Evaluating goodness of fit of discrete distributions: A note on the use of the discrepancy coefficient

Ján Mačutek[1*] (0000-0003-1712-4395), Jozef Kováč[2] (0000-0002-9791-5332)

[1] Mathematical Institute, Slovak Academy of Sciences
[2] Department of Applied Mathematics and Statistics, Faculty of Mathematics, Physics and Informatics, Comenius University
[*] Corresponding author's email: jan.macutek@mat.savba.sk

DOI: Here will be added DOI number by Glottometrics.

ABSTRACT

Goodness of fit of discrete distributions as models in quantitative linguistics is assessed by the usual p-values for relatively small sample sizes, and by the so-called discrepancy coefficient for large sample sizes. Whereas the choice is obvious for very small and very large samples, this methodology leaves a wide gray zone — there are many cases where both approaches could be used. It can happen that the two approaches lead to different conclusion (one of them considers the model to have a good fit, the other says the opposite). In this paper, we present an adjustment of the discrepancy coefficient that can reduce the number of such discrepancies.

Keywords: goodness of fit, discrete distributions, mathematical modelling.

1 Introduction

Goodness-of-fit tests play a very important role in mathematical modelling in any empirical science. Limiting ourselves to quantitative linguistics in the following,[1] the aim of mathematical modelling is an explanation of linguistic phenomena.[2] Such an explanation is possible within a theory, i.e. within a system of interrelated universal laws. However, given that we understand scientific laws as statistical (and that they are, at least in quantitative linguistics, expressed mathematically), a validity of a law can — and should be — statistically corroborated.[3] While no model can fully grasp all facets of reality and

[1] The methodological problems discussed in this paper are valid for discrete distribution models in any empirical science.

[2] "Science does not confine itself to observe phenomena, to describe these observations, and to apply the collected knowledge. The highest aim of any science is the explanation of the phenomena (which also opens up the possibility to predict them)" (Köhler and Altamnn, 2025, p. 278).

[3] We intentionally avoid the formulation "statistically tested". The reason is that statistical tests tend to reject all null hypotheses if the sample size is very large. Consequently, at least in quantitative linguistics, one often uses established rules of thumb rather than proper statistical tests, see, e.g., Mačutek and Wimmer (2013).

can be only its approximation, an approximation can be better or worse, a model can be close to data or far away from them, etc. The famous adage says that "all models are wrong, but some are useful" (Box and Draper, 1987, p. 424), and that "the practical question is how wrong do they have to be to not be useful" (ibid., p. 74).

That "practical question" needs an answer. Often, the answer is obtained by means of statistical tests. If the model is a discrete probability distribution, the chi-square goodness-of-fit test can be used, and the null hypothesis that the data come from a particular distribution is not rejected if the p-value is greater than the significance level (most often 0.05). However, all statistical tests reject practically any hypothesis if the sample size is too large (and nowadays, with many language corpora available, sample sizes can easily be huge). Furthermore, if the differences between data and the model are the same, but the sample size increases, the value of the chi-square statistic grows linearly with the sample size. Therefore, it has become common in quantitative linguistics to use the chi-square test for "reasonably small" samples, and the so called discrepancy coefficient

$$C = \frac{\chi^2}{N}, \tag{1}$$

with χ^2 being the value of the test statistic of the chi-square test and N being the sample size, for "too large" samples (Mačutek and Wimmer, 2013). Usually, the fit is considered satisfactorily good if $C \leq 0.02$. All discrete distributions mentioned in this paper can be found in the discrete distribution dictionary by Wimmer and Altmann (1999).

2 Problems

The first problem is obvious — which approach should one choose if the sample size is neither "really small" nor "too large"? In other words, which sample sizes are so large that a rejection of the null hypothesis is "more a reflection on the size of the sample than on the adequacy of the model" (Browne and Cudeck, 1992)? Or, as Grzybek (2007, p. 74) puts it, "the question is, what is a small text, and where does a large text start? And why do we, in some cases, obtain significant C values when $p(\chi^2)$ is significant, too, but in other cases do not?"

