Revista Mexicana de Análisis de la Conducta,
1977, Vol. 3, Núm. 2, pp. 181-186.

Estimating the reliability of observational data with nonparametric confidence intervals

Estimación de la Confiabilidad de Datos de Observación con Intervalos de Confianza No-Paramétricos

Michael Meighan

Infant Development Laboratory
University of Kansas Medical Center[1]

ABSTRACT

Researchers who use observational methods traditionally report a sample of reliability coefficients generated by having multiple observers score the same data. This sample of coefficients is then used to estimate the reliability of the entire data set. As such, two things must be considered: the zise and variability of the sample, and the techniques used to estimate reliability. A technique is described which allows the researcher to estimate the median reliability of a data set with confidence intervals of 95 and 99 percent.

DESCRIPTORS: reliability coefficient, observational methods, nonparametric intervals, confidence limits, sampling, inferential statistics, statistical hypothesis testing.

RESUMEN

Aquellos investigadores que usan métodos de observación tradicionalmente reportan una muestra de coeficientes de confiabilidad generados por medio de las calificaciones de varios observadores sobre los mismos datos. Esta muestra de coeficientes se usa para estimar la confiabilidad del conjunto completo de datos. De esta manera, se deben considerar dos aspectos: el tamaño y la variabilidad de la muestra y las técnicas usadas para estimar la confiabilidad. El presente artículo describe una técnica que le permite al investigador estimar la mediana de confia-

[1] This research was partially supported by a United States Public Health Service Training Grant (HD-00183) from the National Institute of Child Health and Human Development. Special thanks to Mary Carrigan, Ron Kent, Gary Clark, and Ron Schoenberg for their contributions. Copies may be obtained by writing to Michael Meighan, Infant Development Laboratory, University of Kansas Medical Center, Kansas City, Kansas, U. S. A. 66103.

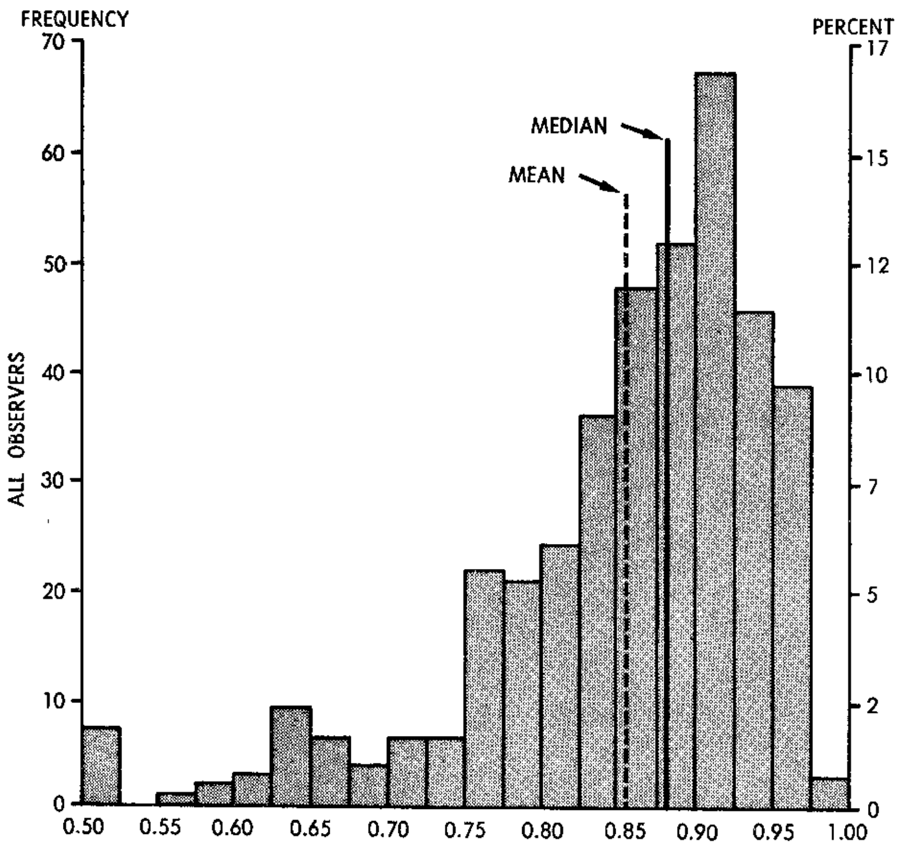
bilidad de un conjunto de datos con intervalos de confianza de 95 y 99 por ciento.
DESCRIPTORES: coeficiente de confiabilidad, métodos de observación, intervalos no-paramétricos, límites de confianza, muestreo, estadística inferencial, prueba de hipótesis estadísticas.

