Dichotomous decisions based on dichotomously scored items: a case study

by G. J. MELLENBERGH,* H. KOPPELAAR** and W. J. VAN DER LINDEN***

Summary In a course in elementary statistics for psychology students using criterion-referenced achievement tests, the total test score, based on dichotomously scored items, was used for classifying students into those who passed and those who failed. The score on a test is considered as depending on a latent variable; it is assumed that the students can be dichotomized into the categories "mastery" (with scores on the latent variable above a cutting score), and "no mastery" (with scores below the cutting score on the latent variable). Two problems are considered: (a) How many students are classified incorrectly? Using the binomial error model a procedure is described for computing the classification proportions: p(mastery, passed), p(mastery, failed), p(no mastery, passed), and p(no mastery, failed). (b) What is the optimal cutting score on a test? Using a loss function a procedure for computing the optimal curring score is described.

1 Introduction

In 1973 the Department of Psychology of the State University at Utrecht started a mastery learning course in elementary statistics. The items of the multiple choice achievement tests used were scored 0 (wrong answer) or 1 (correct answer); the total score on a test was the unweighted sum of the item scores.

It was assumed that the observed score on a test depends on a latent variable that represents the degree of mastery of the tested subject matter. All tests were criterion-referenced implying that a cutting score on the latent variable is fixed in advance; this cutting score dichotomizes the latent variable into the categories "mastery", and "no mastery". The teachers chose a test cutting score in advance; this score represented their informed opinion of the cutting score on the latent variable. Students above the cutting score on the test pass the test and students below fail. The situation can be represented in a twofold table like Table 1.

Table 1

	latent variable	
decision	no mastery	mastery
pass	p_{01}	p_{11}
fail	p_{00}	p_{10}

Proportions (Mis)classifications with Mastery Decisions

The cell entries are proportions: p_{01} and p_{10} are the proportions of misclassifications and p_{11} and p_{00} the proportions of correct classifications. There were two ques-

* University of Amsterdam.
** State University of Utrecht.
*** Twente University of Technology.

tions on the use of the tests. How many students were classified incorrectly? Which cutting score on a test is optimal?

In this article we report a procedure for computing the classification proportions and an optimal cutting score on a test. The procedure is applied to the tests of the mastery learning course. The results for one test are described in detail; the results for other tests are summarized.

2 The Beta-Binomial Model

The following notation will be used:

n = number of items in the test;
x_i = observed total score of subject i ($x_i = 0, 1, \ldots, n$);
τ_i = true proportion of items of the domain that subject i can answer correctly ($0 \leqslant \tau_i \leqslant 1$);
$g(\tau)$ = probability density of τ;
$h(x)$ = probability density of x;
$f(x|\tau)$ = probability density of x, given τ;
$k(x, \tau)$ = joint probability density of x and τ;
c = cutting scores on the test ($c = 1, \ldots, n$);
d = cutting score on the latent variable ($0 \leqslant d \leqslant 1$).

It is assumed that the achievement tests can be considered as random samples drawn from a large domain with independent items of the same content and difficulty. For the process of a fixed student answering the n items of such a test a well-known model is the sequence of n Bernoulli trials. There are indications in the literature that this provides adequate description, even when there is some slight violation of the two basic assumptions of so-called local independence (local, i.e. given the ability of the student) and constant success probability over items within one student. The conditional probability density of observed total score x, given the ability τ (which is the unknown proportion of items in the domain that this fixed student can answer correctly) is now the *binomial* density:

$$f(x|\tau) = \binom{n}{x} \tau^x (1-\tau)^{n-x} \tag{1}$$

It remains to specify the probability density of the ability τ in the population of all students. Usually, in Bayesian statistics, the *beta* density is chosen as the natural conjugate for the binomial model; its flexible form nearly always makes an approximation of prior beliefs possible (NOVICK and JACKSON, [11, pp. 107–113]). Here we do not consider it as a prior density, but we remark that the flexibility is equally useful for characterizing the distribution of τ over the given population of students. Therefore, it is assumed that the distribution of τ follows a beta density:

$$g(\tau) = \mathrm{B}^{-1}(a, b-n+1)\tau^{a-1}(1-\tau)^{b-n} \tag{2}$$

where

$$B(a, b-n+1) = \int_0^1 y^{a-1}(1-y)^{b-n}\,dy = \Gamma(a)\Gamma(b-n+1)/\Gamma(a+b-n+1) \quad (3)$$

and $a > 0$, $b > n-1$ are given real numbers.

