Advances in Methodology, Data Analysis, and Statistics
Anuška Ferligoj (Editor), Metodološki zvezki, 14
Ljubljana: FDV, 1998

The Mixture Index of Fit

Tamás Rudas[1]

Abstract

The mixture index of fit provides a novel way of measuring the fit of a statistical model. It is derived from the assumption that the statistical model of interest can only describe a fraction of the underlying population. The larger is this fraction, the better is the fit of the model. This approach is in sharp contrast with the traditional methods of testing goodness-of-fit, where the model is assumed to describe the entire population and the likelihood of the observations is assessed under this assumption. The paper reviews reasons for which the traditional approach is not satisfactory in many cases, and illustrates applications of the mixture index of fit.

1 Introduction

Traditional approaches to measuring the fit of a statistical model are not satisfactory for the needs of applications in several respects. The mixture index of it offers an alternative method of assessing model fit, which is free from some of these drawbacks.

Tests of fit of a model to a contingency table are based on comparing the actual data to the expectation of the observations under the assumption that the model is true. The resulting statistics can be given a *test of fit*, or a *measure of fit* interpretation. The validity of these procedures can be questioned for two main reasons. First, when the model is not true, a comparison of the data to what could only be expected if it were, is of very little meaning. Second, the actual distribution of the statistic may be very different from the reference distribution if some of the underlying assumptions are violated. These problems will be illustrated in Section 2.

The mixture index of fit is based on an approach which is very different from the idea underlying the classical tests: it is not assumed that a simple model can describe the entire population. Consequently, we will not try to assess whether or not the data provide evidence against this assumption in terms of being too unlikely, or in terms of showing large deviations from what would be expected under the model. More precisely, it is assumed that there are two groups in the population. In one of them, the model of interest holds true, while the other one is completely unrestricted. The sizes of these groups are not known. The larger is the size of the fraction where the model of interest holds true the better the model fits the underlying population. The size of this fraction is the *mixture index of fit*, and it is described, together with several of its properties, in

[1] Central European University, Nádor u. 9., H-1051 Budapest, Hungary

Section 3. Advantages of the mixture index of fit include that its definition does not rely on assumptions that may not be true, and its estimated values do not depend on the sample size in the way in which chi-square-related quantities do.

Some of the published applications of the mixture index of fit are reviewed in Section 4. These include fitting models to social mobility data and the analysis of differential item functioning in educational testing. The results using the mixture index of fit will be compared to those obtained by traditional approaches. It is illustrated, how the mixture index of fit can be used to compare the fits of nested models.

Finally, Section 5 considers briefly some possibilities for the generalization of the mixture index of fit to statistical problems involving continuous data.

2 Difficulties with chi-squared tests of fit

The standard method of testing the fit of a statistical model for a contingency table is to compute the table of estimated, or expected frequencies, and to compare it to the table of observed frequencies. The table of expected frequencies represents the distribution which, if the model were true, would have made the observations more likely than any other distribution. Then, one is interested in knowing whether, under these 'most favorable' conditions, the actual observations belong to the group of the most unlikely observation. If yes, the hypothesis is rejected. To implement this procedure, one needs to know the distribution of the statistic which is used to compare the observed and estimated frequencies. The statistics that are used most frequently for this comparison are the Pearson chi-squared and the likelihood ratio statistics.

The distributions of these statistics depend on several factors, including the size of the table, the sample size and the actual true probabilities. Therefore, using the true distribution of the statistic in the above procedure is impossible and approximations better than the one to be described here are not practical, or at least used to be not practical when sufficient computational capacity was not available to every scientist. Fortunately, as the sample size increases, the true distributions of these statistics are getting closer and closer to a distribution that only depends on one quantity which can be easily computed from certain properties of the model and the size of the table (namely the number of degrees of freedom). This *asymptotic* distribution is a chi-squared distribution on the relevant number of degrees of freedom. The critical values of these distributions are tabulated and routinely used in testing.

When the critical values used in testing are taken from the asymptotic distribution, the actual level of the test may be very different from the nominal level. There have been various 'practical' rules suggested to decide whether this procedure is acceptable.

Some of the criteria suggest that the smallest expected frequency, in order to be able to use the critical value from the asymptotic distribution instead of the actual one, for the Pearson chi-squared statistic, should be at least 5 (Fisher, 1941, p 82); or that it should be 10, at least (Cramer, 1946, p 420). These criteria are not very practical, because whether or not they are satisfied, turns out only after the data collection has been finished. Other criteria have been suggested in terms of the ratio of the sample size to the number of cells of the table, including that this should be at least 4 or 5 (Fienberg, 1979), or that at least 3 (Rudas, 1986).

