20120424 - Misinterpretation of 'p' (1986) (2e)

[Data] [<Normal page] [**PEREZGONZALEZ Jose D [ed] (2012).** *Misinterpretation of 'p' (1986) (2e)[5]*. Journal of Knowledge Advancement & Integration (ISSN 1177-4576), 2012, pages 140-143.]

Misinterpretations of 'p' and 'sig'

Oakes (1986[3]) carried out a study on common misinterpretations of the logic of tests of significance among British psychology academics. Typically, most of these misinterpretations confuse p-values (ie, the probability of the data when assuming that the null hypothesis is true) and, especially, statistical significance[8], with the probability of proving or disproving hypotheses (be this the null hypothesis or an alternative hypothesis).
Another misinterpretation is the so-called "replication fallacy", which occurs when the p-value is assumed to represent the probability of finding similar results if the research were to be repeated.

Oakes found that 96% of the participants held at least one misinterpretation (mean = 2.5) out of the six given to them (see table 1).

Table 1. Frequencies and percentages of misinterpretations regarding tests of significance		
Common misinterpretations[6]	**f**	**%**
Significance disproves the null hypothesis	1	1.4%
The p-value informs of the probability of the null hypothesis	25	35.7%
Significance proves the alternative hypothesis	4	5.7%
The p-value informs of the probability of the alternative hypothesis	46	65.7%
'P' informs of the probability of a wrong decision when rejecting the null	60	85.7%
The p-value informs of the probability of the results if replicated	42	60.0%
(Participants who answered that all of above were false)	3	4.3%

Table 1 reflects potential misinterpretations, potential because they were a product of conscious effort (ie, reflection) introduced by the impromptu research. However, Oakes also asked the participants to point out which of those or alternative interpretations they typically held prior to the research. He found that most participants typically interpreted p-values and the level of significance somehow differently than table 1 shows (ie, they had preferred interpretations rather than exhausting all possible known or "logical" interpretations) (see table 2).

Table 2. Frequencies and percentages of typical interpretations					
Interpretations[6]	**typical**			**upon reflection**	
	f	**%**		**f**	**%**
Significance disproves the null hypothesis	1	1.4%		0	0.0%

The p-value informs of the probability of the null hypothesis	32	45.7%		-7	-10.7%
Significance proves the alternative hypothesis	2	2.9%		+2	+2.8%
The p-value informs of the probability of the alternative hypothesis	30	42.9%		+16	+22.8%
'P' informs of the probability of a wrong decision if rejecting the null	48	68.6%		+12	+17.1%
The p-value informs of the probability of the results if replicated	24	34.3%		+18	+25.7%
The p-value is the probability of the data given the null hypothesis is true[*]	8	11.3%		n/a	n/a

(* Correct interpretation freely provided by some participants)

Most interesting of all, though, is that the ad hoc research prompted them to reflect on other meanings beyond their preferred interpretations. In table 2, the columns under "upon reflection" show changes in interpretation which were prompted by the research (ie, the difference between typical interpretations held before doing the research and interpretations considered 'true', upon reflection, while doing the research). These changes were the following:

- Given the opportunity to freely include any other interpretations, 11.3% of participants claimed to typically use the (correct) interpretation of the p-value as the probability of the data given the null hypothesis is true. Yet, only 3% (2 participants) held it as the only correct interpretation throughout the research while 1.4% (1 participant) came to hold this interpretation as the only correct one upon reflection (these were the 3 participants who thought all research interpretations were false).
- A positive change was that 10% of participants who typically interpreted the p-value as the probability of the null hypothesis being true, thought, upon reflection, that they were wrong in their interpretation.
- Most other changes were negative[7]:
  - 2.8% of the participants who typically did not interpret significance as (absolutely) proving the alternative hypothesis thought, upon reflection, that it could, indeed, support such interpretation.
  - 22.8% of the participants who typically did not interpret the p-value as the probability of the alternative hypothesis to be true thought, upon reflection, that it could support such interpretation.
  - 17.1% of the participants who typically did not interpret the p-value as the probability of making a wrong decision when rejecting the null hypothesis[9] thought, upon reflection, that it could support such interpretation.
  - 25.7% of the participants who typically did not interpret the p-value as the probability of replicating results thought, upon reflection, that it could support such interpretation.