Multiplying (1) and (2) and integrating out τ yields the density of observed total score x, which is known as the *Pólya* distribution P'_0 (BOSCH, [1]):

$$h(x) = \binom{n}{x} B(a+x, b-x+1)B^{-1}(a, b-n+1) = P'_0(x; a, b-n+1, n) \quad (4)$$

LORD ([8], pp. 3–7) has found that this *Pólya* density yields a satisfactory fit with several sets of test scores obtained from large-scale test administrations.

The parameters a and b can be expressed as functions of the mean and variance of this distribution (LORD and NOVICK, [9, p. 517]):

$$a = (-1+\alpha_{21}^{-1})\mu_x \quad (5)$$

$$b = -a-1+n\alpha_{21}^{-1} \quad (6)$$

In these formulas α_{21} is the well-known formula 21 of KUDER and RICHARDSON:

$$\alpha_{21} = n(n-1)^{-1}\{1-n^{-1}(\sigma_x^2)^{-1}\mu_x(n-\mu_x)\} \quad (7)$$

3 Classification Proportions and an Optimal Decision Rule

First, the actual situation of the case study is considered: the cutting score on the test is fixed at c. The educational objective is that the true score of a student is higher than a fixed value d; therefore, the cutting score on the true score is fixed at d. For this case formulas for the classification proportions are developed. After applying the incomplete beta function I, defined as usual:

$$I_z(a, b) = B^{-1}(a, b)\int_0^z y^{a-1}(1-y)^{b-1}\,dy \quad (8)$$

where $a > 0$, $b > 0$, with the property $I_z(a, b) = 1 - I_{1-z}(b, a)$, partial integration yields

$$\sum_{x=c}^{n} f(x|z) = I_z(c, n-c+1)$$

where $f(x|\tau)$ is given by (1). Then from (1), (2) and (8) it follows directly that

$$p_{00} = \sum_{x=0}^{c-1}\int_0^d k(x, \tau)\,d\tau = \int_0^d g(\tau)I_{1-\tau}(n-c+1, c)\,d\tau =$$

$$\sum_{x=0}^{c-1} P'_0(x; a, b-n+1, n)I_d(a+x, b-x+1) \quad (9)$$

$$p_{01} = \sum_{x=c}^{n} \int_{0}^{d} k(x,\tau)\,d\tau = \int_{0}^{d} g(\tau) I_{\tau}(c, n-c+1)\,d\tau =$$

$$\sum_{x=c}^{n} P'_0(x; a, b-n+1, n) I_d(a+x, b-x+1) \tag{10}$$

$$p_{10} = \sum_{x=0}^{c-1} \int_{d}^{1} k(x,\tau)\,d\tau = \int_{d}^{1} g(\tau) I_{1-\tau}(n-c+1, c)\,d\tau =$$

$$\sum_{x=0}^{c-1} P'_0(x; a, b-n+1, n)(1 - I_d(a+x, b-x+1)) \tag{11}$$

$$p_{11} = \sum_{x=c}^{n} \int_{d}^{1} k(x,\tau)\,d\tau = \int_{d}^{1} g(\tau) I_{\tau}(c, n-c+1)\,d\tau =$$

$$\sum_{x=c}^{n} P'_0(x; a, b-n+1, n)(1 - I_d(a+x, b-x+1)) \tag{12}$$

where

$$P'_0(x; a, b-n+1, n) = \binom{n}{x} B(a+x, b-x+1)/B(a, b-n+1)$$

is the *Pólya* distribution.

Second, the question is posed which cutting score on the test is optimal for a fixed cutting score d on the true score. Therefore, the following loss function is introduced:

$$\begin{aligned} &L\ (\textit{suitable, accepted}) &&= w_{11} \\ &L\ (\textit{suitable, not accepted}) &&= w_{10} \\ &L\ (\textit{not suitable, accepted}) &&= w_{01} \\ &L\ (\textit{not suitable, not accepted}) &&= w_{00} \end{aligned} \tag{13}$$

A special case of this loss function is for instance: $w_{11} = w_{00} = 0$, $w_{10} = w_{01} = w$, the loss function sometimes used in testing statistical hypotheses (FERGUSON, [3, p. 199]). The risk is the expected loss:

$$R = EL = \sum_{i=0}^{1} \sum_{j=0}^{1} w_{ij} p_{ij} \tag{14}$$

An optimal cutting score on the test is the value of c that minimizes the risk.

