Soc. Sci. **2015**, *4*, 361–372; doi:10.3390/socsci4020361

OPEN ACCESS

social sciences

ISSN 2076-0760

www.mdpi.com/journal/socsci

Article

Persistent Confusions about Hypothesis Testing in the Social Sciences

Christopher Thron * and Vincent Miller

Department of Mathematics, Texas A&M University-Central Texas, 1001 Leadership Place, Killeen, TX 76549, USA; E-Mail: victormike71@gmail.com

* Author to whom correspondence should be addressed; E-Mail: thron@tamuct.edu; Tel.: +1-254-519-5776; Fax: +1-254-519-5482.

Academic Editor: Martin J. Bull

Received: 24 December 2014 / Accepted: 27 April 2015 / Published: 12 May 2015

Abstract: This paper analyzes common confusions involving basic concepts in statistical hypothesis testing. One-third of the social science statistics textbooks examined in the study contained false statements about significance level and/or *p*-value. We infer that a large proportion of social scientists are being miseducated about these concepts. We analyze the causes of these persistent misunderstandings, and conclude that the conventional terminology is prone to abuse because it does not clearly represent the conditional nature of probabilities and events involved. We argue that modifications in terminology, as well as the explicit introduction of conditional probability concepts and notation into the statistics curriculum in the social sciences, are necessary to prevent the persistence of these errors.

Keywords: statistics; hypothesis testing; inference; *p*-value; significance; conditional probability; conditional event

---

1. Introduction

Although social scientists use statistical hypothesis testing extensively, it is not clear whether they always understand what their test results mean. In 1986 Michael Oakes conducted a survey of academic psychologists concerning their understanding of hypothesis tests, and found that on some of his questions over 70 percent of respondents gave wrong answers [1]. Oakes' survey was almost 30 years ago, but we have found these confusions persist until today. Indeed, in our research we found

several modern statistics textbooks written for social scientists that propagate fundamental interpretative errors. Apparently, social scientists' persistent misunderstandings are due at least in part to the fact that many are being mistaught.

Why have these mistakes not been corrected? Our investigation shows that the conventional terminology is the source of the confusion, because it does not clearly and unambiguously represent the conditional nature of the probabilistic concepts and quantities involved in hypothesis testing. It is incumbent upon the social science community to correct this situation. In any rigorous scientific field, widespread misunderstandings that are taught by textbooks themselves should not be tolerated. Our hope is that this paper will serve to motivate constructive changes that will bring increased clarity and rigor to the practice of statistics within the social sciences.

The remainder of the paper is organized as follows. Section 2 provides brief definitions and explanations of some key concepts in hypothesis testing. Section 3 gives examples of wrong interpretations drawn from social science statistics textbooks. Section 4 analyzes the causes of these misinterpretations. Section 5 gives our conclusions, including our recommendations to remedy to the situation. Included at the end of the paper are a mathematical appendix in which various conditional probability calculations are demonstrated; and a complete list of the textbooks examined in the study.

2. Hypothesis Testing Concepts and Terminology

There are various forms of hypothesis testing, but all require the formulation of a null hypothesis (*H0*) and an alternative hypothesis (*H1*). For example, in evaluating the effectiveness of a certain abstinence education program in reducing venereal disease among high-school students, the null and alternative hypotheses might be expressed as follows:

> *H0*: Students that have been exposed to the program experience the same venereal disease rates as similar students that haven't been exposed;
>
> *H1*: Students that have been exposed to the program experience lower rates of venereal disease compared to similar students that haven't been exposed.

To test a hypothesis, a *statistical experiment* must be conducted. McClave and Sincich [2] define such an experiment as "an act or process of observation that leads to a single outcome that cannot be predicted with certainty". In an experiment that is designed to test a particular *H0*, the probabilities of different possible outcomes must depend on whether or not *H0* is true.

Another required feature of hypothesis testing is the *significance level*, which is usually represented by the Greek letter α. The significance level of a statistical test is defined as the probability that, in a situation where *H0* is true, the experimenter nonetheless rejects *H0* as a result of the test. To clarify this statement, we re-express it in the notation of *conditional probability*. This is not typically done in statistics textbooks for social scientists, but we will see that this will enable us to avoid the confusion that is often introduced (either explicitly or implicitly) in such textbooks. Using conditional probability notation, we may write

$$\alpha := \Pr[\, H0 \text{ is rejected in an experiment} \mid H0 \text{ is true for that experiment}] \quad (1)$$

which may be read as: "Alpha is defined as the conditional probability that the null hypothesis is rejected in an experiment, given that the null hypothesis is actually true for that experiment."

