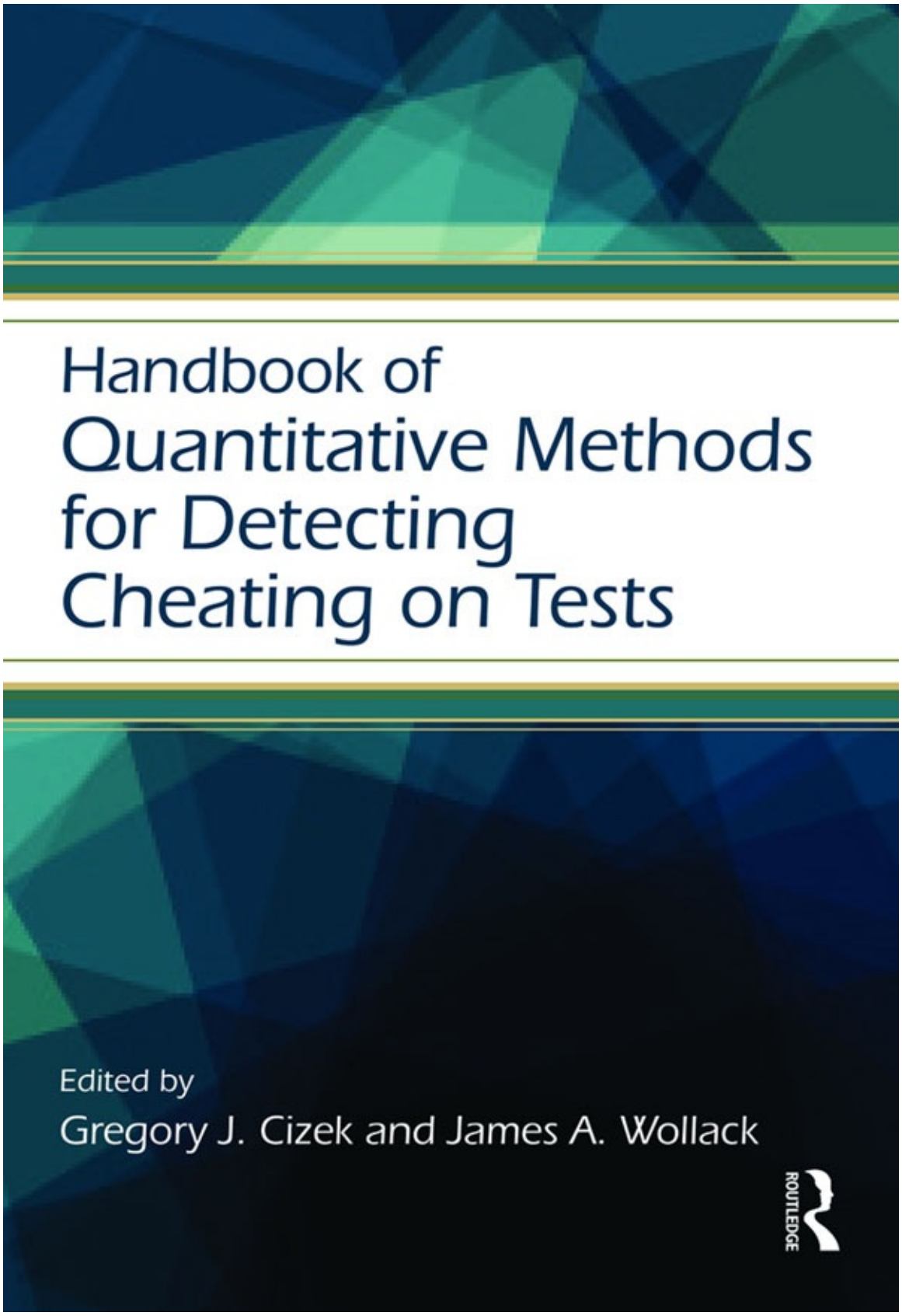
Handbook of
Quantitative Methods
for Detecting
Cheating on Tests
Edited by
Gregory J. Cizek and James A. Wollack
ROUTLEDGE

First published 2017
by Routledge
711 Third Avenue, New York, NY 10017

and by Routledge
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

Library of Congress Cataloging-in-Publication Data
A catalog record for this book has been requested

ISBN: 978-1-138-82180-4 (hbk)
ISBN: 978-1-138-82181-1 (pbk)
ISBN: 978-1-315-74309-7 (ebk)

Typeset in Minion
by Apex CoVantage, LLC

18

The Case for Bayesian Methods when Investigating Test Fraud

William P. Skorupski and Howard Wainer

> "I would never die for my beliefs because I might be wrong."
>
> *Bertrand Russell*

Introduction

Examinee cheating is always a concern for testing programs with high stakes (e.g., Cizek, 1999; Thiessen, 2007). There are obvious issues with fairness involved (students receiving higher scores than they deserve, contaminating criterion- and norm-referenced inferences), as well as concerns for the psychometric integrity of scores from assessments where cheating has occurred. As such, cheating is a validity issue that affects not only individual ability estimates by potentially biasing them upwards but may in fact result in other fairly earned ability estimates being biased downwards (in terms of their relative position in the score distribution).

