Eva Tesaříková

Bayesian estimation of number of the questions unfamiliar with the examinee in variant I of probabilistic model of double choice test

Acta Universitatis Palackianae Olomucensis. Facultas Rerum Naturalium. Mathematica, Vol. 34 (1995), No. 1, 167--174

Persistent URL: http://dml.cz/dmlcz/120326

Acta Univ. Palacki. Olomuc., Fac. rer. nat.,
Mathematica **34** (1995) 167–174

Bayesian Estimation of Number of the Questions Unfamiliar with the Examinee in Variant I of Probabilistic Model of Double Choice Test

Eva TESAŘÍKOVÁ

Department of Algebra and Geometry, Faculty of Science, Palacký University, Tomkova 40, 779 00 Olomouc, Czech Republic
E-mail: tesariko@risc.upol.cz

(Received December 13, 1994)

Abstract

In connection with the analysis of achievement tests with double choice response it is possible to construct two variants of mathematical models describing the probability structure of such tests. Variant I published in 1992 proceeds from the assumption that when the tested person is really familiar with the topic of the question, he or she will select both the correct responses among the offered alternatives. Then on the basis of a single application of the test it is posible to derive Bayesian estimation of the number of the questions with which the examinee is really familiar.

Key words: school-achievement test, double-choice response, probabilistic model, Bayesian estimation of the number of the questions with which the examinee is really unfamiliar.

MS Classification: 62P10, 62P15

1 Introduction

The construction of mathematical models describing the probability structure of school-achievement tests should take into account a certain difference between the real knowledge of the examinee and the result of the test evaluated by the examiner.

This paper follows the study summarized in [1], [2], [3] and [4] where the probabilistic models of the school-achievement test with double choice response and their statistical analysis were given including certain simplifying assumptions. We consider the school-achievement test with compulsory choice of two correct responses from $q > 3$ offerred alternatives two of which are correct. A missing answer is evaluated as an incorrect one. The examinee is informed about the nature of the test before testing. It is also presumed that the test consists of n independent question of the same difficulty, the number of offered alternatives is the same in all the questions and the role of the alternatives is equivalent. Alternative responses to each question should be choosen in order to avoid similarity and discrepancy.

Variant I of the probabilistic model of the school-achievement test with double choice response published in [1] proceeds from the assumption that when the person under examination is really familiar with the topic of the question he or she will select both the correct responses from the offered alternatives.

This variant of probabilistic model suggests the expression of multinomic distribution of the random vector $M = (M_0, M_1, M_2)$ in the form

$$P(M_0 = m_0, M_1 = m_1, M_2 = m_2) = \tag{1}$$
$$= \frac{n!}{m_0!m_1!m_2!}\left(\tau\frac{(q-2)(q-3)}{q(q-1)}\right)^{m_0}\left(\tau\frac{4(q-2)}{q(q-1)}\right)^{m_1}\left(1-\tau\frac{q(q-1)-2}{q(q-1)}\right)^{m_2}$$

with $m_0 + m_1 + m_2 = n$, where M_0, M_1, M_2 is a random variable expressing the number of questions of the test to which no correct response was given, only one correct response was given, both the correct responses were given by the examinee, respectively.

The distribution of the random vector $M = (M_0, M_1, M_2)$ of the registrated results of the test depends on one parameter τ which represents the proportion of the tested topic with which the examinee is unfamiliar.

The maximal likelihood estimator of the parameter τ in the form

$$\hat{\tau} = \begin{cases} \dfrac{m_0+m_1}{n}\dfrac{q(q-1)}{q(q-1)-2} & \text{for } m_0+m_1 \le n\dfrac{q(q-1)-2}{q(q-1)} \\ 1 & \text{for } m_0+m_1 > n\dfrac{q(q-1)-2}{q(q-1)} \end{cases}$$

was given in [2].