In practical hypothesis tests (such as the ubiquitous *z*-test), the value of α is used to set a certain threshold value (which is calculated using probability theory). When the experiment is performed, a statistic is computed based on the results of the experiment. If the computed statistic is greater than the threshold, then *H0* is rejected at significance level α. We shall not describe in detail the computations involved in this procedure, because these specifics are not germane to the issues we wish to consider. A concept that is closely related to the significance level is the notion of *Type I error*, for which textbooks give a variety of definitions. Bakeman ([3], p. 27) says, "A type I error…means rejecting the null hypothesis when it is in fact true". Daly and Bourke ([4], p. 130) says, "If…the null hypothesis is true and a significant result is obtained, …, this form of error is called…a type I error". Hopkins and Glass ([5], p. 221) says, "If we reject *H0*, and it is true, then we have made a type-I error". Welkowitz ([6], p. 126) says, "Rejecting the null hypothesis for an experiment in which it is actually true is called a Type I error". Popular internet resources give similar definitions: Wikipedia [7] states: "type I error is the incorrect rejection of a true null hypothesis"; and alternatively, "A type I error…occurs when the null hypothesis (*H0*) is true, but is rejected".

These definitions appear to be more or less equivalent. However, a possible difference emerges when we speak of the *probability* of a Type I error. If we go with Hopkins and Glass's definition, then it seems we should say

$$\text{Probability of Type I error} = \Pr[\ H0 \text{ is rejected in an experiment and } H0 \text{ is true for that experiment}] \quad (2)$$

On the other hand, it is possible to interpret Bakeman's definition as implying (in conditional probability notation)

$$\text{Probability of Type I error} = \Pr[\ H0 \text{ is rejected in an experiment} \mid H0 \text{ is true for that experiment}] \quad (3)$$

which is equal to α as defined above. Indeed, Bakeman's definition is ambiguous, because his "when" could mean either "at the same time" or "given that". In the former case, then Type I error denotes a possible *final outcome* of the completed experiment; while in the latter case, then Type I error is a *conditional event*. Ordinary English usage would seem to favor the former, simpler interpretation (for example, "seeing a gorilla when at the zoo" would not ordinarily be considered a conditional event)—but statistical authorities (including two reviewers of this paper) indicate that the latter interpretation is correct. In textbooks, generally no mention is made of the possible ambiguity. Many authors refer to α as the "probability of Type I error", which presumes the conditional-event definition; but we shall see that some of these same authors also identify Type I error as a possible experimental outcome, leading to the false conclusion that α is the probability that *H0* is wrongly rejected in a given experiment. Expressed in conditional probability notation, the difference between the two probabilities is clear—$\Pr[A|B]$ is not the same as $\Pr[A \text{ and } B]$—but unfortunately, the ambiguous conventional terminology leads to confusion between the two.

Many statistics texts illustrate the concepts of Type I and Type II error using the following table (Table 1):

Table 1. Hypothesis Testing and Types of Error.

	H0 is Actually True	**H0 is Actually False**
Experimental conclusion: reject H0	Type I error is committed	Experimental conclusion is correct
Experimental conclusion: do not reject H0	Experimental conclusion is correct	Type II error is committed

This table fails to resolve the ambiguity in the definition of Type I error. Under the final-outcome interpretation, the table lists the four possible outcomes of an experiment; while under the conditional-event interpretation, the column headings are preconditions, and the row headings list possible outcomes for each precondition. Under the first interpretation, the probabilities of the four table entries should sum to one; but under the second, the probabilities in each *column* sum to one. Frequently the probabilities of the upper left and lower right corners are identified as α and β, respectively—without clarifying that this identification requires that column headings.be interpreted as preconditions.

Another concept that is closely related to α is the *p*-value. The *p*-value is commonly defined as the probability of obtaining a result at least as "extreme" as the result observed, given the precondition that the null hypothesis is in fact true for the experiment that was conducted. This definition relies on the vague and problematic notion of "extreme"; it also leaves open the question of what *p*-value means if the null hypothesis is not true for the conducted experiment. The following more precise definition avoids these problems. Suppose an experiment has been performed, and based on the results the value of the statistic is calculated and found to be *V*. Then the *p*-value for the experiment may be defined as the probability that *H0* will be rejected in a not-yet-performed experiment (that is, an experiment for which the outcome is as yet unknown) in which both *H0* is true and the threshold value used to determine rejection is *V*. In conditional probability language, we have:

$$\begin{aligned} p\text{-value: } = \Pr[\ & H0 \text{ will be rejected in a not-yet-performed experiment} \mid \text{in this experiment} \\ & \text{both } H0 \text{ is true and the threshold value for the statistic is set equal to the value of the} \\ & \text{statistic that was obtained in the completed experiment}] \end{aligned} \tag{4}$$