Cheating behavior may occur in multiple ways. Many research studies over the years have conceptualized cheating in terms of collusion among examinees, aberrant response patterns characterized by lack of person fit, unexpected score gains, and suspicious answer changes or erasures (e.g., Angoff, 1974; Cannell, 1989; Meijer & Sijtsma, 1995; Wollack, 1997, 2003). Ultimately, cheating may be manifested in a number of different ways, so it is reasonable to consider multiple approaches to detection (while keeping in mind the possible inflation of Type I error that may occur).

The purpose of this chapter is not to focus on individual cheating detection methods, but rather to provide logical and analytic evidence to encourage Bayesian reasoning for cheating detection, regardless of the method employed. Many methods currently rely on traditional null hypothesis testing to flag potential cheaters for further review. We present an argument

here that these frequentist inferences are often misleading (e.g., Gelman, 2011), and that this is especially true with regard to establishing the probability of very-low-incidence events (Savage & Wainer, 2008), such as cheating. Our goal is push social scientists to consider that the traditional frequentist *p*-value is really the wrong *P*. We posit that most cheating detecting investigations are not really interested in the probability of observing data, given the null "not a cheater" condition (Gelman, 2013); rather, the inference of interest is the probability that a condition ("is a cheater") has been met, given what has been observed in the data. This is precisely what the Bayesian paradigm provides. In practical terms, we will demonstrate in the following sections of this chapter that it is much more useful to know, for example, that there is a 90% chance that someone is a cheater, given the data, as opposed to saying that there is an infinitesimal chance of observing this person's data if he or she were a noncheater. Statements like the latter are confusing at best, misleading at worst. In short, we advocate for using Bayesian posterior inferences instead of these troublesome *p*-values.

For any high-stakes operational testing program, it is of vital importance to incorporate powerful methods to detect cheating behavior. Consequently, cheating detection methods and their evaluation have become increasingly prevalent in the literature. However, although statistical power is of great importance, so is maintaining an acceptable Type I error rate that avoids false positives. Whereas the need to catch cheaters is great, the need to protect noncheaters from false accusations is perhaps equally important, if not more so.

This begs the question: What are the consequences associated with false positives and false negatives? That depends, in large part, on the unit of analysis and the course of action taken as a result of being identified as a possible cheater. For example, when a teacher or administrator is accused of cheating (as events in the Atlanta Public Schools case have demonstrated), the cost of being flagged as a cheater may be the loss of one's career and criminal charges. If those are false positives (we are not suggesting they are), those would be dire consequences, indeed. The consequences for an individual examinee accused of cheating might result in having to retake a test, which doesn't sound too harsh, or possibly being barred from further testing, which could mean exclusion from one's chosen career. It should be obvious that extreme caution must be exercised when flagging examinees as potential cheaters. The cost of a false negative—the cheater gets away with it—may or may not be just as serious. One only has to imagine being on the operating table of a surgeon who cheated on a medical board examination to appreciate the potential seriousness of false negatives. Thus, when discussing the performance of a cheating detection statistic, one must sensibly evaluate its sensitivity and specificity in terms of how cheating behavior is flagged, and the relative costs and benefits associated with being wrong or being right.

Consider a motivating example, adapted from a line of reasoning presented in Savage and Wainer (2008), arguing for Bayesian methods when detecting suspected terrorists. The context and numbers used here are different, but the approach is identical. Suppose there is a test-

taking population of 70,000 examinees—a number not uncommon for a statewide testing program. Furthermore, suppose that 5% of the test-taking population is comprised of cheaters. Finally, suppose that a recently developed statistic, x, shows remarkable promise; it has been shown that using a critical value of X is "99% accurate" both in terms of its sensitivity (i.e., statistical power to detect true cheaters) and its specificity (i.e., true negative rate, or ability to accurately classify noncheaters as noncheaters). This sounds too good to be true, and it probably is. In reality, sensitivity and specificity may not be equal, as one is usually sacrificed in favor of the other; indeed, they are inversely related and we make this assumption for the sake of argument, to demonstrate a best-case scenario.

If the observed value of the statistic, x, for examinee i (x_i) is greater than X (the critical value that promises 99% accuracy), that examinee is flagged as a potential cheater. When x_i is large, the magnitude of this departure from expectation is summarized through traditional null hypothesis testing: given the individual is *not* a cheater (~C), what is the probability of observing a value this large (or larger) for x_i? To answer this, we determine $P(x_i \geq X \mid \sim C)$ by calculating the area in the right-hand-tail of the null sampling distribution of X. If that probability is sufficiently small (i.e., it falls below an a priori acceptable Type I error rate), then we make the usual inference that the person is probably a cheater; more formally, we say that we reject the null hypothesis. However, that conclusion is not terribly clear to those untrained in statistical inference. More important, it is not a direct inference about the actual subject of interest: What is the probability that the examinee is a cheater? Our evidence thus far indicates that for a noncheater, x_i is an unusual result. But does that imply that examinee i must be a cheater? The answer is: "it might, but we need to know more about the distribution of cheaters to be confident." The fact is, relying on this traditional frequentist interpretation of p-values alone to flag potential cheaters will result in a lot of incorrect/false positive decisions.