Second, even if the chi-square goodness-of-fit test is chosen, one faces further ambiguities. Denote k the number of classes (i.e., the number of possible or observed values of the phenomenon under study).[4]

[4]If. e.g., phoneme frequencies are studied, one knows the phoneme inventory size in a language and k is the number of all possible classes. On the other hand, the number of all words in a language can never be exactly determined; therefore, one takes the number of words occurring in a text/corpus as the number of classes.

The chi-square test statistic is

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - NP_i)^2}{NP_i}, \tag{2}$$

where f_i is the frequency in the i-th class, P_i is the probability (corresponding to the null hypothesis) that a randomly chosen object is from the i-th class, k is the number of classes, and N is the sample size.

It has, under the null hypothesis, asymptotically the chi-square distribution with $k - 1$ degrees of freedom for fully specified hypothetical distributions, i.e., distributions with all parameters values being known. If r parameters are not known and must be estimated, the test statistic remains asymptotically chi-square distributed, but the degrees of freedom decrease to $k - r - 1$, see Snedecor and Cochran (1989, pp. 76–79).[5]

Regardless of whether some parameter(s) of the distribution from the null hypothesis must be estimated or not, we know only the *asymptotic* (as opposed to exact) distribution of the test statistic. This fact opens a question of the conditions under which the distribution of the test statistic is "close enough" to its limit, i.e., to the chi-square distribution. Answers to this question differ among authors. We provide but a few examples:

- Cochran (1952, p. 328): "[I]t is customary to recommend, in applications of the test, that the smallest expected number in any class should be 10 or (with some writers) 5. (. . .) The numbers 10 or 5 appear to have been arbitrarily chosen."
- Snedecor and Cochran (1989, p. 77): "[N]o expected frequency should be less than 1. Two expected frequencies may be near 1 provided that most of the other expected frequencies exceed 5."
- Kim (2017, p. 153): "The result is relevant only when no more than 20% of cells with expected frequencies < 5 and no cell have expected frequency < 1."

Many other recommendations (some of them more complicated than the ones listed above) can be found in Greenwood and Nikulin (1996, pp. 18–21). We borrow the (non-)conclusion from Agresti (2013, p. 78) according to whom "[i]t is difficult to give a guideline that covers all cases." All choices are (not unreasonable) rules of thumb.

[5] We emphasize that here p is the number of *estimated*, and not of all parameters. Consider, e.g., the test of the null hypothesis data come from the binomial distribution (that has two parameters — n is the number f trials and p the probability of success in each of the trials). The usual situation is that the parameter n is known and one must estimate only one parameter p. In such a case, $r = 1$ and distribution of the test statistic would have $k - 2$ degrees of freedom.

If conditions for the minimum expected frequencies are not satisfied, data classes should be combined so that they conform to the conditions. The number of classes after such combinations is used for counting degrees of freedom of the asymptotic distribution of the test statistic (Snedecor and Cochran, 1989, p. 78). It is obvious that a choice of the minimum expected frequencies has an impact on the degrees of freedom and, consequently, on the resulting p-value. The decision to reject or not to reject the null hypotheses thus partly depends on the choice of one of several rules of thumb.

3 Adjusted discrepancy coefficient

Obviously, one cannot escape a subjective choice of one of rules of thumb for "mid-sized" sample sizes. Different choices can easily lead to contradictory conclusions. Therefore, it seems reasonable to try to develop a more unified approach to assessing goodness of fit of discrete distributions for such sample sizes that would reduce the number of such contradictions. The discrepancy coefficient C in its original form (1) is not a good measure of the fit between data and a model in such cases. The reason is that the chi-square distribution has the mean that is equal to its degrees of freedom (Johnson et al., 1994, p. 420), i.e., its mean is a positive number (depending on the character of data, it can be quite distant from zero). This differs from test statistics that are normally or t-distributed — they have, under the null hypothesis, the mean equal to zero. The difference is a consequence of the definition of the χ^2 test statistic from (2), in which the differences between data and he model are squared.