4 Computations

LORD and NOVICK [9, p. 517] recommend to substitute in formulas (5), (6) and (7), the mean and standard deviation of the observed scores to estimate the parameters a and b of formula (2). Using these estimates, $\hat{a}$ and $\hat{b}$, and formula (4), it is easy to estimate the theoretical frequency distribution of x (LORD and NOVICK, [9,

p. 518]). If the model fits the data, the observed frequency distribution should fit the estimated theoretical frequency distribution. Then, using formulas (9) up to and including (12) and an algorithm for the incomplete beta function, the classification proportions can be computed; the program is described by KOPPELAAR, VAN DER LINDEN and MELLENBERGH [6].

5 Results

The procedure for computing the classification proportions is described for the first achievement test of the course in elementary statistics. The test was composed of 19 three-choice items and was administered to 184 sophomores majoring in psychology. The cutting score was fixed at 15: students passed the tests if they answered 15 or more items correctly. The estimates of the parameters a and b were: $\hat{a} = 14.47$ and $\hat{b} = 21.31$. Using these estimates the theoretical frequency was computed and compared with the observed frequency distribution (Table 2).

Table 2. Observed and Theoretical Distribution of Scores

	observed		negative hypergeometric distribution ($\hat{a} = 14.47$, $\hat{b} = 21.31$)	
score	frequency	proportion	frequency	proportion
0	0	0	0.00	0.0000
1	0	0	0.00	0.0000
2	0	0	0.00	0.0000
3	0	0	0.00	0.0000
4	0	0	0.02	0.0001
5	0	0	0.06	0.0003
6	0	0	0.15	0.0008
7	0	0	0.37	0.0020
8	3	0.0163	0.83	0.0045
9	1	0.0054	1.69	0.0092
10	3	0.0163	3.24	0.0176
11	5	0.0271	5.72	0.0311
12	10	0.0543	9.44	0.0513
13	15	0.0815	14.44	0.0785
14	15	0.0815	20.46	0.1112
15	27	0.1467	26.55	0.1443
16	35	0.1902	31.00	0.1685
17	32	0.1739	31.41	0.1707
18	29	0.1576	25.48	0.1385
19	9	0.0489	13.16	0.0715

An inspection of the table shows that the differences between the frequency distributions were small. To quantify this impression, scores equal to or less than 10 were treated as one class. The following computations were done. The mean of the absolute differences between the observed and theoretical proportions was 0.011, and the value of chi-square was 4.00 with 7 (10 − 1 − numbers of estimated parameters)

degrees of freedom; the right tail probability of this value was greater than 0.77. These computations confirm the hypothesis that the observed frequency distribution fits the theoretical distribution rather well.

The teachers of the course in elementary statistics considered a student as having mastered the subject matter if he could answer correctly at least 80% of the total domain of items. Therefore, d was fixed at 0.80. Using this value of d the classification proportions were computed for all possible values of the cutting score on the test (Table 3). From this table one finds the classification proportions for the actually used cutting score on the test ($c = 15$): $\hat{p}_{00} = 0.236$, $\hat{p}_{01} = 0.156$, $\hat{p}_{10} = 0.070$, and $\hat{p}_{11} = 0.537$.

Table 3. Classification proportions for all possible cutting scores on the test

cutting score on the test (c)	classification propertions $\hat{p}_{00}$	$\hat{p}_{01}$	$\hat{p}_{10}$	$\hat{p}_{11}$
1	0.000	0.392	0.000	0.608
2	0.000	0.392	0.000	0.608
3	0.000	0.392	0.000	0.608
4	0.000	0.392	0.000	0.608
5	0.000	0.392	0.000	0.608
6	0.000	0.392	0.000	0.608
7	0.001	0.391	0.000	0.608
8	0.003	0.389	0.000	0.608
9	0.008	0.385	0.000	0.608
10	0.017	0.376	0.000	0.608
11	0.034	0.359	0.001	0.607
12	0.063	0.330	0.003	0.605
13	0.107	0.285	0.010	0.598
14	0.167	0.225	0.028	0.579
15	0.236	0.156	0.070	0.537
16	0.302	0.090	0.148	0.459
17	0.353	0.040	0.267	0.341
18	0.380	0.012	0.410	0.198
19	0.391	0.002	0.538	0.070