The likelihood ratio statistic appeared in most of the studies to converge slower to the asymptotic distribution than the Pearson chi-squared statistic, however the exact results depend on various conditions.

For a unified treatment of these goodness-of-fit statistics see Read and Cressie (1988).

When a test of fit needs to be performed, and the size of the sample may be too small for the application of the critical values of the asymptotic distribution, computer simulations may be used to approximate the true distribution of the test statistic.

There are several measures of fit which are derived from the Pearson chi-squared statistic. Their correct interpretation poses problems similar to those described above.

When the sample size is 'large', the application of the asymptotic critical values does not appear to be problematic, but one runs into other kinds of difficulties. These come from the fact that the Pearson chi-squared (and many related) statistics are proportional to the sample size. This means that if two observed distributions are the same, one with sample size N_1, and the other one with sample size N_2, then the ratio of the relevant Pearson chi-squared statistic values is N_1/N_2. Practically, with large samples, which are desirable otherwise, one has to reject the simple models of interest most of the time.

This does not mean that something is wrong with chi-squared tests, rather that they do something different from what many of the users expects them to do. The test, correctly, detects relatively weak effects, in terms of deviations from the model, when applied to large samples. The procedure is a test for statistical significance. What most of the users of this procedure are interested in, however, is not statistical but subject matter significance.

Testing for subject matter significance requires a precise specification of how large effects are considered important from the point of view of the scientific problem at hand, and the statistical testing procedure needs to be adjusted accordingly. This usually leads to non-standard statistical problems.

One advantage of the approach presented here is that it makes the handling of subject matter significance possible by separating the two attributes that may make an effect important: the size of the effect and the size of the part of the population where it is present. Other advantages include that it is derived from a framework which is always valid, it has a straightforward interpretation, and it does not depend on the sample size in the usual sense.

3 The mixture index of fit

The mixture index of fit is a measure of the ability of model H to describe the population underlying the observed data. It is assumed that there are two groups in the population. In one of them, the model of interest holds, that is, the distribution describing the first group belongs to model H. There are no assumptions regarding the other group. The sizes of the two groups are unknown. The relative size of the first group (where H holds true) is denoted by $1 - \pi$, and of the second (unrestricted) group by π, where $0 \leq \pi \leq 1$. It follows, that an observation comes from a distribution that belongs to H (that is, from the first group) with probability $1 - \pi$, and with probability π it comes from an other distribution (that is, from the second group). In other words, the distribution P that

describes the population has the following mixture representation:

$$P = (1 - \pi)Q + \pi R, \tag{1}$$

where $Q \in H$ is the distribution in the first group and R is the distribution in the second group. The only restriction in (1) is that the distribution Q belongs to model H. If $1 - \pi = 0$, which is its smallest value, (1) simply says that the size of the fraction where model H is valid, is zero, and the entire population is described by an unrestricted distribution R. That is, if $1 - \pi = 0$, (1) does not restrict the distribution P at all. When $1 - \pi = 1$, then the assumption in (1) is that model H describes the entire population. This is the usual null hypothesis when model H is tested.

As $1 - \pi$ moves from 0 to 1, the assumption in (1) moves from being not restrictive at all, to assuming that model H describes the entire population. In every case, whatever the population may be, model (1) is true for at least one value of $1 - \pi$ (when it is equal to zero). If there are several values of $1 - \pi$ for which (1) holds, then there is a largest one from among these.

To illustrate this, suppose that the model of interest is independence and the distribution in the population is described by the following 2x2 table

0.208	0.162
0.252	0.378

Then, as it is easy to see, the population is the mixture of the following two distributions, with respective probabilities 0.9 (= $1 - \pi$) and 0.1 (= π):

0.12	0.18
0.28	0.42

1	0
0	0

Here the first distribution is independent. The marginal probability in the upper left hand side cell is 0.9*0.12+.01*1=0.208, and in the upper right hand side cell is 0.9*0.18+0.1*0=0.162 and the other probabilities can be obtained similarly.

It can be seen that the above distribution cannot be decomposed into two distributions, one of which is independent, so that the independent distribution would have a larger weight than 0.9. The maximum value of $1 - \pi$ with which the representation (1) is possible, denoted by $1 - \pi^*$, measures the ability of the model of interest to describe the population. This is the mixture index of fit; its value in the above example is 0.9. On the other hand, π^* is an index of misfit, its value in the above example is 0.1.