2 Bayesian estimation of the number of questions with which the examinee is unfamiliar

The aim of this paper is to derive the Bayesian estimation of the number of the questions with which the examinee is really unfamiliar in case of the above described variant I of the probabilistic model.

In the construction of the probabilistic model we use the following notation for random events relative to every question of the test:

Z the examinee is familiar with the topic of the question

N the examinee is unfamiliar with the topic of the question.

According the above mentioned assumptions the random events Z, N have probabilities $P(Z) = 1 - \tau$, $P(N) = \tau$.

In relation to the registered results of the test the following three random events are considered

S_0 no correct answer was given to the question

S_1 only one correct answer was given to the question

S_2 both correct answers were given to the question

In case of familiarity with the given topic the examinee will give the two correct responses, so that

$$P(S_0|Z_i) = 0, \quad P(S_1|Z_i) = 0, \quad P(S_2|Z_i) = 1,$$

In case of unfamiliarity with the topic tested by the question the examinee can only use the random choice. With regard to the presence of two correct responses among $q > 3$ offered alternatives the following relations hold:

$$\begin{aligned} P(S_0|N) &= \frac{\binom{q-2}{2}}{\binom{q}{2}} = \frac{(q-2)(q-3)}{q(q-1)} \\ P(S_1|N) &= \frac{\binom{q-2}{1}\binom{2}{1}}{\binom{q}{2}} = \frac{4(q-2)}{q(q-1)} \\ P(S_2|N) &= \frac{\binom{2}{2}}{\binom{q}{2}} = \frac{2}{q(q-1)} \end{aligned}$$

These conditional probabilities were calculated by the hypergeometric distribution which is applicable in this situation.

Unconditional probabilities of evets S_0, S_1, S_2 were calculated according to the theorem of total probability:

$$\begin{aligned} P(S_0) &= P(Z)P(S_0|Z) + P(N)P(S_0|N) = \tau\frac{(q-2)(q-3)}{q(q-1)} \\ P(S_1) &= P(Z)P(S_1|Z) + P(N)P(S_1|N) = \tau\frac{4(q-2)}{q(q-1)} \\ P(S_2) &= P(Z)P(S_2|Z) + P(N)P(S_2|N) = 1 - \tau\left(1 - \frac{2}{q(q-1)}\right) \end{aligned}$$

From these probabilities was derived the multinomial distribution (1) of the random vektor M.

Now, we are interested in the probability distribution of the random vector $W = (Y, U)$, where Y represents the number of questions of the test with which the exaninee was really unfamiliar and U represents the number of questions with which the examinee was really familiar.

The unconditional distribution of the random variable Y is a binomial probability distribution dependent on the parameter τ in the form

$$P(Y = y) = \binom{n}{y} \tau^y (1-\tau)^{n-y} \tag{2}$$

This formula is equivalent to the formula

$$P(Y = y, U = u) = \frac{n!}{y!u!} \tau^y (1-\tau)^u \qquad \text{for } y + u = n$$

of the multinomial distribution of the random vector W.

By the Bayesian theorem we can derive aposteriori conditional probability distribution of the variable Y in dependence on a single application of the test regisrated by the examiner, i.e. on the values of the random vector M.

We have

$$P(Y = y | M = m) = \frac{P(Y = y) P(M = m | Y = y)}{P(M = m)} = \tag{3}$$

$$= \frac{\binom{n}{y} \tau^y (1-\tau)^{n-y} \frac{y!}{m_0! m_1! (y-m_0-m_1)!} \left(\frac{(q-2)(q-3)}{q(q-1)}\right)^{m_0} \left(\frac{4(q-2)}{q(q-1)}\right)^{m_1} \left(\frac{2}{q(q-1)}\right)^{y-m_0-m_1}}{\frac{n!}{m_0! m_1! (y-m_0-m_1)!} \left(\tau \frac{(q-2)(q-3)}{q(q-1)}\right)^{m_0} \left(\tau \frac{4(q-2)}{q(q-1)}\right)^{m_1} \left(1 - \tau \frac{2}{q(q-1)}\right)^{y-m_0-m_1}} =$$

