Detecting Copying in MCQ based Test

SUBHRAJYOTY ROY
BS – 1613

Problem: There is an MCQ based test with N questions. Each of the question has k many choices of which only one answer is correct. An examinee may unattempt a question. Either the wrong answer or leaving a question unanswered would fetch the examinee 0 marks for that question. On the other hand, a correct answer for any question would fetch 1 marks. Devise a statistical strategy, to test whether two randomly selected examinee are copying from each other or not, stating the statistical model and your hypothesis clearly. You may assume that there are M examinees giving the test.

Answer:

Let us consider two examinees A and B.
One key observation is that, if A and B both are giving the correct answer to the same question, then it is not possible to tell whether they copied from each other or both knew the correct answer to the questions independently. Therefore, we must restrict our attention to only those questions where at least one of them answers wrongly or kept unanswered.
For now, let us assume that A is the source of the answer, and B is the examinee who is trying to copy answers from A. Then, we consider the following assumptions about the copier B;

- If the copier knows a correct response to a question, then he/she will give that correct response whether or not the source has given correct response.
- If the copier does not know the correct response and can copy from A, then he/she exactly copies the response of A.
- If the copier does not know the correct response and cannot copy from A or finds A have not answered it, then he/she guesses blindly among the alternatives or keeping unattempted.

Therefore, the above discussion tells us to restrict our attention only to those question for which the source A has given wrong answer, or kept unanswered.
Let b_i be the answer to the i-th question by the copier B, which takes a value from $0,1,2,\dots k$ each positive integer representing one of the alternatives and 0 representing unattempt, while a_i denotes the answer to the i-th question by the source A. Clearly, as A answers the question wrongly or unattempted, a_i can be anything in $\{0,1,\dots k\}$ except the correct alternative. Let; T_i be the indicator that both a_i and b_i are same, i.e.

$$T_i = \begin{cases} 1 & a_i = b_i \\ 0 & a_i \neq b_i \end{cases}$$

The three assumptions as mentioned before will lead us to the following distribution of T_i;

$$P(T_i = 1) = \begin{cases} 0 & \text{if B knows the answer} \\ (k+1)^{-1} & \text{if B guesses blindly} \\ 1 & \text{if B copies from A} \end{cases}$$

Now, we formally define our hypotheses as follows;

H_0 : The examinee B did not copy anything from A.
H_1 : The examinee B copied some of the answers from A.

Now we shall see how does these hypothesis affects the distribution of T_i. Let, among the total N questions, there are n_A questions where the source A answered wrongly or kept unanswered. Among these n_A questions,

correct response of α_{AB} questions were known to B, while β_{AB} questions were copied by B. Therefore, we have the distribution of T_i as;

$$P(T_i = 1) = \begin{cases} 0 & \text{for } \alpha_{AB} \text{ questions} \\ (k+1)^{-1} & \text{for } (n_A - \alpha_{AB} - \beta_{AB}) \text{ questions} \\ 1 & \text{for } \beta_{AB} \text{ questions} \end{cases}$$

With $\alpha_{AB} \geq 0, \beta_{AB} \geq 0$, and $n_A \geq \alpha_{Ab} + \beta_{AB}$, regardless of whether he/she has copied or not.
Now, it is evident that the null hypothesis and the alternative hypothesis can be regarded as;

$\mathbf{H_0 : \beta_{AB} = 0}$
$\mathbf{H_1 : \beta_{AB} > 0}$

Let us now consider the total number of matching questions, where both has given same wrong answer or kept same question unattempted, as $M = \sum_{i=1}^{n_A} T_i$. Now, observe that, if $m < \beta_{AB}$, then it means the copier has copied more answers than m, hence it is not possible to get m many matches. Also, in a similar way, as the copier knows the answer to α_{AB} questions, hence, it is also not possible to get more than $(n_A - \alpha_{AB})$ matches of wrong response. Therefore, the support of the random variable M is the set of integers, $S = \{\beta_{AB} + 1, \beta_{AB} + 2, \dots n_A - \alpha_{AB}\}$. Now, let us assume we have m matches. Then, the copier has exactly copied β_{AB} many answers, and for others, he/she matches it by pure chance, with a success probability $(k+1)^{-1}$ at each trial. The probability mass function of M is given by something closely related to binomial distribution;

$$P(M = m) = \binom{n_A - \alpha_{AB} - \beta_{AB}}{m - \beta_{AB}} \left(\frac{1}{k+1}\right)^{m-\beta_{AB}} \left(\frac{k}{k+1}\right)^{n_A - \alpha_{AB} - m} \times \mathbf{1}\{m \in S\}$$

Now, we cannot test the above mention hypothesis using the above distribution, unless we know α_{AB}, because, under null hypothesis, when $\beta_{AB} = 0$, then the distribution of M depends on α_{AB}.