The mixture index of fit is defined as a *population parameter*, that is, one that can only be computed if the true distribution in the population is known. In practical applications, it needs to be estimated from a sample. This can be done by computing its value for the observed distribution. The value obtained is the maximum likelihood estimate of the mixture index of fit.

The problem of maximum likelihood estimation of π^*, therefore, reduces to the problem of calculation of the index for a given (estimated) distribution. For this problem, and

for *any* model H, Rudas, Clogg, and Lindsay (1994) suggested an algorithm that involved repeated application of the EM algorithm (Dempster, Laird, and Rubin, 1977). For the case when H is a log-linear model for the contingency table, Xi (1996) suggested a more efficient algorithm.

Because the estimate is a function of the observed distribution (and not of the observed frequencies), the estimated values do not depend on the sample size. For two samples with different sizes but yielding the same distribution, the estimated values of π^* will be identical. This is in sharp contrast with the chi-squared values, which in this case would be proportional to the sample sizes. In what way does sample size effect statistical inference regarding π^* if the estimates do not depend on it? One expects a larger sample to provide more information than a smaller one, and in the present case this means that the confidence interval for the true value π^* will be shorter for the larger sample than for the smaller sample. A method of computing confidence intervals for the mixture index of fit was described in Rudas, Clogg, and Lindsay (1994).

The unrestricted distribution in the mixture representation describes the distribution in that part of the population where the model of interest does not hold true. This distribution can be given a residual interpretation. The overall importance of this residual is measured by its relative size, π^*. The residuals associated with the mixture index of fit are results of an estimation procedure within a valid framework, in contrast with usual residuals, which are based on the assumption that the model of interest describes the entire population. When this assumption is not valid, the residuals in the usual sense have very little meaning.

4 Some applications of the mixture index of fit

Clogg, Rudas, and Xi (1995) and Clogg, Rudas, and Matthews (1997) considered various simple models for social mobility tables, and their respective abilities to account for well-known sets of mobility data, including the mobility table published by Blau, and Duncan (1967), in the 5x5 form, as condensed by Knoke and Burke (1980). The models investigated were independence, quasi-independence and quasi-uniform association, and gave the Pearson chi-squared values of 875.10, 269.07 and 30.78, on 16, 11, and 10 degrees of freedom, respectively. The estimated values of π^* were 0.310, 0.147, and 0.052, respectively. The analysis suggested, that the independence model may account for nearly 70% of the population, the quasi-independence model for nearly 85%, and the model of quasi-uniform association for nearly 95% of the population. The three models are nested in the sense that for two-way tables of a fixed size, all the independent distributions are contained among the quasi-independent ones and these are contained among the ones where quasi-uniform association holds true. The mixture index of fit is monotone in this case in that if $H_1 \subseteq H_2$ then $\pi^*(P, H_1) \geq \pi^*(P, H_2)$, and, consequently, the same inequality holds for the estimated values, no matter what the observations are. Therefore, any model, which contains quasi-uniform association as a special case, can only improve model fit in term of the description of part of the 5% not accounted for by the model of quasi-uniform association. Note that the standard statistical decision based on chi-squared statistics is not monotone in the above sense.

For the model of quasi-uniform association, the 95% confidence interval for the true

value of π^* contains zero, that is, at the 5% level, the hypothesis that the model of quai-uniform association describes the entire population cannot be rejected.

Clogg, Rudas, and Xi (1995) considered further aspects of the analysis based on the mixture index of fit, and Clogg, Rudas, and Matthews (1997) suggested simple graphical methods for the analysis of model fit and of the residual distribution.

Rudas and Zwick (1997) applied an analysis based on the mixture index fit to items from the 1993 Advance Placement Physics B Exam of the *Educational Testing Service*. The question of interest was the lack or presence of differential item functioning for the groups of male and female examinees. In educational testing, differential item functioning is said to be present, if individuals having the same ability, but belonging to different groups in the population, have different chances of responding to a question correctly.

It was found that for the 10 items considered, the hypothesis of no differential item functioning can describe an estimated $94 - 98\%$ of the population, that is the test items appeared to be free of differential item functioning to a great extent. A less restrictive model is the model of uniform differential item functioning. This model assumes that there may be differential item functioning present, but when the amount of differential item functioning is measured by the conditional odds ratio of grouping (male or female) and item response (correct or incorrect), conditioned on ability, then this conditional odds ratio takes on the same value independently of the ability level. This model led to improvement in fit for five of the ten items analyzed, and in these cases the described fraction improved to $97 - 98\%$. The difference in the π^* values is a direct measure of the improvement in fit due to using a less restrictive model.