$$= \frac{\frac{(n-m_0-m_1)!}{(n-y)!(y-m_0-m_1)!} (1-\tau)^{n-y} \left(\frac{2\tau}{q(q-1)}\right)^{y-m_0-m_1} \left(1 - \tau \frac{q(q-1)-2}{q(q-1)}\right)^{y-m_0-m_1}}{\left(1 - \tau \frac{q(q-1)-2}{q(q-1)}\right)^{n-m_0-m_1} \left(1 - \tau \frac{q(q-1)-2}{q(q-1)}\right)^{y-m_0-m_1}} =$$

$$= \binom{n-m_0-m_1}{y-m_0-m_1} \left(\frac{2\tau}{q(q-1) - \tau(q(q-1)-2)}\right)^{y-m_0-m_1} \left(\frac{(1-\tau) q(q-1)}{q(q-1) - \tau(q(q-1)-2)}\right)^{n-y} =$$

$$= \binom{n-m_0-m_1}{r} \left(\frac{2\tau}{q(q-1) - \tau(q(q-1)-2)}\right)^{r} \left(\frac{(1-\tau) q(q-1)}{q(q-1) - \tau(q(q-1)-2)}\right)^{n-m_0-m_1-r}$$

with $y = m_0 + m_1, m_0 + m_1 + 1, \dots, n$ or with $r = 0, 1, \dots, n - m_0 - m_1$, respectively, where r is a value of random variable R representing the number of the questions of the test with which the examinee was really unfamiliar and to which at the same time the both correckt responses were given only by the random choice. We can see, that the derived distribution of the variable R is a binomial probability distribution again. So, on the basis of formula (3) it is not difficult to find the Bayesian estimation of y in the form of a regression function of Y in dependence on the random vector M.

$$E(Y|M=m)= \tag{4}$$

$$=\sum_{y=m_0+m_1}^{n}\binom{n-m_0-m_1}{y-m_0-m_1}\left(\frac{2\tau}{q(q-1)-\tau(q(q-1)-2)}\right)^{y-m_0-m_1}$$

$$\left(\frac{(1-\tau)q(q-1)}{q(q-1)-\tau(q(q-1)-2)}\right)^{n-y}=$$

$$=\sum_{r=0}^{n-m_0-m_1}(m_0+m_1+r)\binom{n-m_0-m_1}{r}\left(\frac{2\tau}{q(q-1)-\tau(q(q-1)-2)}\right)^{r}$$

$$\left(\frac{(1-\tau)q(q-1)}{q(q-1)-\tau(q(q-1)-2)}\right)^{n-m_0-m_1-r}$$

$$=(m_0+m_1)+E(R)=(m_0+m_1)+(n-m_0-m_1)\frac{2\tau}{q(q-1)-\tau(q(q-1)-2)}=$$

$$=n\frac{2\tau}{q(q-1)-\tau(q(q-1)-2)}+(m_0+m_1)\frac{(1-\tau)q(q-1)}{q(q-1)-\tau(q(q-1)-2)}.$$

We can see, that regression (4) is a linear function of the sum m_0+m_1. In case of $m_0+m_1=0$ we can expect

$$E(Y|m_0+m_1=0)=n\frac{2\tau}{q(q-1)-\tau(q(q-1)-2)}$$

question of the test with which the examinee was really unfamiliar. In case of $m_0+m_1=n$ our expectation of the number of these questions will be

$$E(Y|m_0+m_1=n)=n.$$

It correspondents with the assumption in this variant of probabilistic model of the test.