We should clarify that this not-yet-performed experiment could be a repetition of the completed experiment from which the *p*-value was calculated. Note however that *H0* must be true in the repeated experiment for the *p*-value to represent the probability of a Type I error. But in practice we don't know whether *H0* is true for the completed experiment—otherwise we wouldn't be doing the experiment in the first place.

Practically, the *p*-value serves as an indicator of significance. When an experiment is performed the *p*-value computed from the experiment is compared with α to determine whether or not to reject *H0.* The *p*-value is the *minimum* significance level α which would lead to the rejection of *H0* based on the experimental result (this is logically equivalent to the statement that α determines the *maximum* *p*-value which leads to rejection of *H0*). For example, suppose an experimenter first sets $\alpha = 0.05$. He then conducts the experiment, and calculates $p = 0.011$ based on the results. In this case he rejects *H0*, since $\alpha > p$; that is, α exceeds the minimum significance level that indicates rejection of *H0*. On the other hand, if another experimenter first sets $\alpha = 0.01$, then repeats the same experiment and happens to obtain the same result of $p = 0.011$, then she does not reject *H0* since $\alpha < p$. The two experimenters

draw two different conclusions because the experimenter that set the smaller significance level is requiring more convincing evidence before she is willing to reject *H0* (This example points out a somewhat misleading feature of the terminology: if two experimental tests end up rejecting *H0*, it is the test with *larger* significance level that provides *less* significant evidence that *H0* is untrue).

It is quite possible to simply define *p*-value in terms of its function as significance indicator, as we have described in the preceding paragraph. This interpretation is so much simpler than the conditional-probability definition that one might wonder why the former is used at all. Indeed, the conditional-probability definition is all too commonly misunderstood and misapplied, as we shall see in the following discussion.

3. Contemporary Textbooks' Misinterpretations of Hypothesis Testing Concepts

In this section we present some erroneous interpretations of confidence level (α) and *p*-value that we found in contemporary statistics textbooks aimed for social scientists. Most of these texts were taken from the main library at the University of Texas at Austin, while others were texts used by acquaintances in their statistics classes in various branches of social science.

3.1. Erroneous Interpretations of α

The text by Healey [8] gives a practical example of a hypothesis test that compares absenteeism between treated alcoholics and the rest of the community. An alpha level of 0.05 is used, and experimental results corresponded to a *p*-value smaller than 0.05. Healey concludes ([8], p. 191):

> In the example at hand, the null hypothesis was rejected and the probability that this decision was incorrect is 0.05.

At first glance, this statement may appear reasonable. But expressed as a statement of conditional probabilities, it becomes:

$$\Pr[H0 \text{ is wrongly rejected in a given experiment} \mid H0 \text{ is rejected in the experiment}] = \alpha \quad (5)$$

This is entirely different from our original specification of α:

$$\alpha := \Pr[\text{ H0 is rejected in a given experiment} \mid H0 \text{ is true for the experiment}] \quad (6)$$

By neglecting the condition that is implicit in the definition of α, Healy has been led into making a false statement.

Bachman & Paternoster [9] make essentially the same mistake. In their example, they perform a one-tailed *z*-test and use an alpha level of 0.01, which corresponds to a threshold value for the *z*-statistic of 2.33. The *z*-statistic computed based on the experiment is +4.34. They conclude ([9], p. 286):

> As $z_{obt} > 2.33$, you would reject the null hypothesis, knowing there is a 1 in 100 chance that you are making the wrong the decision (Type I error).

Here is one instance where the authors are apparently treating Type I error as an experimental outcome rather than a conditional event, since they equate it with making a wrong rejection of *H0* in a case where *H0* is not known to be true. If they had intended for Type I error to represent a conditional event, they should have said something like the following: "As $z_{obt} > 2.33$, you would reject *H0*,

knowing that in an experiment where *H0* is true there is a 1 in 100 chance that the experimental result will lead to rejection of *H0* under the current decision criterion (Type I error)". This is an entirely different assertion from the one they actually made.

