

Chapter 3

What's Wrong With Statistical Tests— And Where We Go From Here

Reality is that which, when you stop believing in it, doesn't go away.

—Philip K. Dickhey

Overview

- Misinterpretations of outcomes of statistical tests
- Limitations of traditional statistical tests for the behavioral sciences
- Is anything right with statistical tests?
- Where we go from here

Misinterpretations of statistical tests

- Correct interpretation of p values from statistical tests is actually quite narrow
- The term p is the conditional probability of the data (D) assuming H_0 is true:

$$p(D | H_0)$$

- Reviewed next are some false beliefs about p values described by Bakan (1966), Carver (1978), Cohen (1994), Nickerson (2000), Oakes (1986), Rozeboom (1960), Schmidt and Hunter (1997), and Shaver (1993), among others

Misinterpretations of statistical tests

- Common misinterpretations for the case $p < .05$:

Fallacy: There is less than a 5% likelihood that the results are due to chance (sampling error)

Reality: This is the *odds-against-chance fantasy*—because p is derived assuming H_0 is true (i.e., all results are due to chance), it cannot be seen as a measure of the odds of chance

Misinterpretations of statistical tests

- Common misinterpretations for the case $p < .05$:

Fallacy: The probability that H_0 is true is less than 5%

Reality: This is the *inverse probability error*— p measures a characteristic of the data, not H_0

That is, the form of p is not $p(H_0 | D)$

Misinterpretations of statistical tests

- Common misinterpretations for the case $p < .05$:

Fallacy: There is less than a 5% chance that the decision taken to reject H_0 is a Type I error

Reality: This confuses p with the conditional *posterior* probability of a Type I error (i.e., based on the results):

$$p(H_0 \mid \text{reject } H_0)$$

The term p is also not identical to $\alpha = p(\text{reject } H_0 \mid H_0)$, the conditional *prior* probability of a Type I error

One of the most common misconceptions (see Table 3.1)

Misinterpretations of statistical tests

- Common misinterpretations for the case $p < .05$:

Fallacy: The likelihood that H_1 is true is greater than 95%

Reality: This is the *valid research hypothesis fantasy*—the complement of p , $1 - p$, also measures a characteristic of the data, not that of any hypothesis

That is, the form of $1 - p$ is not $p(H_1 | D)$

Misinterpretations of statistical tests

- Common misinterpretations for the case $p < .05$:

Fallacy: The chance of replication exceeds 95%

Reality: This is the *replicability fantasy*—the probability of replication is *not* directly estimated by the complement of p

Replication is a matter of experimental design, sampling, and the magnitude of a true effect

Misinterpretations of statistical tests

- Common misinterpretations for the case $p < .05$:

Fallacy: Low p values indicate large effects

Reality: Statistical tests measure effect size and sample size together

Misinterpretations of statistical tests

- Common misinterpretations for the case $p < .05$:

Fallacy: Rejection of H_0 confirms both H_1 and the substantive hypothesis behind it

Reality: This is the *meaningfulness fallacy*— H_0 and H_1 are merely statistical hypotheses that may be just as consistent with other substantive hypotheses

Rejection of H_0 could be a Type I error

Misinterpretations of statistical tests

- Common misinterpretations for the case $p < .05$:

Fallacy: Rejecting H_0 confirms the quality of the research design

Reality: Failing to reject H_0 can be the product of good science (e.g., a bogus claim is not verified); also, failure to reject H_0 supports the researcher's theory in the accept-support context

Related cognitive distortion—The “law of small numbers” (Tversky & Kahneman, 1971)—the belief that

1. even small samples are typically representative
2. statistically significant results are likely to be found in replication samples half the size of the original

Limitations of traditional statistical tests

- There is ample evidence that misinterpretations are quite widespread (e.g., Mittag & Thompson, 2000; Nelson, Rosenthal, & Rosnow, 1986; Oakes, 1986; see also Table 3.1)
- Nil hypotheses (i.e., H_0 states that the population effect is zero) are usually false
- The probability of data under an implausible H_0 is expected to be low

Limitations of traditional statistical tests

- Lunneborg (2001): There is a mismatch between the *population inference model* on which statistical tests are based (e.g., random samples assumed) and the *randomization model* of many experimental studies (e.g., cases from ad hoc samples are randomly assigned to conditions)
- Distributional assumptions are infrequently verified (e.g., Keselman et al., 1998)
- There may also be few instances in practice when standard statistical tests give accurate results due to violations of assumptions (e.g., Lix, L. Keselman, & H. Keselman, 1996)

Limitations of traditional statistical tests

- Statistical tests bias the research literature:
 - ✓ Publication bias against studies without H_0 rejections
 - ✓ May result in overestimation of population effect size (see Table 3.2)
 - ✓ Leads to the “file drawer problem” (can be estimated in meta-analysis)
- Statistical tests make the research literature difficult to interpret

Example: When power is .50 (typical), only half the studies will show positive results (e.g., Table 3.2)

Limitations of traditional statistical tests

- Use of statistical tests discourages replication?
 - ✓ Replication seems to be undervalued in the behavioral sciences compared with the natural sciences
 - ✓ Statistical tests are rarely used in the natural sciences
 - ✓ Correlation or causation (e.g., due to replicability fantasy)?