If N = 70,000, and 5% are cheaters, then the number of cheaters, N_C, is 3,500 and the number of noncheaters, $N_{\sim C}$, is 66,500. Of the 3,500 cheaters, 99% are accurately flagged by X, resulting in 3,465 correctly identified cheaters. Of the 66,500 noncheaters, only 1% are inaccurately flagged by X, resulting in 665 incorrectly identified cheaters. The total number of flagged individuals is 4,130, or 5.9% of the 70,000 examinees. From this, we can determine the probability that an examinee is a cheater, given she or he was flagged, is 3,465/4,130 = 0.84, or 84% accuracy. That rate is not terribly low, but it's not the "99% accuracy" promised. Worse, the complement to this expression is the probability that an examinee is a noncheater, given that she or he was flagged, which is 665/4,130 = 0.16, a false positive rate of 16%, a far cry from the expected 1% rate.

It gets even worse if the incidence of cheating is far less than the previously assumed 5% marginal proportion. Suppose that the detection statistic, x, and associated critical value, X, can't detect all cheaters, only those who copy answers (in point of fact, different statistical methods are required for detecting different kinds of cheating, because different kinds of

cheating don't leave the same evidentiary trail). If the proportion of answer copiers in the population is only 1%, then of the 70,000 examinees, N_C is down to 700, whereas the number of noncheaters, $N_{\sim C}$, is up to 69,300. Of the 700 cheaters, 99% are accurately flagged by X, resulting in 693 correctly identified cheaters. Of the 69,300 noncheaters, only 1% are inaccurately flagged by X, resulting in 693 false positives. Thus, the total number of flagged individuals is 1,386, just under 2% of the total, half of which are true cheaters, half of which are not. One would hardly be impressed with a detection statistic with only 50% accuracy, given a flag has occurred, especially if 99% accuracy is expected.

How can this be? What happened to the "99% accuracy?" The problem is that the null hypothesis p-value is providing an accurate representation of the wrong probability. The frequentist p-value is notoriously misunderstood. In this hypothetical example, it is the probability of someone being flagged, given that the person is a noncheater: $P(x_i \geq X|\sim C)$. This probability cannot account for the fact that the marginal proportion of cheaters and noncheaters may be dramatically different; it is furthermore unaffected by *how different* the distribution of X may be for cheaters and noncheaters. As a result, the p-value is a fairly obtuse way of making an inference about something very acute. It is difficult for many to understand "the probability of observing a value this large or larger if the null condition were true." People will also often erroneously take 1-p to be the probability that the null hypothesis is false (e.g., Gelman, 2013), which of course it isn't: 1-p is the probability of a true negative; the probability that a noncheater would demonstrate a value less than X. That is also useful information, but it likewise does not really tell us anything about whether examinee i is a cheater or not.

Thus, the claim may be true that only "1% of the noncheaters will be erroneously flagged as cheaters," but this does *not* mean that only 1% of the flagged are noncheaters. This is a typical, and potentially very serious, misinterpretation. This misinterpretation is responsible for a number of erroneous conclusions. For example, as Wainer (2011) has shown, even though mammography is as high as 90% accurate (in terms of sensitivity and specificity), because the incidence rate of breast cancer is so small *compared to* the rate at which women are screened, the false positives far outweigh the true positives. He estimated that the probability of a woman having breast cancer, given a positive mammogram, is a shockingly low 5%, meaning that 95% of women who receive a positive mammogram do not have breast cancer. As we have seen here, the false positive rate *for those flagged* is sure to be much larger than the marginal false positive rate, so claims of "99% accuracy" are potentially very misleading. To estimate the correct probability—that is, the probability that someone is a cheater, given that he or she has been flagged, $P(C|x_i \geq X)$—one needs to employ Bayesian reasoning.

Bayes' Rule

Bayes' Rule, shown in Equation 1, is a formula that deals with conditional probability statements:

$$P\left(\theta \mid x\right) = \frac{P(\theta, x)}{P(x)} = \frac{P(x \mid \theta)P(\theta)}{P(x)}\,. \tag{1}$$

It states that the probability of θ, given x (termed the *posterior distribution of θ given x*) is equal to the joint probability of θ and x divided by the marginal probability of x. This expression is equivalent to the probability of x given θ, (often referred to as the *likelihood function of x*, especially in the context of statistical analysis) multiplied by the marginal probability of θ (referred to as the *prior distribution*, as it represents the distribution of θ without regard to x) divided by the marginal probability of x. This result seems benign in and of itself. It is, in fact, an effective way of solving a number of important conditional probability problems, such as, "what are the chances someone is a terrorist, given some suspicious behavior?" (Savage & Wainer, 2008), or, "what is the probability a women has breast cancer, given a positive mammogram result?" (Wainer, 2011).