To illustrate the practical impact of these errors, we give two examples in the context of criminal justice. First, suppose that an accused criminal has a certain rare parasite, and a statistical test on biological matter known to be from the criminal with $\alpha = 0.1$ yields $p = 0.07$, where *H0* corresponds to the assertion that the actual criminal is parasite-free. At the trial the defense attorney, citing Healey, claims that there is a 10% chance that the actual criminal is parasite-free, meaning there is a 10% chance that his client is being wrongly convicted. But the test's finding shows nothing of the kind. The value of α gives the probability that a parasite-free individual will test positive for the parasite; it says *nothing* about the probability that person that has tested positive is actually parasite-free. As a second example, suppose that a corrupt prosecutor wants to find a scapegoat for a crime that he knows was committed by the mayor's son, who happens to be a college student. In this case, a sample of the actual criminal's handwriting is part of the evidence gathered from the crime scene. So the prosecutor obtains 10,000 writing samples from male high school students (preferably from "undesirable" neighborhoods). Then one by one, he randomly selects samples and conducts statistical tests with $\alpha = 0.0002$, where *H0* corresponds to the assertion that the writing sample is not from the criminal. He continues until he finds a student's writing sample for which the null hypothesis is rejected. The prosecutor then arrests the student, and at the trial bases his case on Healey's assertion that there is only a 1 in 5000 chance that the decision to convict is the wrong decision. Once again this is a false argument. The probability which α represents applies only to *not-yet-performed* experiments—it *cannot* be interpreted as a probability concerning the conclusion from a completed experiment. Note that in this case, the precondition that *H0* is true actually applies—and the decision to reject *H0* is wrong with probability 1, not α as implied by Healey.

3.2. Erroneous Interpretations of p-Value

In some textbooks, misinterpretations of significance level are compounded by misinterpretations of *p*-value. Bhatterjee [10] provides a good example of how easy it is to fall into error when one fails to recognize that alpha is a *conditional* probability. He writes ([10], pp. 129–30: italics ours):

> Sir Ronald A. Fisher, established the basic guidelines for significance testing. *He said that a statistical result may be considered significant if it can be shown that the probability of it being rejected due to chance is 5% or less.* In inferential statistics, this probability is called the *p*-value, 5% is called the significance level (α), and the desired relationship between the *p*-value and α is denoted as: $p \leq 0.05$. The significance level is the maximum level of risk that we are willing to accept as the price of our inference from the sample to the population. If the *p*-value is less than 0.05 or 5%, it means that we have a 5% chance of being incorrect in rejecting the null hypothesis or having a Type I error.

Before discussing interpretive errors in this passage, we must first address the sentence in italics, which is incorrectly stated. One possible correction is: "The result of a statistical test leads to the rejection of a null hypothesis if the probability of the result given the null hypothesis is true is 5% or less."

Note that Bhatterjee unfortunately fails to mention the condition "given the null hypothesis is true", which immediately leads him into trouble. He goes on to identify the *p*-value as the probability that the null hypothesis is wrongly rejected due to chance (presumably he is referring to the probability of an incorrect rejection in the experiment under consideration, although he is not entirely clear on this). Thus Bhatterjee's claim can be expressed mathematically as:

$$\Pr[H0 \text{ is wrongly rejected in an experiment} \mid p\text{-value} = q] = q \quad (7)$$

This is completely different from either of the above definitions of *p*-value, and can be shown mathematically to be incorrect (see Appendix). Bhatterjee then repeats the mistake of Bachman and Paternoster by claiming that α gives the probability that the null hypothesis is incorrectly rejected in experiments where the *p*-value is less than α (like Bachman and Paternoster, he treats Type I error as an experimental outcome). He does not appear to realize that this is inconsistent with his earlier claim that the *p*-value gives this probability.

Hopkins and Glass [5] similarly misinterpret *p*-value, although their mistake is phrased somewhat more subtly. In comparing the mean I.Q. of a sample of children to a hypothesized mean, they conduct a *z*-test and obtain the result $|z| = 2.67$ (which corresponds to a *p*-value of 0.0064). They conclude ([5], p. 222): "The probability of making a type-I error in such situations, when the absolute value of *z* is 2.67, is less than 0.01". Once again, type-I error is being identified here as the experimental result of wrongly rejecting *H0* rather than a conditional event. In a footnote on the next page they say, "The probability of a type-I error is generally not reported as 0.0064. The high degree of precision implicit in this 0.0064 value is accurate only if all statistical assumptions are perfectly achieved". Apparently Hopkins and Glass think that in a "statistically ideal" world, it is the *p*-value that determines the probability of a Type I error for the associated experiment.