If the sample size is large, the value of the discrepancy coefficient C from (1) is very low if the model fits data well. However, if a sample is small, the relatively small value of N (relatively if compared with degrees of freedom) is often not able to make the value of C lower than 0.02 (i.e., the conventionally set threshold for good models).

We therefore suggest its adjustment in the form of

$$C_A = \frac{\chi^2 - df}{N}, \tag{3}$$

where df is the number of degrees of freedom of the asymptotic distribution of the χ^2 test statistic (or, in other words, the mean of the asymptotic distribution of the χ^2 test statistic). The adjusted discrepancy coefficient C_A is thus made smaller (if compared with the original C) for small samples, while the change is negligible for large samples.[6]

There are datasets for which the chi-square goodness-of-fit test is not applicable because so many classes must be combined that there do not remain enough degrees of freedom (technically, the value of $k - r - 1$ is not a positive number). In such cases, one can either use a "more tolerant' criterion for what is a

[6] Pederson and Johnson (1990) suggest a very similar adjustment of model discrepancy measures mentioned already by Mosteller (1968) and Moore (1984).

minimum expected frequency (such as. i.e., all expected frequencies must be at least 1). If even that is not suffiecient, a "rough approach" would be not to combine classes at all and to take

$$C_A = \frac{\chi^2 - (k-1)}{N}, \tag{4}$$

with k being the number of classes in the original data (i.e., without any combing of classes).

The adjusted discrepancy coefficient C_A can be used regardless of the sample size, and the value of 0.02 can remain as the threshold. It can have negative values (this is impossible for the original discrepancy coefficient C), but they only hint towards an acceptable fit. However, if the value of the chi-square test statistic (2) is smaller than the number of degrees of freedom, the value of C_A is necessarily negative, and in such a situation any model would be considered good. We recommend using C_A especially for mid-sized samples, or for datasets where the p-value and the original discrepancy coefficient C give contradictory results.

4 Example

In this section, we show an example of frequencies of sentence length (measured on the number of words) from Kelih (2002, p. 94). The data (326 sentences altogether) are presented in Table 1.

Table 1: Distribution of sentence length in a Russian text (x - sentence length in words, $f(x)$ - frequency of sentence length x).

x	$f(x)$	x	$f(x)$	x	$f(x)$
1	49	14	2	27	1
2	48	15	3	28	0
3	44	16	1	29	0
4	32	17	2	30	0
5	32	18	1	31	0
6	14	19	3	32	0
7	22	20	2	33	0
8	17	21	0	34	0
9	22	22	0	35	0
10	8	23	1	36	0
11	10	24	1	37	0
12	5	25	1	38	0
13	3	26	1	39	1

Kelih (2002) modelled the data with the hyperpascal distribution. The distribution is defined by

$$P_x = \frac{\binom{k+x-2}{x-1}}{\binom{m+x-2}{x-1}} q^{x-1}, \quad x = 1, 2, \ldots, \tag{5}$$

with k, m, and q being its parameters. The distribution (5) achieves a good fit in terms of the p-value — we have $\chi^2 = 22.883$ as the value of the test statistic. The asymptotic distribution of the test statistic has (after combining classes with too low expected frequencies) 19 degrees of freedom, which corresponds

to the p-value of 0.243. The optimized parameter values (i.e., ones for which the test statistic achieves the minimum value) are $k = 2.455$, $m = 1.979$, and $q = 0.778$.[7]

However, if one decides to apply the discrepancy coefficient C from (1), its value is

$$C = \frac{22.883}{326} = 0.07, \tag{6}$$

i.e., a value clearly above the conventionally established threshold of 0.02. Recall, however, that the mean of the (asymptotic) distribution of the test statistic is equal to its degrees of freedom, i.e., 19 in this case. With the data from Table 1, $C \leq 0.02$ only if $\chi^2 \leq 6.52$, i.e., only value much lower than the mean of the distribution are interpreted as a good fit.

