NOTICE
This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Department of Energy, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights.

THE DEGREE OF SIMILARITY WITHIN SEVERAL
BINOMIAL SAMPLES: A MEASURE AND TEST THEREOF

by

Joseph L. Fleiss
Division of Biostatistics, Columbia University
School of Public Health, 600 W. 168 St.,
New York, New York 10032, U.S.A.
and
Biometrics Research Department
New York State Psychiatric Institute

and

Jack M. Cuzick
Department of Mathematical Statistics
Columbia University, New York, New York 10027

March 9, 1978

STUDY ON STATISTICS AND ENVIRONMENTAL
FACTORS IN HEALTH

PREPARED UNDER SUPPORT TO SIMS FROM

DEPARTMENT OF ENERGY (DOE)
ENVIRONMENTAL PROTECTION AGENCY (EPA)
NATIONAL SCIENCE FOUNDATION (NSF)
ROCKEFELLER FOUNDATION
SLOAN FOUNDATION

COLUMBIA UNIVERSITY
NEW YORK, NEW YORK

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

EAB

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

The Degree of Similarity Within Several
Binomial Samples: A Measure and Test Thereof

Summary

A measure is proposed for the degree of similarity within several binomial samples. Correspondences with other measures are pointed out, and a test is presented for its statistical significance. A numerical example is given.

Key words: kappa statistics, binomial data, intraclass correlation

1. Introduction

A number of apparently unrelated problems may be posed in terms of the degree of similarity within N binomial samples where, within the i th sample ($i=1, \dots, N$), there are n_i observations of which x_i are "successes" and $n_i - x_i$ are "failures."

a) Consider the assessment of the degree of concordance between the readings at several mobile air pollution monitoring stations in a city. On each of a sample of N days, the stations are situated throughout the city at random, and each is characterized by whether or not the readings on a specified pollutant (e.g., sulphur dioxide) exceeded an officially designated threshold. The numbers of stations providing data may vary from day to day because of reasons such as equipment failure. Let n_i denote the total number of stations reporting on day i , and x_i the number of stations with at least one reading above the threshold. The more similar are the readings (i.e., the closer the x_i 's are to 0 or to n_i), the more concordant they may be taken to be.

b) Consider a study of the importance of familial factors in the development of a condition such as adolescent hypertension. A sample of N sibships is studied, with n_i denoting the number of siblings in the i th sibship and x_i the number of them who are affected. Familial factors would be said to be strong if each x_i were either near 0 or near n_i .

c) Consider a study of the reliability of judgments as to whether subjects are positive or negative on a trait. A sample of N subjects is studied, with n_i denoting the total number of judgments on the i th subject

and x_i the number of positive judgments. The closer the judgments are to one another (i.e., the closer each x_i is to 0 or n_i), the more reliable they may be taken to be.

In this paper, a measure of similarity is proposed as appropriate to each of the above problems. The measure varies from negative values for the case of dissimilarity, through zero for a degree of similarity exactly in accordance with chance, to unity for perfect similarity. Correspondences with other measures are pointed out, and a test for its significance presented.

2. The Measure of Similarity

Let N denote the number of binomial samples under study, n_i the number of trials in the i th sample, and x_i the number of successes in the i th sample. Define

$$\bar{n} = \frac{1}{N} \sum n_i, \quad (1)$$

the mean number of trials per sample;

$$\bar{p} = \frac{1}{Nn} \sum x_i, \quad (2)$$

the overall proportion of success; and $\bar{q} = 1 - \bar{p}$. Then

$$\kappa = 1 - \frac{\sum \frac{x_i (n_i - x_i)}{n_i}}{N(\bar{n}-1)\bar{p}\bar{q}} \quad (3)$$

is a measure with the following properties.

(a) If there is no inter-sample variation in the proportion of successes (i.e., if $x_i/n_i = \bar{p}$ for all i), then there is less similarity (or, more dissimilarity) within than between samples. In this case κ may be seen to assume its minimum value of $-1/(\bar{n}-1)$.

b) If the several proportions x_i/n_i vary as binomial proportions with parameters n_i and a common probability $\bar{p}$, then there is as much similarity within as between samples. In this case, the expected value of κ is approximately equal to 0.

c) If each proportion x_i/n_i assumes either the values 0 or 1, then there is perfect similarity within samples. In this case, κ may be seen to assume the value 1.