Weisburd and Britt ([11], p. 125) make the following statement: "The estimate of the risk of Type I error that is associated with rejecting the null hypothesis in a test of statistical significance (based on a sample statistic) is called the observed significance level and is ordinarily represented by the symbol *p*". Here they interpret the *p*-value not as an *actual* probability, but an *estimated* probability (since they never define "risk", we can only infer that "risk" means "probability").

Suffice it to say that there is no mathematical justification for any of these authors' claims, no matter whether Type I error is interpreted as a conditional event or as an experimental result. In the former case, the (conditional) probability of Type I error is equal to the value of α set for the experiment, regardless of the experimental outcome. In the latter case, the *p*-value is insufficient to determine the probability that *H0* has been wrongly rejected, as we show in the Appendix. In fact, in any experiment in which *H0* has been rejected, depending on the circumstances the probability that this rejection is wrong can be anything from 0 to 1 regardless of the *p*-value. Consider for instance two drug-testing scenarios, the first of which uses urine samples from 1-year old infants, while the second uses urine samples from participants at the Cannabis Cup event in Denver, Colorado. In the former case, any rejection of *H0* (no recent marijuana use) is almost certain to be wrong; while in the latter case, virtually all rejections of *H0* are in fact correct.

Sometimes authors do their damage in "drive-by" remarks which are intended to give readers an intuitive idea of the meaning of *p*-value. Thus Anthony ([12], p. 226) states, "…you should remember what it means to have a *p*-value of 0.05. It means a Type I error will occur roughly 5% of the time."

Similarly, Suchmacher and Geller ([13], p. 44) assert that "Therefore, we can express a conclusion by stating that *p* (Type I error probability) inferred from a study is less than α (statistical significance level that corresponds to the highest tolerable cutoff for Type I error—often 5%, or 0.05), as pre-established by the investigator". When expressed as conditional probability equations, both Anthony's and Suchmacher and Geller's claims turn out to similar to Bhatterjee's.

4. Analysis of Causes of Erroneous Interpretations

In our examination of various textbooks, we have repeatedly encountered mistakes due to the mishandling of conditional events and conditional probabilities. Indeed, neither the notation nor the concept of a conditional probability is mentioned in the majority of statistics textbooks for the social sciences, whether or not they make erroneous statements. Such textbooks often take a cavalier attitude towards clarifying the preconditions for conditional probabilities. Consider this excerpt from Bakeman ([3], p. 27):

> If we set our alpha level to the conventional .05, then the probability that we will reject the null hypothesis wrongly, that is, make a Type I error, is also .05. After all, by setting the alpha level to .05 for a statistical test we commit ourselves to rejecting the null hypothesis if the result we obtain would occur 5% of the time or less given that the null hypothesis is true. If we did the same experiment again and again, and in fact there is no effect in the population, over the long run 95% of the time we would correctly claim no effect. But 5% of the time, just by the luck of the draw, we would wrongly claim an effect.

This paragraph, except for the first sentence, is precisely correct (though students may find it hard to follow). But the first sentence, leads the reader to believe that (in mathematical language)

$$\alpha = \Pr[H0 \text{ is wrongly rejected in an experiment} \mid \text{confidence level for the experiment} = \alpha] \quad (8)$$

which is false. The very title of the section ("Type I error: the risk of making a false claim") reinforces this same wrong impression. Bakeman never defines "risk" as being predicated on the condition that *H0* is true, which easily leads the reader to infer that "risk" is the same as "probability". The less careful reader will come away from this section with a faulty understanding of α.

The way *p*-value is ordinarily presented (including the fact that the letter "*p*" is used) suggests that *p*-value is a probability associated with the experiment it is computed from. But it is actually a probability associated with a hypothetical experiment *that in practice is never performed*. Our intuition also hinders rather than helps us in the interpretation of *p*-value. It seems obvious that given two different experiments that reject their respective null hypotheses, the experiment with the smaller *p*-value should have a smaller probability of Type I error. But in fact, if Type I error is interpreted as an event conditioned on the null hypothesis, then the probability of Type I error has nothing to do with the *p*-value. If on the other hand Type I error is identified as an incorrect rejection of the null hypothesis in the completed experiment, then correct analysis of this situation requires Bayes' theorem and calculations involving conditional probabilities. These paradoxical features of *p*-value show that it does not lend itself to an intuitive interpretation as a probability, and that attempts at such interpretations should be avoided.