However, the implementation of Bayes' Rule for statistical data analysis in the social sciences can be contentious. Using the notation above, the values of θ may be the parameters of a statistical model, or some other estimand of interest, and the values in x are the observed data. In terms of cheating detection, θ could be characterized as "examinee i is a cheater" (which will henceforth be denoted with a "C"), and x could be characterized as "observing a value for our cheating detection statistic, x_i, greater than or equal to X." Equation 2 reframes Bayes' Rule in that context:

$$P\left(C \mid x_i \geq X\right) = \frac{P(x_i \geq X \mid C)P(C)}{P(x_i \geq X)}\,. \tag{2}$$

A few definitions follow, leading to an expression for how to solve Equation 2 in practical terms. First, the denominator of this formula, $P(x_i \geq X)$, represents the marginal distribution of extreme X values; that is, it is the proportion of all observed x_i values which are greater than or equal to X. If one has established a detection threshold via an acceptable marginal Type I error rate (i.e., the a level) and associated critical value, then $P(x_i \geq X)$ is simply the proportion of observed values that meets or exceeds that threshold; this is easily determined from the data distribution. In the motivating example, this probability came from counting up the total

number of flagged examinees. (The example used ratios of frequencies, both of which would have been divided by a constant, *N*, to make them probabilities, so the *N* was left out.)

The numerator of the formula contains a product of two terms that are not directly observed, but may be estimated based on weak assumptions. $P(x_i \geq X|C)$ is the likelihood of observing a value of x_i greater than or equal to X, given examinee i is a cheater. In this context the likelihood represents the power of the statistical test, the probability that a flag is obtained for a true cheater. $P(C)$ is the *prior* probability for cheaters, which represents the proportion of cheaters in the population. This may be known a priori, based on previous experience, or the analyst may make a reasoned estimate of its value; regardless, the impact of the prior tends to be the most contentious issue in Bayesian analysis, so this topic will be addressed directly in a subsequent section of this chapter. For now, we continue on with estimating the numerator of the posterior density expression. To arrive at the estimate, we first note that the marginal density, $P(x_i \geq X)$, by definition is equal to:

$$\begin{aligned} P(x_i \geq X) &= \sum_{\theta} P\left(x_i \geq X \mid \theta\right) P\left(\theta\right) \\ &= P(x_i \geq X \mid C)P(C) + P(x_i \geq X \mid \sim C)P(\sim C). \end{aligned} \tag{3}$$

In Equation 3, θ represents any unknown parameter and all of its possible values; in this case, those values are "examinee i is a cheater," represented by C, and "examinee i is a noncheater," represented by ~C. Those are the only two possible values for the parameter of interest. Solving for the posterior's numerator, $P(x_i \geq X|C)P(C)$, we obtain the Equation 4:

$$P(x_i \geq X \mid C)P(C) = P(x_i \geq X) - P(x_i \geq X \mid \sim C)P(\sim C). \tag{4}$$

In the motivating example, the numerator was estimated by counting up the number of correctly flagged cheating examinees (as previously stated, that example used ratios of frequencies, not probabilities, but operated on the same principles).

There is good news and bad news here for the analyst who is unconvinced that the Bayesian approach is preferred. The good news is that there are some familiar terms that can be estimated directly from the data. As mentioned, $P(x_i \geq X)$ is simply the marginal proportion of x_i values greater than or equal to X. $P(x_i \geq X|\sim C)$ is the probability that $x_i \geq X$, given examinee i is a noncheater. This expression is the well-known p-value from a null hypothesis test. The potential bad news is that $P(\sim C)$ and $P(C)$, the prior probabilities of noncheating and cheating, respectively, cannot be directly estimated from the data. In the motivating example, we posited the marginal incidence of cheating was 5% or 1%. In practice, this probability would have to be estimated. Informed judgment or previous experience and data may supply

reasonable estimates, but the analyst only has to estimate one of these: because $P(\sim C) + P(C) = 1$, every examinee is by definition either a noncheater or a cheater. Assuming for now an estimate can be obtained, we return to the solution for the posterior probability of cheating, given the data, substituting into the formula the result for its numerator:

$$P\left(C \mid x_i \geq X\right) = \frac{P(x_i \geq X \mid C)P(C)}{P(x_i \geq X)} \tag{5}$$

$$= \frac{P(x_i \geq X) - P(x_i \geq X \mid \sim C)P(\sim C)}{P(x_i \geq X)} \tag{6}$$

$$= \frac{P(x_i \geq X)}{P(x_i \geq X)} - \frac{P(x_i \geq X \mid \sim C)P(\sim C)}{P(x_i \geq X)} \tag{7}$$

$$= 1 - \frac{P(x_i \geq X \mid \sim C)P(\sim C)}{P(x_i \geq X)} \tag{8}$$

$$= 1 - P(\sim C \mid x_i \geq X). \tag{9}$$

Equation 9 is just another application of Bayes' Rule. This result is mathematically consistent and logical because $P(C|x_i \geq X) + P(\hat{C}|x_i \geq X) = 1$. For any given value of x_i greater than or equal to X, examinee i must either be a cheater or a noncheater.

