KM 12(1983)
pag 42-48

TESTING FOR INDEPENDENCE OF BINARY RESPONSES

By Vaclav Fidler and Andre de Jonge

Summary

In the context of experiments involving visual inspection of random dot patterns the problem of testing for independence of binary responses is considered. A flexible model for dependence between binary responses is proposed. Two tests, optimal under different versions of the model, are derived. These two tests turn out to involve the same computations as the Wilcoxon two sample test and the runs test respectively.

Medical Faculty, University of Groningen
Bloemsingel 10, 9712 KZ Groningen
telephone 050-114964

1. Introduction

Consider the following experiment conducted in the context of a study of processes involved in visual pattern recognition (DeJonge and Rashbass, 1982). An observer views a pair of static random dot patterns presented to him on two oscilloscope screens placed alongside each other. For each pair he points at the pattern consisting - according to his judgement - of more dots; he has been instructed to force his choice if necessary. The two patterns in a pair are generated independently from the same probability distribution. There are m different pattern pairs, possibly generated from different distributions. Each pattern pair is n times replicated, equally often in the left-right and in the right-left orientation. The n.m pattern pairs are ordered at random.

The patterns consist of several hunderds of dots so that a judgement cannot be a result of a conscious enumeration. The purpose of the experiment is to find out whether there exists a consistency in replicated assessment of the same pattern pair and to infer on differences in consistency for different pattern pairs. Obviously, an operational definition of consistency and a method for its evaluation is required - these are the problems dealt with in this paper. We restrict our attention to inference based on replicated assessments of a single pattern pair.

2. Model

Consider a specific pattern pair consisting of patterns A and B. The observer's assessment of this pair for the i-th time is denoted by X_i, $i=1,\ldots,n$. If the observer concludes that dots are more numerous on A then X_i equals 1, if he concludes that it is the pattern B which contains more dots then X_i equals 0. Note that the observations $X_1,\ldots,X_n$ are only a part of the experiment in which m different pattern pairs are included.

If the numbers of dots on the two patterns in the pair are nearly the same the observer's first assessment is likely to be a pure guess. If he recognizes some structure in this pattern pair his next assessment will be probably positively correlated with the first one. On the other hand, if the observer fails to recognize the pattern pair he

will presumably use the same guess mechanism again, independently of previous assessments.

To express the dependence of responses we propose the following model for the conditional probability of X_i given $X_1,\ldots,X_{i-1}$:

$$P(X_i \mid X_1,\ldots,X_{i-1}) = \frac{\exp\left[X_i(a+b.\ S_{i-1}) \right]}{1 + \exp(a+b.\ S_{i-1})}, \quad i=1,\ldots,n \qquad (1)$$

where a and b are model parameters and where the statistic S_{i-1} determines influence of $X_1,\ldots,X_{i-1}$ on the conditional distribution of X_i ,

$$S_i = \sum_{j=1}^{i} w_{ij}[X_j - 1/2\ ], \quad i \geq 1; \quad S_0 = 0.$$

If b=0 the X_i's are independent with the same distribution as X_1 , $P(X_1 =0 \mid b=0\ ) = 1/[\ 1 + \exp(a)\ ]$. If $b\neq 0$ the X_i's are dependent. If b>0 there exists a positive dependence between X_i's in the sense that $P(X_i=1 \mid S_{i-1}>0) > P(X_1=1)$ and $P(X_i=0 \mid S_{i-1}<0) > P(X_1=0)$.

The choice of weights $\{w_{ij}\}$ should be preferably based on theoretical considerations concerning the involved process of visual perception. We shall consider the following two possibilities:

$$w_{ij} = 1 \qquad (2)$$

and

$$w_{ij} = 1 \text{ if } i=j \text{ and } w_{ij} = 0 \text{ if } i\neq j\ . \qquad (3)$$

With weights given by (2) all observations prior to X_i are equally relevant, with weights (3) only the latest observation is employed.

The choice of values zero and one as the two possible outcomes of an assessment is arbitrary. It is therefore of interest to know what happens if the role of zero and one is interchanged. With

$$Y_i = 1 - X_i\ , \quad i=1,\ldots,n \qquad (4)$$

we find from (1)

$$P(Y_i \mid Y_1, \ldots, Y_{i-1}) = \frac{\exp\left[Y_i (-a+b\, S^*_{i-1}) \right]}{1 + \exp(-a+b\, S^*_{i-1})}$$

where $S^*_{i-1} = S(Y_1, \ldots, Y_{i-1})$. Thus the dependence parameter b is invariant under the transformation (4).