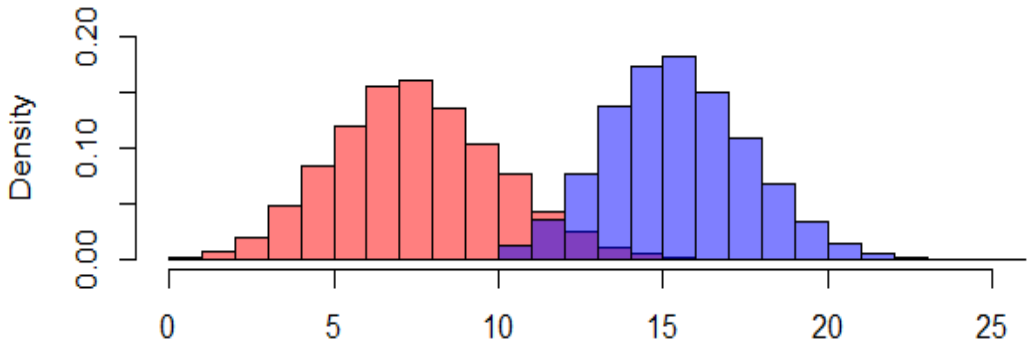
Observe that, this α_{AB} should "solely" depend on the ability of the copier. Hence, using other candidates' answers to those n_A questions would not be a reasonable way to estimate α_{AB}, as the distribution of the ability of answering questions are not identically distributed for each candidate, and it is NOT reasonable to assume it.

Therefore, we consider a different approach now. Observe that, if among n_A questions, B answered α_{AB} many correctly, then for those questions, B clearly has not copied anything, and hence, it would make sense to consider this $\alpha_{AB} = 0$. On a different note, since we do not know who is the source and who is the copier among A and B in a practical scenario, then clearly, as the questions correctly answered by A would not be considered when we assume A is the source, and in a similar manner, the questions correctly answered by B would not be considered when we assume B is the source, **we further need to restrict our attention to only those questions not correctly answered by any one of them.** Let the number questions of this type be n_{AB}.

Therefore, it is reasonable to assume $\alpha_{AB} = 0$, and then we would have; S under H_0, is the set of all integers from 1 to n_{AB}, while under H_1, the support S is the set of integers from $(\beta_{AB} + 1)$ to n_{AB}. Observe that, under both hypothesis, the random variable M follows a Binomial distribution on S with success probability $(k+1)^{-1}$.

The following diagram shows the histogram under null hypothesis and alternative hypothesis for a simulated scenario with $n_{AB} = 40$, $\beta_{AB} = 5$ and the presence of 4 alternatives to choose from for each question.

From the plot it is evident to see that we should reject H_0 when M, i.e. the number of matching is "too" large. To figure out the cutoff exactly, we consider the Type 1 error under the following testing rule:



Reject H_0 iff $M \geq c$ for some suitably chosen $c \in \{1, 2, \ldots n_A\}$

$$P(\text{Type I error}) = P(M \geq c | H_0) = \sum_{m=c}^{n_{AB}} \binom{n_{AB}}{m} \left(\frac{1}{k+1}\right)^m \left(\frac{k}{k+1}\right)^{n_{AB}-m}$$

Therefore, c is the smallest integer such that, the above sum is less than the given level of significance α, generally taken as 0.05. In case we want a test satisfying the level α exactly, we can achieve it using a randomization performed at $m = c$.

To compute the power of the test, or Type II error, we obtain the following;

$$\text{Power of the test at } \beta_{AB} = P(M \geq c | H_1) = \sum_{m=c}^{n_{AB}} \binom{n_{AB} - \beta_{AB}}{m - \beta_{AB}} \left(\frac{1}{k+1}\right)^{m-\beta_{AB}} \left(\frac{k}{k+1}\right)^{n_{AB}-m}$$

Then,

$$P(\text{Type II error}) = (1 - \text{Power of the test at } \beta_{AB})$$

Now, consider the power function of the test, observe that, the power function increases as a function of β_{AB}, and achieves 1 when $\beta_{AB} \geq c$. Therefore, it somewhat justifies why this testing rule is "good". Also note that, the power function increases with respect to n_A, the effective sample size we have in the problem. Note the following diagrams for the simulated power functions;