Thus, there is a relatively simple formula to solve now, and all it requires is to estimate (1) the marginal proportion of x_i values greater than or equal to X, (2) the null hypothesis p-value associated with this threshold, and (3) a reasonable estimate of the proportion of cheaters (and, correspondingly, the proportion of noncheaters) in the population. We can then turn a frequentist p-value into a Posterior Probability of Cheating (PPoC) as follows:

$$PPoc = 1 - \frac{P(x_i \geq X \mid \sim C)P(\sim C)}{P(x_i \geq X)} \tag{10}$$

$$= 1 - \frac{\text{p-value} \times P(\text{non} - \text{cheater})}{P(\text{data above threshold})}, \quad (11)$$

where the *p*-value in the numerator of Equation 11 is the null hypothesis *p*-value. That value is multiplied by an estimate of the proportion of the population who are not cheaters, and the result is divided by the proportion of examinees who demonstrate x_i values above the threshold, *X*. This is the posterior probability of being a *noncheater*, given the data; subtracting that value from one provides the PPoC. A useful feature is that from these two probabilities – $P(C|x_i \geq X)$ and $P(\hat{C}|x_i \geq X)$ – one can construct Bayes factors (Jeffreys, 1960), which are odds ratios, in this case, $P(C|x_i \geq X)/P(\hat{C}|x_i \geq X)$. This simple transformation conveys the probability that a flag indicates a true cheater on the odds scale.

Method

The benefits of calculating the PPoC, as opposed to relying on traditional *p*-values, is demonstrated by means of a series of analytic examples. This analysis does not focus on particular cheating statistics, but rather generalizes detection to any such statistic. These analytic examples demonstrate the problem with null hypothesis testing and "statistical significance" for very low-incidence events.

First, consider a hypothetical cheating detection statistic, *x*. Under the null hypothesis (H_0), the sampling distribution of *x* is standard normal: $x|{\sim}C \sim N(0, 1)$. Further, assume that 1% of the population is comprised of cheaters; thus $P(C) = 0.01$ and $P(\sim C) = 0.99$. Sixteen conditions were created to represent various cheating detection scenarios: four detection threshold values, X_c, representing increasing specificity (threshold X_c values = 2, 3, 4, 5), crossed with four expected values for the sampling distribution of *x* for cheaters. Thus, $x|C \sim N(\mu, 1)$, with $\mu = 2, 3, 4, 5$. For each condition, a unique mixture distribution of normal distributions is constructed and the PPoC is calculated by using true population values for three proportions: $P(x_i \geq X_c)$, the marginal proportion of x_i values greater than or equal to *X;* $P(x_i \geq X_c|{\sim}C)$, the null hypothesis *p*-value associated with this threshold; and $P(\sim C)$, the true proportion of noncheaters, equal to 0.99. Because all supplied values are analytically derived, the resulting PPoC values are true. As previously stated, with real data $P(x_i \geq X_c)$ and $P(x_i \geq X_c|{\sim}C)$ could be calculated directly, but $P(\sim C)$ would, in practice, have to be estimated. As such, the influence of correctly or incorrectly specifying $P(\sim C)$ is then further considered.

Results and Discussion

Tables 18.1–18.3 contain true values for the likelihood values, $P(x_i \geq X|C)$, marginal distributions for $x_i \geq X$, $P(x_i \geq X)$, and the PPoC values, $P(C|x_i \geq X)$, respectively. Each table contains these values for the 4 × 4 crossed conditions. For each of the four X thresholds, the corresponding column is labeled by its null hypothesis alpha level (i.e., the area in the right-hand-tail of the null distribution, or the minimally sufficient p-value to reject H_0). The true likelihood values and marginal distributions of $x_i \geq X$ are included in Tables 18.1 and 18.2 as reference points. PPoC values in Table 18.3 are of primary interest for inference making. These can be constructed by multiplying a corresponding likelihood by 0.01 (the prior probability of cheating) and dividing by $P(x_i \geq X)$.

Table 18.1 True Likelihood (Power) Values, $P(x_i \geq X_c|C)$, by Condition

	Detection threshold X			
μ for cheaters	**X = 2, p-value ≤ 2.275×10^{-2}**	**X = 3, p-value ≤ 1.350×10^{-3}**	**X = 4, p-value ≤ 3.167×10^{-5}**	**X = 5, p-value ≤ 2.867×10^{-7}**
2	0.500	0.159	0.023	0.001
3	0.841	0.500	0.159	0.023
4	0.977	0.841	0.500	0.159
5	0.999	0.977	0.841	0.500

The likelihood values in Table 18.1 represent the statistical power of each null hypothesis test. That is, they represent the probability that the threshold will correctly identify a true cheater. The likelihood values show that when the X threshold is equal to the expected value of the cheaters' distribution, there is only a 50% chance that $x_i \geq X$ for cheaters (i.e., the threshold is the median of the score distribution for cheaters). When the expected value exceeds the threshold, the power is considerably higher. Conversely, when the threshold is below the expected value, the probability that $x_i \geq X$ for cheaters is quite low. These changes are noteworthy because their magnitudes are considerably larger than the corresponding changes in p-values. Type I error and power are always directly related, though not linearly.

