picture (2)



- Power = 1 – beta
- Tradeoffs between alpha and beta
- The role of effect size in power

Sensitivity and Specificity

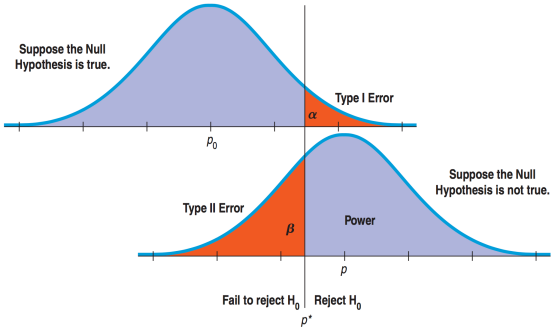
https://en.wikipedia.org/wiki/Sensitivity_and_specificity

- **Specificity** = # true negatives / (# true negatives + # false positives) = 1 - α
- **Sensitivity** = # true positives / (# true positives + # false negatives) = 1 - β = Power

Confusion matrix

		True condition	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error
	Predicted condition negative	**False negative**, Type II error	**True negative**

Reducing both Type I and II errors



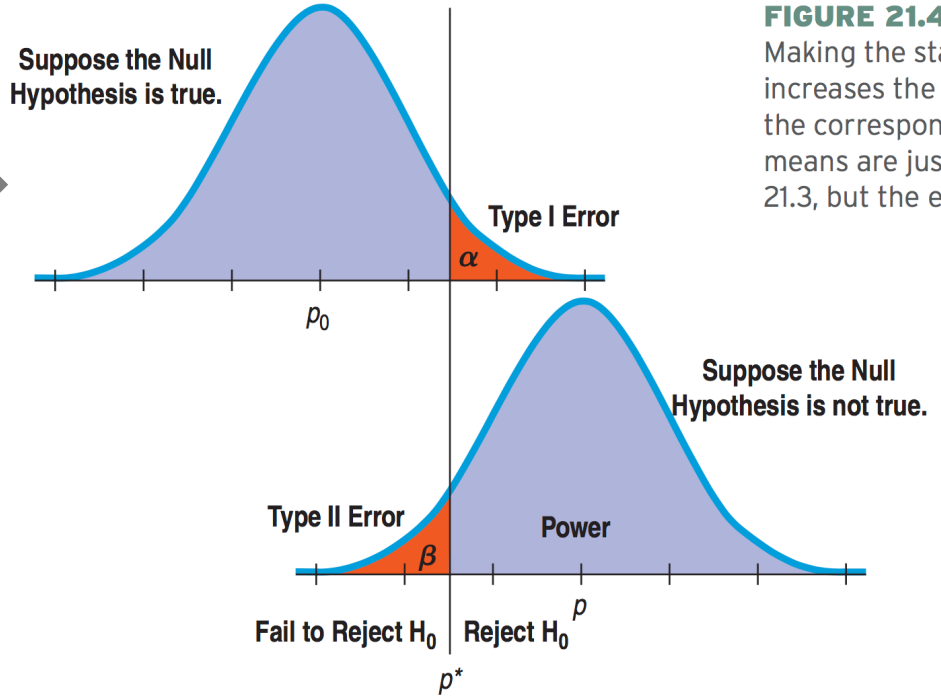


FIGURE 21.4
Making the standard deviations smaller increases the power without changing the corresponding critical value. The means are just as far apart as in Figure 21.3, but the error rates are reduced.

Increasing Power without increasing sample size:

er, these resampling methods free psychologists from the restrictive assumptions of normal statistical theory and permit them to gauge the reliability of chosen statistics by making thousands, even millions, of calculations on many data points. Tukey (e.g., 1980) has long suggested that data analysis is essentially a two-stage compound process where the patterns suggested by exploratory data analysis are critically checked through the use of the jackknife rather than classical inferential procedures. With psychologists at last being encouraged to use exploratory data analysis procedures, it would be reassuring to think that with the increasing availability of suitable software, statistical resampling methods might soon become a companion resource in the psychological researcher's tool kit.

Guidelines for Theory Construction

My concluding comment is a plea directed to the Board of Scientific Affairs: Granted, there are many different tasks in scientific research, but the Board of Scientific Affairs and subsequently the *Publication Manual of the American Psychological Association* (American Psychological Association, 1994) should do for other parts of the research process what they have done and continue to do for statistical data analysis. In particular, they need to provide helpful guidelines and decent explanatory justifications for carrying out the various tasks of theory construction—that half of scientific research whose methodology is steadfastly ignored in American Psychological Association-sponsored deliberations and recommendations. Good science is as much concerned with the explanation of empirical phenomena as it is with their initial detection. Although psychologists' statistical methods speak primarily to the detection of phenomena (cf. Woodward, 1989), the past three decades have witnessed significant advances in the methodology of theory construction. With the codification of methods of theory generation, theory development, and theory appraisal, psychology is now positioned to give its researchers helpful methodological advice on how to build worthwhile explanatory theories.