In view of the fact that Bayesian estimation (4) depends on the values m_0+m_1 of the random sum M_0+M_1 we can express its expectation. According to the multinomial probability distribution (1) of M we get

$$E[E(Y|M)]= \tag{5}$$

$$=n\frac{2\tau}{q(q-1)-\tau(q(q-1)-2)}+(E(M_0+M_1))\frac{(1-\tau)q(q-1)}{q(q-1)-\tau(q(q-1)-2)}=$$

$$=\frac{1}{q(q-1)-\tau(q(q-1)-2)}(2n\tau+(EM_0+EM_1)(1-\tau)q(q-1))=$$

$$=\frac{1}{q(q-1)-\tau(q(q-1)-2)}\left(2n\tau+n\tau\frac{q(q-1)-2}{q(q-1)}(1-\tau)q(q-1)\right)=$$

$$=\frac{n\tau}{q(q-1)-\tau(q(q-1)-2)}(2+(q(q-1)-2)(1-\tau))=n\tau$$

Consequently, the expectation (5) of the Bayesian estimation (4) equals to the unconditional expectation of the variable Y.

Now, we will be interested in the variance of the Bayesian estimation (4). We have

$$D[E(Y|M)] = E[E(Y|M)] - E[E(Y|M)]^2 =$$

$$= E\Big[n\frac{2\tau}{q(q-1)-\tau(q(q-1)-2)} + + (M_0+M_1)\frac{(1-\tau)q(q-1)}{q(q-1)-\tau(q(q-1)-2)} - n\tau\Big]^2 =$$

$$= E\Big[n\tau\frac{2-q(q-1)+\tau(q(q-1)-2)}{q(q-1)-\tau(q(q-1)-2)} + + (M_0+M_1)\frac{(1-\tau)q(q-1)}{q(q-1)-\tau(q(q-1)-2)}\Big]^2 =$$

$$= \left(E(M_0+M_1)^2\right)\frac{[(1-\tau)q(q-1)]^2}{[(q-1)-\tau(q(q-1)-2)]^2} + 2n\tau\left(E(M_0+M_1)\right)\cdot$$

$$\cdot\frac{2-q(q-1)+\tau(q(q-1)-2)}{q(q-1)-\tau(q(q-1)-2)}\frac{(1-\tau)q(q-1)}{q(q-1)-\tau(q(q-1)-2)} + $$

$$+ n^2\tau^2\frac{[2-q(q-1)+\tau(q(q-1)-2)]^2}{[q(q-1)-\tau(q(q-1)-2)]^2}.$$

According the equality

$$E(M_0+M_1)^2 = D(M_0+M_1) + E^2(M_0+M_1) =$$

$$= n\tau\frac{q(q-1)-2}{q(q-1)}\left(1-\tau\frac{q(q-1)-2}{q(q-1)}\right) + \left(n\tau\frac{q(q-1)-2}{q(q-1)}\right)^2 =$$

$$= n\tau\frac{q(q-1)-2}{q(q-1)}\left(1-\tau\frac{q(q-1)-2}{q(q-1)} + n\tau\frac{q(q-1)-2}{q(q-1)}\right) =$$

$$= \left(n^2-n\right)\tau^2\left(\frac{q(q-1)-2}{q(q-1)}\right)^2 + n\tau\frac{q(q-1)-2}{q(q-1)}$$

which follows from the expression of expectation and variance of sum $M_0 + M_1$, we have

$$D[E(Y|M)] = \tag{6}$$

$$= \left(n^2-n\right)\tau^2\left(\frac{q(q-1)-2}{q(q-1)}\right)^2\frac{[(1-\tau)q(q-1)]^2}{[(q-1)-\tau(q(q-1)-2)]^2} +$$

$$+ n\tau\frac{q(q-1)-2}{q(q-1)}\frac{[(1-\tau)q(q-1)]^2}{[(q-1)-\tau(q(q-1)-2)]^2} +$$

$$+2n^2\tau^2\frac{q(q-1)-2}{q(q-1)}\frac{2-q(q-1)+\tau(q(q-1)-2)}{q(q-1)-\tau(q(q-1)-2)}\frac{(1-\tau)q(q-1)}{q(q-1)-\tau(q(q-1)-2)}+$$