Table 4. Optimal cutting scores on the test and estimated risks for different loss functions

loss function	optimal cutting score on the test (c)	estimated risk
$w_{11} = w_{00} = 0,\ w_{10} = 3,\ w_{01} = 1$	14	0.309
$w_{11} = w_{00} = 0,\ w_{10} = 2,\ w_{01} = 1$	14	0.281
$w_{11} = w_{00} = 0,\ w_{10} = w_{01} = 1$	15	0.226
$w_{11} = w_{00} = 0,\ w_{10} = 1,\ w_{01} = 2$	16	0.328
$w_{11} = w_{00} = 0,\ w_{10} = 1,\ w_{01} = 3$	17	0.387

Furthermore, Table 3 was used for computing the optimal cutting scores on the test for the different loss functions reported in Table 4. It can be concluded that the optimal cutting score is 15 in case the loss associated with a passed student without mastery is equal to the loss associated with a failed student with mastery. The effect of raising the cutting score is to give a higher loss to passed students without mastery

compared with failed students with mastery. The effect of lowering the cutting score is the opposite: to give a higher loss to failed students with mastery.

DE BRUYNE [2, p. 97] tested the model for seven other tests used in the course. For five of these tests the fit of the data to the model was acceptable. For these tests the classification proportions were computed. Using the loss function $w_{11} = w_{00} = 0$, $w_{10} = w_{01} = 1$ DE BRUYNE [2, p. 99] also computed the optimal cutting scores. The results are reported in Table 5.

Table 5. Results for seven tests (DE BRUYNE, 1976)

	test						
	A	B	C	D	E	F	G
number of students	127	106	163	147	150	167	153
number of items	18	20	20	19	20	20	20
cutting score on the test	14	16	16	15	16	16	16
chi-square	4.474	2.832	7.888	4.942	10.640	24.971	44.405
degrees of freedom	6	7	10	8	7	7	9
right tail probability	0.61	0.90	0.64	0.76	0.16	0.000	0.000
mean absolute difference	0.0196	0.0141	0.0148	0.0153	0.0172	0.0322	0.0350
$\hat{p}_{00}$	0.410	0.526	0.914	0.607	0.418	–	–
$\hat{p}_{01}$	0.217	0.170	0.067	0.202	0.138	–	–
$\hat{p}_{10}$	0.043	0.052	0.005	0.033	0.057	–	–
$\hat{p}_{11}$	0.330	0.252	0.015	0.158	0.387	–	–
optimal cutting score	15	17	20	17	17		

The table shows that for these tests the estimated total proportions of misclassifications ($\hat{p}_{01} + \hat{p}_{10}$) have values between 0.072 (test C) and 0.260 (test A). These values are rather high and the tests can be improved for making pass-fail decisions; for instance by using a larger sample from the domain of test items. Moreover, all optimal cutting scores on these five tests were higher than the cutting scores the teachers chose. This indicates that the teachers implicitly used a loss function giving higher loss to failed students with mastery than to passed students without mastery.

It is remarked that it is not clear why the tests F and G do not fit the model. The content of these tests is of the same type as the other tests. However, the frequency distribution of test G has a larger variance and is more negatively skewed than the other tests; the frequency distribution of test F is not unimodal.

6 Discussion

In this article the optimal cutting score on the test is determined by minimizing the risk or expected loss. However, other approaches are possible. KLAUER [5], MILLMAN [10], FHANÉR [4], and WILCOX [13], concerned with the problem of determining optimal test length, have proposed classical testing of the hypothesis $H_0: \tau \geqslant d$ against the alternative hypothesis $H_1: \tau < d$. According to this approach the optimal cutting

score on the test is the critical value c^* that splits the total set of possible test scores into an acceptance region $x \geqslant c^*$ (students pass) and a rejection region $x < c^*$ (students fail). It is determined by specifying a maximum value α for the probability of a type I error (students with $\tau \geqslant$ d fail) and requiring the smallest possible value for the probability of a type II error (students with $\tau <$ d pass) in case a type I error is considered as most serious. The opposite can be done, if a type II error is considered as most serious. The probability model involved in this approach is the binomial distribution for τ fixed at d: $f(x|\tau = d)$; NOVICK and LEWIS [12] have advocated a Bayesian version of this approach.