3. Correspondence with Other Measures

The statistic presented above is one of a class of measures called kappa statistics available for indexing the degree of agreement or similarity beyond chance among categorical variables. Originally proposed for the case of a fixed set of judges (Cohen, 1960, 1968; Fleiss, 1975; Landis and Koch, 1977a; Light, 1971), kappa has been extended to the case of varying sets of judges (Fleiss, 1971; Fleiss, Nee and Landis, 1978; Landis and Koch, 1977b). When the number of trials in each of the N samples is constant, the statistic given in (3) is identical to the one proposed earlier by Fleiss (1971).

Landis and Koch (1977b) approached the measurement of similarity within several binomial samples by applying the algebra of a one-way random effects analysis of variance to the data obtained by coding successes 1 and failures 0. In the present notation, the mean square between samples is

$$BMS = \frac{1}{N-1} \sum \frac{1}{n_i} (x_i - n_i \bar{p})^2 \quad (4)$$

and the mean square within samples is

$$WMS = \frac{1}{N(\bar{n}-1)} \sum \frac{x_i(n_i - x_i)}{n_i} \quad (5)$$

They proposed as a measure of similarity the sample intraclass correlation coefficient,

$$r = \frac{BMS - WMS}{BMS + (n_o - 1)WMS} \quad (6)$$

where

$$n_o = \bar{n} - \frac{s_n^2}{N\bar{n}} \quad (7)$$

and

$$s_n^2 = \frac{1}{N-1} \sum (n_i - \bar{n})^2, \quad (8)$$

the variance of the sample sizes.

It may easily be checked that, if BMS is redefined to have a divisor N instead of N-1, then

$$r = \frac{\kappa}{1 - f} \quad (9)$$

where

$$f = \frac{s_n^2}{N\bar{n}^2} (1 - \kappa). \quad (10)$$

Thus, if N is even moderately large, Landis and Koch's measure is virtually identical to κ . This represents yet another correspondence between kappa statistics and intraclass correlation coefficients (Fleiss, 1975; Fleiss and Cohen, 1973). Given this correspondence, one may describe the degree of similarity to be good if κ exceeds 0.60 or so (Landis and Koch, 1977a).

4. The Distribution of Kappa

Consider testing the hypothesis that the underlying probability of success in each sample is constant. A test statistic for this hypothesis is

$$\chi^2 = \frac{\sum \frac{(x_i - n_i \bar{p})^2}{n_i}}{\bar{p} \bar{q}} . \quad (11)$$

When the hypothesis is true and when N is large, χ^2 is approximately normally distributed with mean $N-1$ and variance

$$\text{Var } (\chi^2) \sim \frac{2N(\bar{n}_H - 1)}{\bar{n}_H} + \frac{N(\bar{n} - \bar{n}_H)(1 - 4\bar{p}\bar{q})}{\bar{n} \bar{n}_H \bar{p} \bar{q}} \quad (12)$$

(Cuzick, 1977; Haldane, 1939), where $\bar{n}_H$ is the harmonic mean of the numbers of trials,

$$\bar{n}_H = \frac{N}{\sum \frac{1}{n_i}} . \quad (13)$$

Given the easily verified identity

$$\kappa = \frac{\chi^2 - N}{N(\bar{n} - 1)} , \quad (14)$$

it follows that, when N is large and when the N samples are subject to the same underlying probability of success, κ is approximately normally distributed with mean

$$E(\kappa) \sim \frac{-1}{N(\bar{n} - 1)} \sim 0 \quad (15)$$

and variance

$$\text{Var } (\kappa) \sim \frac{2(\bar{n}_H - 1)}{N\bar{n}_H(\bar{n} - 1)^2} + \frac{(\bar{n} - \bar{n}_H)(1 - 4\bar{p}\bar{q})}{N\bar{n}\bar{n}_H(\bar{n} - 1)^2\bar{p}\bar{q}} \quad (16)$$

When the arithmetic and harmonic means of the numbers of trials are approximately equal, or when $\bar{p}$ is close to $\frac{1}{2}$ (say, $0.35 \leq \bar{p} \leq 0.65$), then

$$\text{Var } (\kappa) \sim \frac{2(\bar{n}_H - 1)}{N\bar{n}_H(\bar{n} - 1)^2} \quad (17)$$

Finally, if the numbers of trials are equal, say $n_i = n$ for $i=1, \dots, N$, then

$$\text{Var } (\kappa) \sim \frac{2}{Nn(n-1)} \quad (18)$$

a result presented by Fleiss, Nee and Landis (1978).