Methods

Research approach

Not much detail. It appears to have been an ad hoc, exploratory study.

Sample

A convenient sample of 70 participants from, probably, a psychology department at a British university. The participants were university lecturers, research fellows and postgraduate students with two or more years' experience doing research.

Materials

A questionnaire consisting of a small scenario and six statements. The scenario described a research with two-independent samples, and provided the relevant results: a t-test with 'p=0.01'.

- The six statements asked for a true / false decision regarding whether each particular statement reflected a logical interpretation of the results. Unbeknownst to the participants, all statements were false, representing six common misinterpretations regarding tests of significance.
- Either the researcher or the questionnaire itself also provided the 'hint' that *"no particular ratio of true to false statements should be anticipated"*.

Following the questionnaire, participants were asked to pinpoint those interpretations they typically held before the research, including any other interpretations not provided during the research. At the end of the research, participants were debriefed about misinterpretations and the correct interpretation of p-values and the level of significance.

Analysis

Descriptive statistics.

Generalization potential

This particular research appears to be severely limited to an unknown population of psychology academics (probably Sussex University, in the UK). Yet the results might, at least, serve as a working hypothesis for generalizing to other populations such as the following (in order of decreasing generalizability):

- British psychology academics and researchers (including students).
- British academics and researchers (including students) who also use null hypothesis significance testing (NHST) or Fisher's tests of significance (such as academics from other social sciences, medicine, biology, etc).
- Professionals trained in British universities (especially psychologists, social scientists, etc).
- (See also partial replication studies by Haller and Krauss, 2000[2], in Germany, and by Falk and Greenbaum, 1995[1], in Israel, for potential generalizability beyond the UK).

References

1. **FALK Ruma & Charles W GREENBAUM (1995).** *Significance tests die hard: the amazing persistence of a probabilistic misconception.* Theory & Psychology, 1995, volume 5, number 1, pages 75-98. DOI 10.1177/0959354395051004.
2. **HALLER Heiko & Stefan KRAUSS (2000).** *Misinterpretations of significance. A problem students share with their teachers.* Methods of Psychological Research Online, 2002, volume 7, number 1, pages 1-20.
3. **OAKES Michael (1986).** *Statistical inference: a commentary for the social and behavioral sciences.* John Wiley & Sons (Chichester, UK), 1986.
4. **PEREZGONZALEZ Jose D [ed] (2011).** *Misinterpretation of 'p' (1986).* Journal of Knowledge Advancement & Integration (ISSN 1177-4576), 2011, pages 99-102.

+++ **Notes** +++

5. This second edition updates the original edition[4] by reducing confusion between p-values and statistical significance (see tests of significance).
6. The original research statements have been rephrased here.
7. Notwithstanding this, all participants were debriefed about the correct and incorrect interpretations after the research.
8. The example provided by Oakes to his participants used p=0.01, thus it can be interpreted as having the dual role of a 'p-value' and a 'conventional level of significance'.
9. They could have interpreted it as the probability of making a type I error, though. As Oakes puts it, *"Statement (v) may seem like a textbook definition of the type 1 error, but on close inspection it can be seen that it is identical to statement (ii). It is therefore a statement of inverse probability and should be described 'false'."*

Want to know more?

Wiki of Science - Hypotheses testing (disambiguation)
　　　This Wiki of Science page lists alternative methods for testing the probability of data or hypotheses.
Wiki of Science - Null hypothesis significance testing
　　　This Wiki of Science page reflects on the pseudoscientific bases of the null hypothesis significance testing (NHST) procedure typically used in the social sciences and medicine.
Wiki of Science - Studies which replicate Oakes's
　　　You can find more information on two studies that partially replicated Oakes's one, in Wiki of Science. One study was done by Falk and Greenbaum in 1995; the other study was done by Haller and Krauss in 2000.

Editor

Jose D PEREZGONZALEZ (2012). Massey University, Turitea Campus, Private Bag 11-222, Palmerston North 4442, New Zealand. (JDPerezgonzalez JDPerezgonzalez).

Like　　Tweet 0　　+1 0

Other interesting sites

Journal KAI	Wiki of Science	AviationKnowledge	A4art	The Balanced Nutrition Index

page revision: 2, last edited: 24 Apr 2012, 17:45 GMT+12 (15 seconds ago)