According to our opinion a flaw of the above outlined method is that it is exclusively based on the subpopulation of students with values of τ in the neighboorhood of $\tau = d$ and does not take into account the entire population of students for whom the decision is made. Furthermore, using this method it is impossible to specify the loss for combinations of possible values of τ and the decisions. In the study presented here, the principle of minimizing the risk was chosen and this principle takes into account the entire distribution of values of τ for the given population of students. Moreover, loss function (13) was specified for combinations of values of τ and the decisions. It is also possible to specify other loss functions; for example, VAN DER LINDEN and MELLENBERGH [7] used a linear loss function based on the difference between τ and the cutting score; this yielded an analytical solution for the optimal cutting score on the test.

It is important to note the following feature of a method for determining an optimal cutting score, which draws upon the total distribution of τ values for a given population of students. Suppose an optimal cutting score is determined for a population of which nearly all have a mastery level below the cutting score d and just a few have a mastery level above d. Loss function (13) is chosen with $w_{00} = w_{11} = 0$ and $w_{01} = w_{10} = 1$. Therefore, $p_{01} = \text{Prob}\{\tau < d, x \geqslant c\}$ and $p_{10} = \text{Prob}\{\tau \geqslant d, x < c\}$ will not differ much from Prob $\{x \geqslant c\}$, respectively 0, and the risk $p_{01} + p_{10}$ will be minimal for a very large value of c. As a consequence some of the few students with a value of τ above d and an expected test score above $x =$ dn will fail the test just because other students have a low level of mastering the subject matter. From an individual point of view this is a serious drawback of the method used in the present study: a student with a mastery level above the required level d should have a fair chance of passing the test, irrespective of the mastery levels of the other students From the point of view of the institute that organizes the educational program, however, the procedure is correct: the decision is not only made for a few students with mastery level above d, but for the entire population including the large part of low achieving students. Therefore, two nonzero contributions to the risk are considered. The contribution of the small part of the population with $\tau \geqslant d$ equals the small probability p_{10}, which will increase for larger values of c. The contribution of the large part of the population with $\tau < d$ equals the large probability p_{01}, which will decrease for larger values of c. Accordingly, it is for a very large value of c that the risk, which is the sum of both contributions, will be minimal.

Acknowledgements

R. GILL, F. N. KERLINGER, W. MOLENAAR, and W. SCHAAFSMA for their valuable comments.

References

[1] BOSCH, A. J., The Pólya distribution. Statistica Neerlandica, 1963, 17, 201–213.
[2] BRUYNE, H. C. D. DE, Blokken in het onderwijs. Groningen: Tjeenk Willink, 1976.
[3] FERGUSON, T. S., Mathematical statistics: a decision theoretic approach. New York: Academic Press, 1967.
[4] FHANÉR, S., Item sampling and decision-making in achievement testing. British Journal of Mathematical and Statistical Psychology, 1974, 27, 172–175.
[5] KLAUER, K. J., Zur Theorie und Praxis des binomialen Modells lehrzielorientierter Tests. In K. J. Klauer, R. Fricke, M. Herbig, H. Rupprecht & F. Schott, Lehrzielorientierter Tests: Beiträge zur Theorie, Konstruktion und Anwendung, Düsseldorf: Pädagogischer Verlag Schwann, 1972.
[6] KOPPELAAR, H., W. J. VAN DER LINDEN and G. J. MELLENBERGH, A computerprogram for classification proportions in dichotomous decisions based on dichotomously scored items. Tijdschrift voor Onderwijsresearch, 1977, 1, 32–37.
[7] VAN DER LINDEN, W. J. and G. J. MELLENBERGH, Optimal cutting scores using a linear loss function. Applied Psychological Measurement, 1977, 1, (in press).
[8] LORD, F. M., The negative hypergeometric distribution with practical application to mental test scores. Princeton: Educational Testing Service, 1960.
[9] LORD, F. M. and M. R. NOVICK, Statistical theories of mental test scores. Reading, mass.: Addison-Wesley, 1968.
[10] MILLMAN, J., Passing scores and test lengths for domain-referenced measures. Review of Educational Research, 1973, 43, 205–216.
[11] NOVICK, M. R. and P. H. JACKSON, Statistical methods for educational and psychological research. New York: McGraw-Hill, 1974.
[12] NOVICK, M. R. and C. LEWIS, Prescribing test length for criterion-referenced measurement. In C. W. Harris, M. C. Alkin and W. J. Popham, Problems in criterion-referenced measurement (CSE monograph series in evaluation, No. 3). Los Angeles: Center for the Study of Evaluation, University of California, 1974.
[13] WILCOX, R. R., A note on the length and passing score of a mastery test. Journal of Educational Statistics, 1976, 1, 359–364.