Table 18.2 True Marginal Distributions of $x_i \geq X_c$ $P(x_i \geq X)$, by Condition

	Detection threshold for x

ft for cheaters	X = 2, p-value ≤ 2.275×10^{-2}	X = 3, p-value ≤ 1.350×10^{-3}	X = 4, p-value ≤ 3.167×10^{-5}	X = 5, p-value ≤ 2.867×10^{-7}
2	0.028	0.003	<0.001	<0.001
3	0.031	0.006	0.002	<0.001
4	0.032	0.010	0.005	0.002
5	0.033	0.011	0.008	0.005

Table 18.3 True PPoC Values, P(C|$x_i \geq X_c$), by Condition

	Detection threshold for *x*			
μ for cheaters	X = 2,p-value ≤ 2.275×10^{-2}	X = 3,p-value ≤ 1.350×10^{-3}	X = 4,p-value ≤ 3.167×10^{-5}	X = 5,p-value ≤ 2.867×10^{-7}
2	0.18	0.54	0.88	0.98
3	0.27	0.79	0.98	0.99
4	0.30	0.86	0.99	0.99
5	0.31	0.88	0.99	0.99

The $P(x_i \geq X)$ values in Table 18.2 are consistent with expectation. As μ for cheaters increases, so does the marginal proportion of extreme x values. Conversely, as the detection threshold is increased, the marginal proportion of extreme x values decreases. The fact that these proportions are so small (ranging from practically zero to no higher than 3.25%) is important for explaining the PPoC values in Table 18.3.

The PPoC values are presented in Table 18.3 and illustrated in Figures 18.1 and 18.2. Figure 18.1 illustrates how the PPoC is calculated, using the example of detection threshold $X_c = 2$, and $\mu|C = 5$. In this example, even though the likelihood, $P(x_i \geq X_c|C)$, is 0.999, only 1% of the total population is comprised of cheaters, $P(C) = 0.01$. With the threshold set at $X_c = 2$, there are 2.275 times more incorrectly flagged noncheaters than there are correctly flagged cheaters. Thus, the PPoC is only 0.31. Figure 18.2 shows this same relationship, with the area around the flagging region magnified to enhance the details of the two distributions.

As evidenced by these results, there can be a considerable difference between finding a statistically significant *p*-value (i.e., because $x_i \geq X_c$) from a null hypothesis test, and having a reasonably high probability of correctly identifying a cheater. For low-incidence events such as these, if the detection threshold is relatively low (e.g., only two standard errors above the mean), the probability that examinee i is a cheater, given $x_i \geq X$, is considerably lower than the probability that he or she is a noncheater. For any value of μ, as the detection threshold is

increased, so does the PPoC. The increase

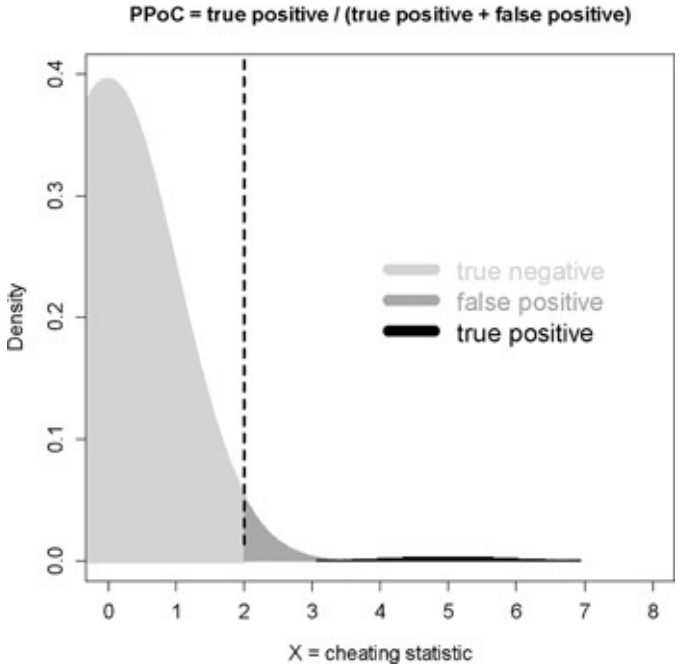

Figure 18.1 The PPoC, $P(C \mid x_i \geq X)$, is the Marginal Proportion of true positives, $P(x_i \geq X \mid C)\, P(C)$, divided by the probability of being flagged, $P(x_i \geq X)$. This is equivalent to the number of true positives divided by the number of flagged examinees. In this example, $X = 2$ and $m|C = 5$.