$$+n^2\tau^2\frac{[2-q(q-1)+\tau(q(q-1)-2)]^2}{[q(q-1)-\tau(q(q-1)-2)]^2}=$$

$$=-n\tau^2\frac{[q(q-1)-2)]^2(1-\tau)^2}{[q(q-1)-\tau(q(q-1)-2)]^2}+n\tau\frac{[q(q-1)-2)](1-\tau)^2q(q-1)}{[q(q-1)-\tau(q(q-1)-2)]^2}=$$

$$=n\tau\frac{[q(q-1)-2)](1-\tau)^2}{[q(q-1)-\tau(q(q-1)-2)]^2}[q(q-1)-\tau(q(q-1)-2)]=$$

$$=n\tau(1-\tau)\frac{[q(q-1)-2)](1-\tau)}{q(q-1)-\tau(q(q-1)-2)}=$$

$$=n\tau(1-\tau)\left(1-\frac{2}{q(q-1)-\tau(q(q-1)-2)}\right)$$

The expression (6) represents the variability of the regression function $E(Y|M=m)$ around the unconditional expectation $E(Y)$. The whole variability $D(Y)$ is then posible divide into two components. One of them is the just calculated variance (6), the other one is the expectation $E[D(Y|M)]$ of the residual variance $D(Y|M=m)$. In view of the fact that the conditional distribution (3) is a distribution of binomial type we have

$$D(Y|M=m)=$$
$$=(n-m_0-m_1)\frac{2\tau}{q(q-1)-\tau(q(q-1)-2)}\frac{(1-\tau)q(q-1)}{q(q-1)-\tau(q(q-1)-2)}$$

The expectation of this variance is expressed in the form

$$E[D(Y|M)]= \tag{7}$$
$$=[n-E(M_0-M_1)]\frac{2\tau}{q(q-1)-\tau(q(q-1)-2)}\frac{(1-\tau)q(q-1)}{q(q-1)-\tau(q(q-1)-2)}=$$
$$=n\left(1-\tau\frac{q(q-1)-2}{q(q-1)}\right)\frac{2\tau(1-\tau)q(q-1)}{[q(q-1)-\tau(q(q-1)-2)]^2}=$$
$$=n\tau(1-\tau)\frac{2}{q(q-1)-\tau(q(q-1)-2)}$$

The unconditional variance $D(Y)$ is derived as a variance of binomial probability distribution (2) in the form $D(Y)=n\tau(1-\tau)$. This expression can be obtained by the addition both the components (6) and (7) of the whole variability of Y, too. As we can see, the proportion of the two components (6), (7) of the variance $D(Y)$ is independent on the number n of the questions in the test. It depends only on the number q of the offered alternatives and on the proportion τ of the whole topic of the test with which the examinee is really unfamiliar. The expectation (7) of the residual variance $D(Y|M)$ is a measure of the influence of random which is present in the possibility of random choice. This influence decreases with the increasing number q of offered alternatives.

References

[1] Tesaříková, E.: *Probabilistic model of school-achievement test with double-choice response—variant I.* Acta Univ. Palacki. Olomuc., Fac. rer. nat. **105**, Math. 31 (1992), 119–126.

[2] Tesaříková, E.: *Statistical analysis of probabilistic model—variant I of the school-achievement test with double choice response.* Acta Univ. Palacki. Olomuc., Fac. rer. nat. **105**, Math. 31 (1992), 126–147.

[3] Tesaříková, E.: *Variant II of probabilistic model of school-achievement test with double-choice response.* Acta Univ. Palacki. Olomuc., Fac. rer. nat. **110**, Math. 32 (1993), 159–164.

[4] Tesaříková, E.: *Statistical analysis of variant II of probabilistic model of the school-achievement test with double choice response.* Acta Univ. Palacki. Olomuc., Fac. rer. nat. **110**, Math. 32 (1993), 165–175.