On the other hand, using the adjusted discrepancy coefficient C_A from (3) yields

$$C_A = \frac{22.883 - 19}{326} = 0.012, \tag{7}$$

a value suggesting a good fit (we note that the same suggestion is arrived at if one chooses the p-value approach, and that the sample size of 326 is certainly not very large).

5 Conclusion

In this paper, we suggest a modification of the discrepancy coefficient. It can be used regardless of the sample size. It is, admittedly, a rule of thumb, and the threshold for a good fit (the value of 0.02) is conventionally set. However, a very similar coefficient has been used for decades for large samples, and even the chi-square goodness-of-fit test involves some rules of thumb (criteria when some classes should be combined, and, after all, the threshold for the p-value is also a convention).

The adjusted discrepancy coefficient is especially suited to assessing goodness of fit of discrete distributions (not only) in quantitative linguistics in the situation when the sample size is neither clearly small nor large, and the p-value contradicts the original discrepancy coefficient C (i.e., the fit is good according to one of them, but not according to the other).

Acknowledgments

J. Mačutek was funded by EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I03-03-V04-00748.

References

Agresti, A. (2013). *Categorical Data Analysis*. Wiley.

[7] We rounded all values to three decimal places; Kelih (2002) presents more exact values.

Box, G. E. P., Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces*. Wiley.

Browne, M. W., Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*(2), 230–258. https://doi.org/10.1177/0049124192021002005

Cochran, W. G. (1952). The χ^2 test of goodness of fit. *The Annals of Mathematical Statistics*, *23*(3), 315–345. https://doi.org/10.1214/aoms/1177729380

Greenwood, P. E., Nikulin, M. S. (1996). *A Guide to Chi-Squared Testing*. Wiley.

Grzybek, P. (2007). History and methodology of word length studies. In P. Grzybek (Ed.), *Contributions to the Science of Text and Language: Word Length Studies and Related Issues* (pp. 15–90). Springer Netherlands. https://doi.org/10.1007/978-1-4020-4068-9_2

Johnson, N. L., Kotz, S., Balakrishnan, N. (1994). *Continuous Univariate Distributions, Volume 1*. Wiley.

Kelih, E. (2002). Untersuchungen zur Satzlänge in slowenischen und russischen Prosatexten. Band 2 [Master's thesis, University of Graz]. https://homepage.univie.ac.at/emmerich.kelih/wp-content/uploads/p2002_ke_satzlaenge_band2.pdf

Kim, H.-Y. (2017). Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test. *Restorative Dentistry & Endodontics*, *42*(2), 152–155. https://doi.org/10.5395/rde.2017.42.2.152

Köhler, R., Altamnn, G. (2025). Aims and methods of quantitative linguistics. In E. Kelih, J. Mačutek, M. Koščová (Eds.), *Quantification in Linguistics and Text Analysis: Selected Papers of Gabriel Altmann* (pp. 261–286). De Gruyter Mouton.

Mačutek, J., Wimmer, G. (2013). Evaluating goodness-of-fit of discrete distribution models in quantitative linguistics. *Journal of Quantitative Linguistics*, *20*(3), 227–240. https://doi.org/10.1080/09296174.2013.799912

Moore, D. S. (1984). Measures of lack of fit from tests of chi-squared type. *Journal of Statistical Planning and Inference*, *10*(2), 151–166. https://doi.org/10.1016/0378-3758(84)90067-3

Mosteller, F. (1968). Association and estimation in contingency tables. *Journal of the American Statistical Association*, *63*(321), 1–28. https://doi.org/10.1080/01621459.1968.11009219

Pederson, S. P., Johnson, M. E. (1990). Estimating model discrepancy. *Technometrics*, *32*(3), 305–314. https://doi.org/10.1080/00401706.1990.10484683

Snedecor, G. W., Cochran, W. G. (1989). *Statistical Methods*. Iowa State University Press.

Wimmer, G., Altmann, G. (1999). *Thesaurus of Univariate Discrete Probability Distributions*. STAMM.