5. Conclusions

Out of 24 social science statistics textbooks we surveyed, 8 made statements about α or *p*-value that were definitely false. Several others made ambiguous statements that could easily be misunderstood by students. This problem is not uncommon; neither is it unimportant. We have given examples in which misunderstandings lead to significant misinterpretations of experimental results.

We have showed that in general these misinterpretations are due to confusions over terminology. No matter how carefully explained, if terminology too closely resembles expressions in common use, then the common use will prevail and the careful explanations will be forgotten. The only effective solution to this problem is to reform the terminology. Certainly a less ambiguous definition of Type I error is called for, which makes it plain that it refers to a conditional event: for example, "In the case where *H0* is true, a *Type I error* is said to occur when *H0* is wrongly rejected".

If furthermore social scientists would start referring to α as a *conditional* probability, we believe this would put an end to most if not all of the persistent confusions concerning α.

As far as *p*-value, the situation is more problematic. It is unfortunate that *p*-values are so entrenched in the literature, for it quite possible to do statistical hypothesis testing without them. Realistically however, it seems that *p*-values are probably here to stay. So we propose that textbooks (and instructors) stick to defining *p*-value as an indicator of significance (as explained in Section 2), and avoid identifying *p*-value as a (conditional) probability.

Our study illustrates the subtle yet enormously powerful influence that terminology exerts on its users' thought process. Scientists cannot afford to become complacent in their terminology, but rather should encourage scrutiny for hidden biases—including scrutiny by experts outside their own field. Indeed, our results argue for closer collaboration between mathematicians and social scientists in the educational process. In many (if not most) American universities, statistics for social scientists is taught outside of the mathematics or statistics departments. No doubt this is because social science departments want to focus particularly on the methods and applications that are important to their own respective fields. The drawback to this approach is that errors such as we have described in this paper may go undetected. Our results suggest that social science departments should consider increased input from the math and/or statistics departments, perhaps (for instance) in the form of team-taught courses.

Acknowledgments

We wish to thank the reviewers for their valuable comments.

Author Contributions

The second author (Miller) performed the textbook research and furnished the preliminary draft; while the first author (Thron) is responsible for the mathematical analysis and the final form of the paper.

Abbreviations

α: Significance level;
H0: Null hypothesis;
H1: Alternative hypothesis.

Appendix A: Conditional Probability Calculations

In this section, we calculate several of the conditional probabilities mentioned in the foregoing discussion. For justification of these calculations, we refer the reader to any textbook on mathematical probability.

First, we compute

$$\Pr[H0 \text{ wrongly rejected in a given experiment} \mid H0 \text{ is rejected in the experiment}] \tag{A1}$$

which was mentioned in our discussion of Healey [8] in Section 3.1. In the following, we leave off "in the given experiment" for brevity of expression. So we have

$$\begin{aligned}&\Pr[H0 \text{ is true} \mid H0 \text{ is rejected}] = \Pr[H0 \text{ is true and } H0 \text{ is rejected}]/\Pr[H0 \text{ is rejected}]\\ &= \frac{\Pr[H0 \text{ is rejected} \mid H0 \text{ is true}]\,\Pr[H0 \text{ is true}]}{\Pr[H0 \text{ is rejected} \mid H0 \text{ is true}]\,\Pr[H0 \text{ is true}]+\Pr[H0 \text{ is rejected} \mid H0 \text{ is false}]\,\Pr[H0 \text{ is false}]}\\ &= \frac{\alpha\cdot\Pr[H0 \text{ is true}]}{\alpha\cdot\Pr[H0 \text{ is true}]+(1-\beta)\,\Pr[H0 \text{ is false}]}\\ &= \left(1+\frac{(1-\beta)(1-\Pr[H0 \text{ is true}])}{\alpha\cdot\Pr[H0 \text{ is true}]}\right)^{-1}\end{aligned} \tag{A2}$$

where we have denoted the (conditional) probability of Type II error by β according to the usual convention. In practice, it is impossible to determine β or Pr[*H0* is true], and these values could be anything from 0 to 1. Clearly this expression is not equal to α, as some writers claim.