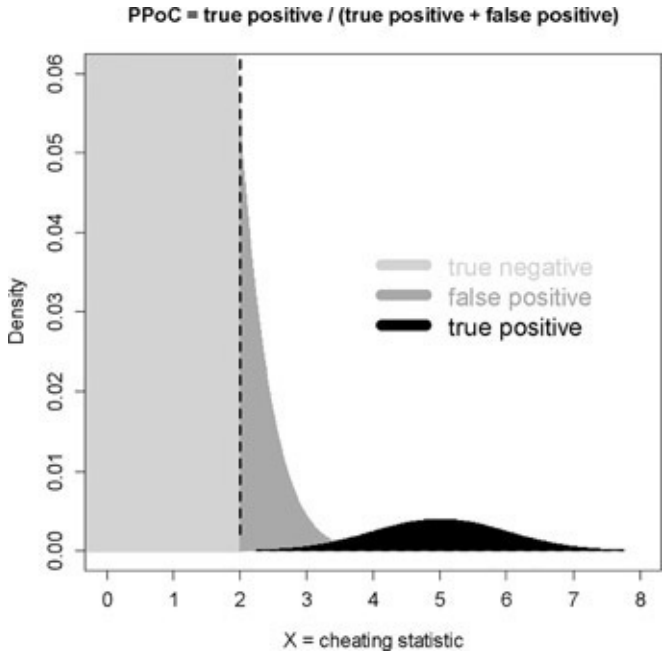

Figure 18.2 Magnification of the Flagging Region Shown in Figure 18.1 to Enhance Detail

in PPoC is larger for increases in the threshold than it is for increases in μ. When the detection threshold is set to $X = 5$ (five standard errors above the mean), the probability that examinee i is a cheater, given $x_i \geq X$, is nearly one. Of course, such a large threshold may very well

sacrifice sensitivity for the sake of specificity; in this example, thresholds of three and four are higher-powered but still have fairly large PPoC values.

This analytic exercise is instructive, but it assumes one knows the prior probability of cheating, which is unlikely to be true in practice. How does one proceed without knowing that value? One answer is to consider a range of plausible priors, and see how that choice influences the posterior inferences. If the inferences are greatly influenced by the priors, then the analyst knows that the information in the likelihood is not very strong, whereas small changes in the posterior probabilities mean that the prior's influence is not very strong. To demonstrate these differences, we extend the analytic demonstration to a case that approximates reality, whereby $P(C)$ is not known and must be estimated.

Table 18.4 contains a series of PPoC values derived by iteratively changing the value of $P(C)$. In this example, the cheater distribution expected value is held constant ($\mu = 5$, corresponding to the last row in Table 18.3), but the choices for $P(C)$, ranging from 0.001 to 0.5, are completely crossed with the four threshold conditions. We see that when the correct prior is used (in the highlighted row), we obviously obtain the correct PPoC value. But another interesting pattern is clear: The estimate of PPoC is much less influenced by the choice of prior when the true PPoC value is closer to one. For the $X_c = 2$ threshold, PPoC estimates ranged from 0.30 for $P(C) = 0.001$ up to 0.65 for $P(C) = 0.5$. Estimating the percentage of cheaters at 50% is almost certain to be an overestimate (one would hope), but these extreme cases give us insight into the stability of the PPoC. Conversely, for the $X = 3$ threshold, PPoC estimates ranged from 0.88 for $P(C) = 0.001$ up to 0.94 for $P(C) = 0.5$, a range of only 0.06. For the $Xc = 3$ and $Xc = 4$ thresholds, the range of PPoC estimates is practically zero.

Table 18.4 Estimated Bayesian PPoC Values by Detection Threshold (X) (and Associated Null P-Value) Crossed With Various Prior Specifications for Cheating Prevalence

Prior specification for Marginal Probability of Cheating, *P(C)*	$X_c = 2$, p-value $\leq 2.275 \times 10^{-2}$	$X_c = 3$, p-value $\leq 1.350 \times 10^{-3}$	$X_c = 4$, p-value $\leq 3.167 \times 10^{-5}$	$X_c = 5$, p-value $\leq 2.867 \times 10^{-7}$
0.001	0.301	0.879	0.996	0.999
0.01	0.307	0.880	0.996	0.999
0.05	0.335	0.885	0.996	0.999
0.1	0.370	0.891	0.997	0.999
0.2	0.440	0.903	0.997	0.999
0.3	0.510	0.915	0.997	0.999
0.4	0.580	0.927	0.998	0.999
0.5	0.650	0.939	0.998	0.999

Note: Calculations are computed here for the true cheating distribution with $\mu = 5$. The highlighted row, $P(C) = 0.01$, is correctly specified.

Table 18.3 and Figure 18.1 helped establish that PPoC values are closer to one when the detection threshold moves away from the null distribution (and, to a lesser extent, when the distribution of x for cheaters moves further away from the null distribution). The reason why the threshold is more influential than the expected value of the cheater distribution is due to the very large discrepancy in the prior probabilities of cheating and not cheating. When 99% of the frequency distribution is noncheaters, the only way to confidently say a flag probably indicates a cheater is to make the threshold so high that practically all noncheaters would be excluded. At that point, within the distribution of examinees for whom $x_i \geq X$, there are more cheaters by far than noncheaters. It is precisely this kind of inference making that the Bayesian paradigm encourages and that the frequentist reliance on p-values completely ignores.