3. Inference

Given observations $X_1, \ldots, X_n$ distributed according to (1), with S_i given by (2) or by (3) we wish to test $H_0 : b=0$ against the one sided alternative $H_1 : b>0$. As

$$P(X_1, \ldots, X_n) = P(X_1)\, P(X_2 \mid X_1) \ldots P(X_n \mid X_1, \ldots, X_{n-1})$$

we obtain by substituting from (1)

$$\log[P(X_1, \ldots, X_n)] =$$

$$= a\, X + b \cdot \Sigma\, X_i S_{i-1} - \Sigma \log[1 + \exp(a + b\, S_{i-1})]$$

where $X = \Sigma\, X_i$. The statistic X is under H_0 sufficient for the nuisance parameter a. Thus for inference on b the standard theory - see for example Cox and Hinkley (1974) - suggests to consider the conditional distribution of $(X_1, \ldots, X_n)$ given X. This conditional distribution still involves the nuisance parameter a. Thus no uniformly most powerful test exists. It is however possible to look for a test with maximal power for alternatives close to H_0. Such a test - see Cox and Hinkley (1974) - rejects H_0 for large values of

$$\lim_{b \to 0} \partial \log[P(X_1, \ldots, X_n)] / \partial b$$

from which its critical region is derived as

$$\Sigma\, X_i . S_{i-1} - \Sigma \frac{e^a}{1 + e^a} . S_{i-1} \geq \text{const.} \qquad (5)$$

To derive specific tests we substitute the two choices of weights, (2) and (3), in (5). With weights w_{ij} given by (2) we have

$$\Sigma\, X_i S_{i-1} = (X^2 - T)/2$$

$$\Sigma\, S_{i-1} = X.n - n(n-1)/4 - T$$

$$T = \Sigma i X_i$$

and the critical region (5) can be written as

$$T \left(\frac{e^a}{1 + e^a} - \frac{1}{2} \right) \geq K(X),$$

where K(X) has to be determined so that the test is of a required size. If $a \neq 0$ it follows from invariance considerations - under transformation (4) - that H should be rejected if T is either too large or too small. This test is computationally equivalent to the two-sample Wilcoxon test. The vector $(X_1,\ldots,X_n)$ can be viewed as indicating sample membership of ordered observations from pooled samples consisting respectively of X and of n-X observations. It follows that T equals to the sum of the ranks from the sample labeled by ones.

If a = 0 the above test can still be used. However, in this case there is no need to condition on X; a locally most powerful test can be derived directly from $P(X_1,\ldots,X_n)$. This test turns out to reject H_0 for large values of X(X-n), that is for too small or too large values of X.

Next consider weights (3). We have

$$\Sigma X_i S_{i-1} = \Sigma X_i X_{i-1} - X/2 + X_1/2$$

$$\Sigma S_{i-1} = X - (n-1)/2 - X_n .$$

Substitution in (5) results in

$$\Sigma X_i X_{i-1} + \frac{1}{2} X_1 + \frac{e^a}{1+e^a} X_n \geq K(X)$$

and invariance considerations lead to use this critical region with a=0. Thus H_0 is to be rejected for too large values of

$$U = \Sigma X_i X_{i-1} + (X_1 + X_n)/2.$$

Here $\Sigma X_i X_{i-1}$ equals to the number of runs of ones. This test is equivalent to the runs test as described for example by Lehmann (1975). The runs test statistic equals to the number of runs of ones plus the number of runs of zeros,

$$\Sigma (1-X_i)(1-X_{i-1}) + \Sigma X_i X_{i-1} = 2U - 2X + (n-1)$$

The nulldistributions of the above derived test statistics are well known, they can be found for example in Lehmann's (1975) book.

Alternatively, the likelihood ratio test for H_0 could be derived. Such an approach is however less attractive as it requires numerical maximalization of the likelihood function.

4. Example

To illustrate the two tests derived in the preceding section we consider an example with n=10 repeated assessments of the same pattern in which X=4. The number of different vectors $(X_1,\ldots,X_n)$ is $\binom{10}{4}=210$. The vector of outcomes (00000 01111) is labeled as extreme by both tests while the vector (01010 01010) is not. The vector (00011 11000) is labeled as extreme by the runs test, not by the Wilcoxon test. The situation is reversed for the outcome (00010 10101) which is not extreme under the runs test while being relatively extreme under the Wilcoxon test.

5. Discussion and conclusion

The model (1) proposed for the experimental design of Section 1 is parsimonious as only two parameters are involved. In view of possible choices for function S_i the model nevertheless appears to be flexible. The choice of the function S_i clearly affects the power of the test for a particular class of alternatives. Further research on goodness-of-fit tests and parameter estimation is needed.

The parameter b, which controls the dependence structure of the model, may be interpreted as describing consistency of replicated binary responses. Derived tests of hypothesis of no dependence - and thus of no consistency - of observers judgements are equivalent to well-known distribution-free tests. Both tests are easy to use and their working is readily understood.

Acknowledgement

We would like to thank Dr. N.J.D. Nagelkerke and Professor Dr. Ir. L. de Pater for helpful discussions and a critical reading of the manuscript.

References

Cox,D.R. and Hinkley,D.V. (1974).
Theoretical Statistics. Chapman and Hall, London.

deJonge,A.B. and Rashbass,C.(1982).
Apparent numerosity differences in pairs of static random dot patterns.
Perception, 11,p.32-34.

Lehmann,E.L. (1975).
Nonparametrics: Statistical Methods Based on Ranks.
Holden-Day, San Francisco.

Ontvangen: 11-8-1983