Next we compute

$$\Pr[H0 \text{ is wrongly rejected in an experiment} \mid p\text{-value from the experiment} = q] \tag{A3}$$

which Hopkins and Glass [5] claimed was equal to q in a "statistically ideal" world. Again using abbreviated notation, we find

$$\begin{aligned}&\Pr[\text{H0 is wrongly rejected} \mid p\text{-value} = q]\\ &= \Pr[\ H0 \text{ is true and } p\text{-value} = q\ ] \,/\, \Pr[p\text{-value} = q]\\ &= \frac{\Pr[p\text{-value}=q \mid H0 \text{ is true}]\cdot\Pr[H0 \text{ is true}]}{\Pr[p\text{-value}=q \mid H0 \text{ is true}]\cdot\Pr[H0 \text{ is true}]+\Pr[p\text{-value}=q \mid H0 \text{ is false}]\cdot\Pr[H0 \text{ is false}]}\\ &= \frac{q\cdot\Pr[H0 \text{ is true}]}{q\cdot\Pr[H0 \text{ is true}]+\Pr[p\text{-value}<q \mid H0 \text{ is false}]\cdot\Pr[H0 \text{ is false}]}\\ &= \left\{1+\left(\frac{\Pr[p\text{-value}=q \mid H0 \text{ is false}]}{\Pr[p\text{-value}=q \mid H0 \text{ is true}]}\right)\left(\frac{1}{\Pr[H0 \text{ is true}]}-1\right)\right\}^{-1}\end{aligned} \tag{A4}$$

In general, both numerator and denominator in the expression $\left(\frac{\Pr[p\text{-value}=q \mid H0 \text{ is false}]}{\Pr[p\text{-value}=q \mid H0 \text{ is true}]}\right)$ are zero, but the expression still corresponds to a well-defined ratio of probability densities which in practice are impossible to estimate (as is the probability that *H0* is true). Certainly, the expression is not equal to q in general.

Finally we compute

$$\Pr[H0 \text{ is wrongly rejected in an experiment} \mid \text{confidence level for the experiment} = \alpha] \tag{A5}$$

which we mentioned in our discussion of Bakeman [3] in Section 4. The expression can be rewritten as (with our usual abbreviation)

$$\begin{aligned}&\Pr[H0 \text{ is wrongly rejected} \mid \text{confidence level} = \alpha \text{ and } H0 \text{ is true}]\cdot\Pr[H0 \text{ is true}]\\ &+ \Pr[H0 \text{ is wrongly rejected} \mid \text{confidence level} = \alpha \text{ and } H0 \text{ is false}]\cdot\Pr[H0 \text{ is false}]\end{aligned} \tag{A6}$$

The second term is zero, since it is impossible to wrongly reject *H0* in an experiment where *H0* is false. On the other hand, the first conditional probability is equal to α. In summary we have:

$$\begin{aligned}&\Pr[H0 \text{ is wrongly rejected in an experiment} \mid \text{confidence level for the experiment} = \alpha] \\ &\quad = \alpha \cdot \Pr[H0 \text{ is true in the experiment}]\end{aligned} \tag{A7}$$

Since Pr[*H0* is true in the experiment] is less than or equal to 1, one could say that the confidence level α for an experiment could be interpreted as a "maximum risk" of wrongly rejecting *H0*. We concur, but this means something completely different from what many writers appear to think. In particular, it has nothing whatsoever to do with the probability that rejection of *H0* in an already-completed experiment constitutes a Type I error. In the district attorney example discussed in the text, this probability is equal to 1; and it is equally possible to construct a scenario where the probability is equal to 0.

Appendix B: Textbooks Examined in the Study

Anthony, Denis. *Understanding Advanced Statistics: A Guide for Nurses and Health Care Researchers*. Edinburgh: Churchill Livingstone, 1999.

Bachman, Ronet, and Raymond Paternoster. *Statistical Methods for Criminology and Criminal Justice*. New York: Mcgraw-Hill, 1997.

Bakeman, Roger. *Understanding Social Science Statistics: A Spreadsheet Approach.* Hillsdale: Lawrence Erlbaum Associates, 1992.

Bhattacherjee, Anol. *Social Science Research: Principles, Methods, and Practices*. Tampa: Global Text Project, 2012. Available online: http://scholarcommons.usf.edu/oa_textbooks/3 (accessed on 7 May 2015).

Blalock, Hubert M. *Social statistics*, rev. 2nd ed. New York: McGraw-Hill, 1979.

Bland, Martin. *An Introduction to Medical Statistics*. Oxford: Oxford University Press, 1987.

Cotton, John W. *Elementary Statistical Theory for Behavioral Scientists*. Reading: Addison Wesley, 1967.

Couch, James V. *Fundamentals of Statistics for the Behavioral Sciences*. New York: St. Martin's, 1982.

Daly, Leslie, and Geoffrey J. Bourke, *Interpretation and Uses of Medical Statistics*, 5th ed. Oxford: Blackwell Science, 2000.