5. Numerical Example

Insert Table 1 about here

Consider the hypothetical data in Table 1. The number of samples is $N=15$, the mean number of trials is $\bar{n}=47/15 = 3.133$, and the overall proportion of success is $\bar{p} = 32/47 = 0.681$. The value of kappa is

$$\kappa = 1 - \frac{5.050}{15(2.133)(0.681)(0.319)} = 0.274. \quad (19)$$

The expected value of κ is, by (15), equal to -0.031.

Coding successes 1 and failures 0, the mean squares in the resulting analysis of variance table are

$$\text{BMS} = \frac{1}{14} (5.163) = 0.369 \quad (20)$$

and

$$\text{WMS} = \frac{1}{15(2.133)} (5.050) = 0.158. \quad (21)$$

The value of the constant n_o (7) is

$$n_o = 3.133 - \frac{0.838}{47} = 3.115, \quad (22)$$

so the value of Landis and Koch's intraclass correlation coefficient of similarity (6) is

$$r = \frac{0.369 - 0.158}{0.369 + (2.115)(0.158)} = 0.300, \quad (23)$$

slightly larger than the value of kappa (19). If BMS is redefined to have 15 instead of 14 as its divisor, it becomes equal to 0.344 and the value of r becomes 0.274, identical to the value of kappa.

The harmonic mean of the number of trials is

$$\bar{n}_H = \frac{15}{5.200} = 2.885, \quad (24)$$

and the approximate variance of kappa (16) is equal to 0.0193. The statistical significance of kappa may be tested by referring the value of

$$z = \frac{\kappa - E(\kappa)}{\sqrt{\text{Var}(\kappa)}} = \frac{0.274 + 0.031}{0.14} = 2.18 \quad (25)$$

to the standard normal distribution. Kappa is seen to be significantly different from 0 at the 0.05 level, although the degree of similarity is weak. If equation (17) were used to find the variance, the value $\text{Var}(\kappa) = 0.0191$ would be obtained, and the resulting value of z (25) would be 2.18, the same as above.

Acknowledgment

This work was supported in part by a grant from the SIAM Institute for Mathematics and Society to conduct biostatistical and epidemiological studies into the health effects of air pollution.

References

- Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20, 37-46.
- Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin 70, 213-220.
- Cuzick, J. (1977). Asymptotic normality of χ^2 in $m \times n$ tables with n large and small cell expectations. Unpublished technical report, SIAM Institute for Mathematics and Society Study of Environmental Factors and Health.
- Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. Psychological Bulletin 76, 378-382.
- Fleiss, J.L. (1975). Measuring agreement between two judges on the presence or absence of a trait. Biometrics 31, 651-659.
- Fleiss, J.L. and Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and Psychological Measurement 33, 613-619.
- Fleiss, J.L., Nee, J. and Landis, J.R. (1978). The large sample variance of kappa in the case of different sets of raters. In preparation.
- Haldane, J.B.S. (1939). The mean and variance of χ^2 , when used as a test of homogeneity, when expectations are small. Biometrika 31, 346-355.
- Landis, J.R. and Koch, G.G. (1977a). The measurement of observer agreement for categorical data. Biometrics 33, 159-174.
- Landis, J.R. and Koch, G.G. (1977b). A one-way components of variance model for categorical data. Biometrics 33, 671-679.
- Light, R.J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. Psychological Bulletin 76, 365-377.

TABLE 1. Hypothetical Data on the Degree of Similarity Within
15 Binomial Samples

Sample (i)	Number of Trials (n_i)	Number of Successes (x_i)	Sample	Number of Trials	Number of Successes
1	2	2	9	3	0
2	2	0	10	3	3
3	3	2	11	3	2
4	4	3	12	5	4
5	3	1	13	2	2
6	4	1	14	4	3
7	2	2	15	3	3
8	4	4	Total	47	32