One can also estimate the common value of the conditional odds ratio, postulated by the model of uniform differential item functioning, by choosing the value which leads to the smallest value of π. These conditional odds ratios were remarkably close to the odds ratio estimates obtained by other methods. It was also possible to pinpoint those parts of the population where differential item functioning occurs. For most items, this was the part of the population with lower ability levels.

5 Generalizations to other statistical problems

The idea underlying the mixture index of fit applies to any kind of data, and to any statistical model. The application of the π^*-approach may lead to appealing interpretations of well-known statistical quantities. Rudas, Clogg, and Lindsay (1994) considered the relationship between the mixture index of fit and the correlation coefficient.

When two variables have a joint normal distribution, their correlation coefficient can be used as a measure of the strength of their association. If the correlation is equal to any of the values -1, 0, 1, then its interpretation is straightforward. The correct assessment of the amount of association when the correlation coefficient takes on other values is difficult, because of the lack of an intuitive interpretation. As the correlation coefficient is a measure of association, one may expect that the smaller is its absolute value, the more similar is the joint distribution of the two variables to independence. Or, the larger is the absolute value of the correlation coefficient, the lesser is the ability of the model of independence to describe the joint distribution. This suggests the application of the mixture index of fit.

In fact, π^* is the following function of the correlation coefficient

$$\pi^* = 1 - \sqrt{\frac{1-|\rho|}{1+|\rho|}},$$

where $|\rho|$ is the absolute value of the correlation coefficient. For example, when the correlation is 0.6, at most 50% of the population can be described by independence. This is an intuitively clear interpretation of the meaning of the given value of the correlation coefficient.

More generally, to any set of observation, maybe after an initial smoothing step, one may try to find the distribution from a given model which is the closest in the sense of minimizing π^*. This leads naturally to a minimum distance estimation problem based on the mixture index of fit. These generalizations are subject to ongoing research now. Initial results suggest that under suitable conditions, the solution of this minimum distance estimation problem coincides with the solution of other minimization or estimation problems and the mixture index of fit provides a natural measure of fit in these cases.

References

[1] Blau P.M. and Duncan O.D. (1967): *The American Occupational Structure.* The Free Press.

[2] Clogg C.C., Rudas T., and Xi L. (1995): A new index of structure for the analysis of models for mobility tables and other cross classifications. In: P. Marsden (Ed.) *Sociological Methodology, 1995,* Blackwell, 197-222.

[3] Clogg C.C., Rudas T., and Matthews S. (1997): Analysis of contingency tables using graphical displays based on the mixture index of fit. In: M. Greenacre and J. Blasius (Eds.): *Visualization of Categorical Data,* Academic Press, in press.

[4] Cramer H. (1946): *Mathematical Methods of Statistics.* Princeton University Press.

[5] Demster A.P., Laird N.M., and Rubin D.B. (1977): Maximum likelihood from incomplete data via the EM algorithm, *J. Roy. Statist. Soc., Ser B,* **39**, 1-38.

[6] Fienberg S.E. (1979): The use of chi-squared statistics for categorical data problems. *J. Roy. Statist. Soc., Ser B,* **41**, 54-64.

[7] Knoke D. and Burke P.J. (1980): *Log-Linear Models.* Sage.

[8] Read T.R.C. and Cressie N.A.C. (1988): *Goodness-of-fit Statistics for Discrete Multivariate Data.* Springer.

[9] Rudas T. (1986): A Monte Carlo comparison of the small sample behaviour of the Pearson, the Likelihood Ration and the Cressie-Read statistics. *J. Statist. Comput. Simul.*, **24**, 107-120.

[10] Rudas T., Clogg C.C., and Lindsay B.G. (1994): A new index of fit based on mixture methods for the analysis of contingency tables. *J. Roy. Statist. Soc., Ser. B*, **56**, 623-39.

[11] Rudas T. and Zwick R. (1997): Estimating the importance of differential item functioning. *J. Educ. Behav. Statist.*, **22**, 31-45.

[12] Xi L. (1996): *The Mixture Index of Fit.* Ph. D. Thesis, Department of Statistics, The Pennsylvania State University.