Conclusion

In 1976, George Box famously stated "all models are wrong" (p. 792) but may nonetheless be useful, especially when parsimonious. Bayesian methods persist because they are eminently useful, and can be applied parsimoniously, even if they are sometimes wrong. An important feature of these procedures is they come with a built-in way of evaluating how wrong they are. If one is concerned that the prior may be too influential, one can consider other, competing priors and see how much, if at all, the answer changes. There are three great benefits to this way of thinking: (1) one attends to the continuous nature of probability, as opposed to focusing on yes/no decisions from the null hypothesis (Gelman, 2011, 2013); (2) one has a practical way to construct meaningful credible intervals for the PPoC; and (3) ultimately, one estimates the probability that is of actual interest, the probability that someone is a cheater, given the observed value on a statistic of interest.

As the final example in Table 18.4 demonstrates, the choice of a prior may be fairly influential on the estimate of the Bayesian posterior probability. This is often cited as a concern regarding the Bayesian paradigm, but it in fact allows the Bayesian analyst to treat the prior as just another source of information, not unlike the data themselves. Initially, there may be little evidence to suggest an obvious value for $P(C)$. An informed guess could be the starting point, for example, cheating is probably less prevalent than not cheating, so start with 0.5 or lower for $P(C)$. Furthermore, the Bayesian paradigm actually invites the analyst to consider a range of possible values for the prior probability, as was demonstrated in Table 18.4. Evaluating the influence of the prior provides insight into the nature of the posterior, as with the comparison of ranges in posterior probabilities from the $X_c = 2$ column in Table 18.4 with

the same range from the $X_c = 5$ column. With a relatively low threshold for detection, there is greater uncertainty in the PPoC as a function of the prior information, but as the threshold increases, the prior probability has less and less influence, because the likelihood dominates the equation. This enlightening information is provided by considering multiple priors, while the traditional *p*-value remains a constant for a given threshold.

One can not only evaluate a range of priors and examine their influence, one can also change these values as new information comes to light. For example, perhaps a series of investigations turns up evidence that certain flagged examinees were cheaters and others were not. If subsequent data analysis and experience demonstrate that the prior probability is surely somewhere between 0.01 and 0.10, for example, then posterior inferences can be that much more influenced by this information. Priors can be updated accordingly and posteriors may be recalculated to make stronger inferences. Traditional null hypothesis testing does not consider the distribution of the actual parameter of interest (in this case, the prevalence of cheating), so it cannot benefit from such updated information.

Although the context of this discussion is cheating detection, the arguments contained herein really apply to all statistical diagnostic decision-making practices, particularly those dealing with low-incidence phenomena. Moreover, the reasons for adopting a Bayesian approach in a wide variety of parameter estimation contexts are merely extensions of this line of reasoning. In recent years, Bayesian statistics have enjoyed a renaissance, as pragmatists have seized upon techniques like Markov Chain Monte Carlo as a means to estimate parameters when traditional maximum likelihood might not work. However, many of these analysts nonetheless limit the role of the prior to the extent possible. In the words of L. J. Savage (1961), they attempt to "make the Bayesian omelet without breaking the Bayesian eggs" (p. 575). We go a step further to say that Bayesian methods are both theoretically sound and of practical advantage because they incorporate prior information. We say: "Break the eggs, make the omelet, and use all the information available to make as informed an inference as possible." In closing, we began with a quote from Bertrand Russell, and will conclude with another: "In all affairs it's a healthy thing now and then to hang a question mark on the things you have long taken for granted" (unknown source). We suggest that it is time to hang a question mark on null hypothesis testing and our dogged reliance on estimating the wrong probability.

References

Angoff, W. H. (1974). The development of statistical indices for detecting cheaters. *Journal of the American Statistical Association, 69,* 44–49.

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association, 71,* 791–799.

Cannell, J. J. (1989). *How public educators cheat on standardized achievement tests: The "Lake Wobegon" report.* Albuquerque, NM: Friends for Education.

Cizek, G. (1999). *Cheating on tests: How to do it, detect it, and prevent it.* Mahwah, NJ: Lawrence Erlbaum Associates.

Gelman, A. (2011). Induction and deduction in Bayesian data analysis. *Rationality, Markets, and Morals, 2,* 67–68.

Gelman, A. (2013). *P* values and statistical practice. *Epidemiology, 24*(1), 69–72.

Jeffreys, H. (1960). *Theory of probability* (3rd Ed.). Oxford: Clarendon.

Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review of recent developments. *Applied Measurement in Education, 8,* 261–272.

Savage, J. L. (1961). The foundations of statistical inference reconsidered. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, 575–586.

Savage, S., & Wainer, H. (2008). Until proven guilty: False positives and the war on terror. *Chance, 21*(1), 59–62.

Thiessen, B. (2007). Case study—Policies to address educator cheating. Retrieved from: http://homepage.mac.com/bradthiessen/pubs/format.pdf

Wainer, H. (2011). How should we screen for breast cancer? *Significance, 8*(1), 28–30.

Wollack, J. A. (1997). A nominal response model approach to detect answer copying. *Applied Psychological Measurement, 21,* 307–320.

Wollack, J. A. (2003). Comparison of answer copying indices with real data. *Journal of Educational Measurement, 40,* 189–205.