Healey, Joseph F. *Statistics: A Tool for Social Research*, 9th ed. Stamford: Cengage Learning, 2011.

Hopkins, Kenneth D., and Gene V. Glass. *Basic Statistics for the Behavioral Sciences*. Englewood Cliffs: Prentice-Hall, 1978.

Jackson, Sherri. *Statistics and Research Design (Custom edition for Psychology 418 at the University of Texas at Austin)*. Mason: Cengage Learning, 2008.

Jordan, Kelvin, Bie No Ong, and Peter Croft. *Mastering Statistics: A Guide for Health Service Professionals and Researchers*. Cheltenham: Stanley Thornes Ltd., 1998.

Mueller, John H., and Karl F. Schuessler. *Statistical Reasoning in Sociology*, 3rd ed. Boston: Houghton Mifflin, 1977.

Oakes, Michael W. *Statistical Inference*. Chestnut Hill: Epidemiology Resources, Inc., 1990.

Runyon, Richard, Kay Coleman, and David Pittenger. *Fundamentals of Behavioral Statistics*, 9th ed. Hawkins: McGraw-Hill Higher Education, 2000.

Sedlmeier, Peter. *Improving Statistical Reasoning Theoretical Models and Practical Implications Account*. Mahwah: Lawrence Erlbaum Associates, 1999.

Suchmacher, Mendel, and Mauro Geller. *Practical Biostatistics: A User-Friendly Approach for Evidence-Based Medicine*. Amsterdam: Academic, 2012.

Schefler, William C. *Statistics for Health Professionals*. Reading: Addison-Wesley, 1984.

Weinbach, Robert W., and Richard M. Grinnell. *Statistics for Social Workers*, 4th ed. New York: Longman, 1998.
Weisburd, David, and Chester Britt. *Statistics in Criminal Justice*, 3rd ed. New York: Springer Science & Business Media, LLC, 2007.
Welkowitz, Joan, Barry H. Cohen, and R. Brooke Lea. *Introductory Statistics for the Behavioral Sciences*. New York: Wiley, 2011.
Wilcox, Rand R. *New Statistical Procedures for the Social Sciences: Modern Solutions to Basic Problems*. Hillsdale: Lawrence Erlbaum Associates, 1987.
Willemsen, Eleanor W. *Understanding Statistical Reasoning*. San Francisco: W. H. Freeman, 1974.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Oakes, Michael W. *Statistical Inference*. Chestnut Hill: Epidemiology Resources, Inc., 1990.
2. McClave, James, and Terry Sincich. *Statistics*, 12th ed. Boston: Pearson, 2013.
3. Bakeman, Roger. *Understanding Social Science Statistics: A Spreadsheet Approach.* Hillsdale: Lawrence Erlbaum Associates, 1992.
4. Daly, Leslie, and Geoffrey J. Bourke. *Interpretation and Uses of Medical Statistics*, 5th ed. Oxford: Blackwell Science, 2000.
5. Hopkins, Kenneth D., and Gene V. Glass. *Basic Statistics for the Behavioral Sciences*. Englewood Cliffs: Prentice-Hall, 1978.
6. Welkowitz, Joan, Barry H. Cohen, and R. Brooke Lea. *Introductory Statistics for the Behavioral Sciences*. New York: Wiley, 2011.
7. "Type I and Type II Errors." *Wikipedia, the Free Encyclopedia*. Available online: http://en.wikipedia.org/wiki/Type_I_and_type_II_error.s (accessed on 22 April 2015).
8. Healey, Joseph F. *Statistics: A Tool for Social Research*, 9th ed. Stamford: Cengage Learning, 2011.
9. Bachman, Ronet, and Raymond Paternoster. *Statistical Methods for Criminology and Criminal Justice*. New York: Mcgraw-Hill, 1997.
10. Bhattacherjee, Anol. *Social Science Research: Principles, Methods, and Practices*. Tampa: Global Text Project, 2012. Available online: http://scholarcommons.usf.edu/oa_textbooks/3 (accessed on 7 May 2015).
11. Weisburd, David, and Chester Britt. *Statistics in Criminal Justice*, 3rd ed. New York: Springer Science & Business Media, LLC, 2007.
12. Anthony, Denis. *Understanding Advanced Statistics: A Guide for Nurses and Health Care Researchers*. Edinburgh: Churchill Livingstone, 1999.
13. Suchmacher, Mendel, and Mauro Geller. *Practical Biostatistics: A User-Friendly Approach for Evidence-Based Medicine*. Amsterdam: Academic, 2012.