REFERENCES

American Psychological Association. (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington, DC: Author.

Curd, M. V. (1980). The logic of discovery: An analysis of three approaches. In T. Nickles (Ed.), *Scientific discovery, logic, and rationality* (pp. 201–219). Dordrecht, the Netherlands: Reidel.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Giere, R. N. (1983). Testing theoretical hypotheses. In J. Earman (Ed.), *Testing scientific theories* (pp. 269–298). Minneapolis, MN: University of Minnesota Press.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66,* 195–244.

Rozeboom, W. W. (1997). Good science is abductive, not hypothetico-deductive. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 335–391). Hillsdale, NJ: Erlbaum.

Tukey, J. W. (1980). We need both exploratory and confirmatory. *American Statistician, 34,* 23–25.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Woodward, J. (1989). Data and phenomena. *Synthese, 79,* 393–472.

Correspondence concerning this comment should be addressed to Brian D. Haig, Department of Psychology, University of Canterbury, Private Bag 4800, Christchurch, New Zealand. Electronic mail may be sent to b.haig@psyc.canterbury.ac.nz.

DOI: 10.1037//0003-066X.55.8.963

Increasing Statistical Power Without Increasing Sample Size

Gary H. McClelland
University of Colorado at Boulder

The American Psychological Association Task Force on Statistical Inference (TFSI; Wilkinson & the TFSI, August 1999) admirably emphasized the importance of statistical power for the design and reporting of research. However, by discussing power primarily in the section entitled *Power and Sample Size* (p. 596), the TFSI may have created an unintended false impression that increasing sample size is the best or perhaps only remedy for improving statistical power. Psychologists must consider design strategies other than augmenting sample size for increasing statistical power to reduce cost and minimize the burden on human or animal subjects.

The TFSI's recommendation that "confidence intervals replace calculated power in describing results" (Wilkinson & the TFSI, 1999, p. 596) provides the key to understanding alternative and equally effective strategies researchers can use to increase statistical power. The formula for confidence intervals identifies the ingredients that influence confidence interval width and that hence determine statistical power. If X is a quantitative independent variable or a numerical coding of a categorical independent variable, then the $100(1-\alpha)\%$ confidence interval for the slope (or the coded difference among the means) relating it to the dependent variable in a linear model with m independent variables is given by the following equation:

$$b_X \pm t_{n-m-1;\alpha}\sqrt{\frac{MSE}{nV_X(1-R^2_X)}} \qquad (1)$$

The square root term in Equation 1 represents the standard error of the slope and contains the primary ingredients of the confidence interval: the mean square error (*MSE*), the number of observations (n), the variance of X (V_X), and the proportion of the variation in X not shared with other variables in the model ($1-R^2_X$). Decreasing any term in the numerator or increasing any term in the denominator narrows the confidence interval and increases statistical power. Proportional changes in any of the terms are equipotent. For example, doubling the variance of X or doubling the proportion of unique variation in X (i.e., reducing its redundancy with other predictor variables) narrows the confidence interval by exactly the same amount as would doubling sample size. Hence, each term in the confidence interval equation suggests alternative strategies for increasing statistical power; each is considered briefly in turn, beginning with the ingredients of the standard error.

First, researchers can decrease the mean square error by using more reliable and accurate measures, by implementing better experimental controls, and by collecting data of higher quality. It is important to note that they can also reduce the mean square error by including concomitant variables in more sophisticated statistical models of their data (Judd & McClelland, 1989; Maxwell & Delaney, 1990). Many good computer programs recommend sample sizes for, say, two-sample t tests, but these programs fail to recommend considering analyses of covariance or within-subject designs as alternatives to increasing sample size for improving inadequate statistical power.

Second, changing the range of X can increase the variance of X substantially. Even for a fixed range, changing the distribution of X by unequally allocating observations to levels in an experiment or by oversampling extreme cases in field studies can substantially increase the variance of X (McClelland, 1997; McClelland & Judd, 1993).

Third, striving for nonredundant predictor variables or using orthogonal contrasts greatly reduces the need for larger sample sizes to achieve adequate statistical power. The reduced redundancy also allows a less ambiguous interpretation of a variable's effect.

Fourth, researchers can change the critical value ($t_{n-m-1,\alpha}$) of Student's t distribution by increasing sample size or by increasing the value of alpha. However, the critical values from Student's t distribution do not appreciably decrease as n changes for most sample sizes (e.g., $t_{20;.05} = 2.09$ and $t_{40;.05} = 2.02$), and even increasing alpha from .05 to .1 does not help much ($t_{120;.05} = 1.98$ and $t_{120;.10} = 1.66$). Thus, although many textbooks discuss increasing the risk of a Type I error as a means of increasing power, other strategies for increasing power are likely to be more effective. More important, when using Student's t one implicitly assumes that the errors are independent and normally distributed with a common variance. Although violations of the normality and common variance assumptions sometimes do not appreciably alter Type I error probabilities, they can substantially reduce statistical power (Wilcox, 1996). Ensuring that the assumptions are satisfied (either by using certain procedures when collecting the data or by performing certain transformations after the data are collected) or using alternative statistical methods (see Wilcox, 1996) improves statistical power.