Limitations of traditional statistical tests

- Statistical tests encourage dichotomous thinking, especially about p values (e.g., Rosnow & Rosenthal, 1989)
- Diverts attention away from the data and the measurement process
- Diverts resources away from learning about other methods of data analysis or inference (e.g., Bayesian statistics, methods for establishing replication)

Limitations of traditional statistical tests

- There has been little change in undergraduate and graduate education in psychology statistics over the last 20 years (e.g., Frederich, Buday, & Kerr, 2000)
- Facilitates research about “fad” topics that clutter the literature, but have little scientific value (Meehl, 1990)
- The researcher can change the rules of the game (e.g., increasing α from .01 to .05 in order to reject H_0)
- These problems may limit the impact of behavioral science research (e.g., Beutler, Williams, Wakefield, & Entwistle, 1995; Miller, 1999)

Is there is anything right with statistical tests?

- There are some variations on standard statistical tests that may be useful in certain specialized situations
- One example is *equivalence testing*, which deals with hypotheses about the equivalence between two conditions
- Equivalence testing is based in part on the idea of *good-enough belts*, a range of results specified by the researcher that establishes minimum results necessary for further analysis or indicate no appreciable difference

Is there is anything right with statistical tests?

- The specification of range null hypotheses is rare in the behavioral sciences, but it is better known in other areas, such as the environmental sciences (e.g., McBride, 1999)
- Tryon (2001) described a method based on *inferential confidence intervals* for the testing of hypotheses about statistical difference or equivalence; a result that is neither is indeterminant
- These approaches, however interesting, are probably not a general solution for many behavioral researchers

Where we go from here

- There are basically three options:
 1. Do nothing; that is, continue using statistical tests just as we have for the last 50 years
 2. Ban the use of statistical tests: This option is not as radical as it may first appear because it has been seriously discussed—it is merely impractical (e.g., Abelson, 1997)
 3. Change the way we use statistical tests—including not using them at all
- The third option is already happening, but we should plan these changes more systematically

Where we go from here

- Recommendations:

1. Only in very exploratory research may a primary role for statistical tests be appropriate (i.e., a marker of a more advanced research area is that it does not depend on statistical tests)
2. Report estimated a priori power for *any* use of statistical tests, and specify only plausible null hypotheses
3. It is *not* acceptable anymore to describe results solely in terms of outcomes of statistical tests

Where we go from here

- Recommendations:

4. Drop the word “significant” from our data analysis vocabulary (i.e., use it only as everyone else does—to designate something as meaningful)
5. Whenever possible, researchers should report and interpret effect size estimates and confidence intervals for primary results
6. The researcher must demonstrate the *substantive* significance of the results—statistical tests are inadequate for this purpose

Where we go from here

- Recommendations:

7. Replication is the best way to deal with sampling error
8. Reform education in statistics—specifically, de-emphasize statistical tests and instead show students how to replicate results and evaluate their substantive significance
9. Make statistical software less NHST-centric—for example, provide better support for effect size estimation and interval estimation

Where we go from here

- Aim of these recommendations: To make the behavioral sciences more like the natural sciences in that we will
 1. report the directions and magnitudes of our effects
 2. determine whether they replicate, and
 3. evaluate them for substantive, not just statistical, significance (e.g., Kirk, 1996)

References

- Abelson, R. P. (1997). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented). In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 117-141). Mahwah, NJ: Erlbaum.
- Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*, 423-437
- Beutler, L. E., Williams, R. E., Wakefield, P. J., & Entwistle, S. R. (1995). Bridging scientist and practitioner perspectives in clinical psychology. *American Psychologist*, *50*, 984-994.
- Carver, R. P. (1978). The case against significance testing. *Harvard Educational Review*, *48*, 378-399.
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997-1003.
- Frederich, J., Buday, E., & Kerr, D. (2000). Statistical training in psychology: A national survey and commentary on undergraduate programs. *Teaching of Psychology*, *27*, 248-257
- Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., et al. (1998). Statistical practices of education researchers: An analysis of the ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, *68*, 350-386.
- Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*, 746-759.
- Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumptions violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. *Review of Educational Research*, *66*, 579-620.
- Lunneborg, C. E. (2001). Random assignment of available cases: Bootstrap standard errors and confidence intervals. *Psychological Methods*, *6*, 402-412.
- McBride, G. B. (1999). Equivalence testing can enhance environmental science and management. *Australian and New Zealand Journal of Statistics*, *41*, 19-29.

- Meehl, P. E. (1990). Why summaries on research on psychological theories are often uninterpretable. *Psychological Reports*, 66 (Monograph Suppl. 1-V66), 195-244.
- Miller, D. W. (1999, August 6). The black hole of education research: Why do academic studies play such a minimal role in efforts to improve the schools? *Chronicle of Higher Education*, 45(48), A17-A18.
- Mittag, K. C., & Thompson, B. (2000). A national survey of AERA members' perceptions of statistical significance tests and other statistical issues. *Educational Researcher*, 29, 14-20.
- Nelson, N., Rosenthal, R., & Rosnow, R. L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist*, 41, 1299-1301.
- Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, 5, 241-301.
- Oakes, M. (1986). *Statistical inference*. New York: Wiley.
- Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, 44, 1276-1284.
- Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin*, 57, 416-428.
- Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 37-64). Mahwah, NJ: Erlbaum.
- Shaver, J. P. (1993). What significance testing is, and what it is not. *Journal of Experimental Education*, 61, 293-316.
- Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods*, 6, 371-386.
- Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, 76, 105-110.