Finally, although a large coefficient b_X does not affect the width of the confidence interval, it does, all else being equal, make it less likely that the confidence interval will include zero. Hence, one of the best strategies for increasing statistical power is to use theory and prior research to identify variables that are likely to have potent effects. Doubling one's thinking is likely to be much more productive than doubling one's sample size.

In summary, researchers have available a number of strategies other than increasing sample size for improving statistical power. Adopting those strategies not only increases statistical power but also produces sounder research and better data analyses. Before incurring the expense of increasing sample sizes, researchers should consider their powerful alternatives.

REFERENCES

Judd, C. M., & McClelland, G. H. (1989). *Data analysis: A model comparison approach*. San Diego, CA: Harcourt Brace Jovanovich.

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data: A model comparison perspective*. Belmont, CA: Wadsworth.

McClelland, G. H. (1997). Optimal design in psychological research. *Psychological Methods, 2,* 3–19.

McClelland, G .H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin, 114,* 376–390.

Wilcox, R. R. (1996). *Statistics for the social sciences*. San Diego, CA: Academic Press.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Correspondence concerning this comment should be addressed to Gary H. McClelland, Department of Psychology, CB 345, University of Colorado, Boulder, CO 80309-0345. Electronic mail may be sent to gary.mcclelland@colorado.edu.

DOI: 10.1037//0003-066X.55.8.964

Significance Testing and the Science

David Sohn
University of North Carolina at Charlotte

From the report of the Task Force on Statistical Inference (a committee of the American Psychological Association Board of Scientific Affairs; Wilkinson & the Task Force on Statistical Inference, August 1999), one would never guess that the impetus for the report was controversy over a particular way of drawing statistical inferences, namely, the statistical significance test. The test is barely mentioned. What discussion there is of it is found in a section of fewer than 100 words that does not even use the term *significance test*. Further, what is suggested in the single sentence that actually deals with the test is debatable, to be charitable. It is suggested that the accept–reject decision of the test be dropped and in its place p values be reported. But where are reports of p values greater than .05 going to be published, given the editorial bias against nonsignificant findings? To make matters absolutely maddening, the report, in another place, strongly urges the use of power analysis, which is premised on there being an accept–reject approach with alpha levels.

Thus, if readers are looking for enlightenment on the matter of statistical inference, they will not find it in a report that takes up every statistical question other than the one that is its reason for being. Readers are regaled with advice to "never use the unfortunate expression 'accept the null hypothesis'" (Wilkinson & the Task Force on Statistical Inference, 1999, p. 599) and to avoid empty statements like "further research needs to be done" (p. 602). What the report amounts to is a vote of confidence for business as usual in the conduct of the science of psychology. Psychologists may go on doing as they have been, if they throw in an effect size or a confidence interval when discussing the statistical analysis. Fifty years of collective anguish over the test and the test's value to the science, the report implies, are just an aberration that one need not lose sleep over.

But does it really mean nothing that, in the context of considering the value of the significance test to the science, a leading scientist is brought to the following view of the state of the science? "I have developed a certain angst over the intervening 30-something years—a constant, nagging feeling that our field spends a lot of time spinning wheels without really making much progress" (Loftus, 1996, p. 161). Does it mean nothing that another laments, "There is a chronic pessimistic feeling in the social and behavioral sciences that . . . our progress has been exceedingly slow, if indeed there has been any progress at all" (Rosenthal, 1991, p. 3)? Is it insignificant that another proposes that because scientific research has been so ineffective, the critical responsibility for deciding what research means should be taken out of the hands of empirical researchers and put into those of literature reviewers, specifically, meta-analysts (Schmidt, 1992)? What does this next exclamation suggest about its author's perception of the success of the science? "Go build a quantitative science with p values!" (Cohen, 1994, p. 1001). Because the author does not say where in psychology a quantitative science is being built, readers can only assume that it is not being built anywhere.

Concern for data analysis may be the occasion for such expressions, but this does not make the concern for the science itself less real. The fact remains that for whatever reasons, leading scientists are profoundly disappointed with the progress of the science. Is it not time for those who practice psychological science to look beyond data analysis in their attempt to discover what is responsible for the disappointing results of their efforts?

REFERENCES

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003.

Key Points

Chapter 22: More about Tests and Intervals

- **Type I error:** the null hypothesis is true, but we mistakenly reject it (false positive)
- **Type II error**: The null hypothesis is false, but we fail to reject it (false negative)
- Alpha: how small the P-value should be, P(Type I error)
- Beta: the probability of Type II error
- Power = 1 – beta
- Winner's curse: increased bias in low powered studies
- Effect size: the distance between the null hypothesis value and the truth, but similar to signal-to-noise ratio