Researchers who use observational methods usually report on the reliability of their measurements by quoting coefficients of agreement between two observers. These coefficients are important because they estimate both the precision and accuracy (Topping, 1962; Stalling & Gillmore, 1971) of the measurements. Precision being the closeness with which measurements agree with one another, while accuracy refers to the closeness of the measurements to the "real" or "actual" value of the thing being measured. The fact that reliability coefficients are only estimates of these two attributes is, however, often overlooked, but researchers rarely report a coefficient of reliability for each data point, and since reliability is assessed only infrequently then the researcher may be said to have *sampled* the set of possible coefficients, which sample is then used to estimate the precision and accuracy of the whole set. Three things must therefore be considered: the parameter to be estimated, the size and variability of the sample, and the rules for estimation.

It is quite common in the literature that the mean of the sample of coefficients is used as the estimate of the mean of the population, a form of point estimation. Figure 1 shows the distribution of percentage agreement coefficients from a large-scale observational project (Meighan, Burgess, & Lovitt, 1977). The distribution is clearly heavily skewed toward the higher coefficients. Blalock (1960) notes that "whenever there are considerably more extreme cases in one direction than the other, the median will generally be more appropriate than the mean" (p. 58) as a measure of central tendency. Further, the even split of the sample around the median allows the use of the binomial to estimate the median of the population using the techniques of non parametric interval estimation which are superior to the tech techniques of point estimation since they allow us to set confidence intervals.

Edgington (1969) gives us one way to set confidence intervals using only the assumption of random sampling. In his discussion of distribution-free confidence intervals, the author sets forth the method of calculating intervals for the mean, median, and various ratios of the hypothesized population. An easier way, however, is to refer to Table 1 wich is drawn from a table published by Ciba-Geigy Limited (1970) based upon the binomial distribution. The theory and mathematics behind the use of this table can be found in Conover (1971, p. 110 ff).

To use the table we arrange our coefficients in ascending order, from the lowest to the highest. Reading down the column under N we find the sample size, and reading across the table gives us the beginning and the end of the interval which estimates the median of the population of all possible reliability coefficients. For example, if we sample reliability 12 times and arrange the



DISTRIBUTION OF PERCENTAJE AGREEMENT SCORES FOR 16 OBSERVERS (N = 717) SHOWING MEAN AND MEDIAN OF THE DISTRIBUTION

TABLE 1

Confidence Limits for Estimating the Median of a Population at the .05 and .01 Levels for Sample

Size N = 1 to 50

N	.05 level b*	.05 level e**	.01 level b	.01 level e	N	.05 level b	.05 level e	.01 level b	.01 level e
1	—	—	—	—	26	7	19	6	20
2	—	—	—	—	27	7	20	6	21
3	—	—	—	—	28	8	20	6	22
4	—	—	—	—	29	8	21	7	22
5	—	—	—	—	30	9	21	7	23
6	0	6	—	—	31	9	22	7	24
7	0	7	—	—	32	9	23	8	24
8	0	8	0	8	33	10	23	8	25
9	1	8	0	9	34	10	24	9	25
10	1	9	0	10	35	11	24	9	26
11	1	10	0	11	36	11	25	9	27
12	2	10	1	11	37	12	25	10	27
13	2	11	1	12	38	12	26	10	28
14	2	12	1	13	39	12	27	11	28
15	3	12	2	13	40	13	27	11	29
16	3	13	2	14	41	13	28	11	30
17	4	13	2	15	42	14	28	12	30
18	4	14	3	15	43	14	29	12	31
19	4	15	3	16	44	15	29	13	31
20	5	15	3	17	45	15	30	13	32
21	5	16	4	17	46	15	31	13	33
22	5	17	4	18	47	16	31	14	33
23	6	17	4	19	48	16	32	14	34
24	6	18	5	19	49	17	32	15	34
25	7	18	5	20	50	17	33	15	35

* beginning of estimating interval **end of interval

Adapted from *Documenta Geigy, Scientific Tables, 7th edition*, Page 105, Courtesy of Ciba-Geigy Limited, Basle, Switzerland.

coefficients in ascending order we might get the set: .79, .83, .84, .87, .91, .94, .94, .95, .95, .97, .98, 1.00, & 1.00. Selecting the .05 level we find that the interval which estimates our population median begins with the 2nd coefficient in the series and ends with the 10th. We say with 95% confidence, in other words, that the population median is between .83 and .98, or, more precisely, that it is greater than .83 and less than .98 (Edgington, 1969, p. 54).

Notice that with a sample size of 5 or less coefficients the table tells us that we cannot legitimately make an estimate of the median whith 95 or 99% confidence. As the sample size increases, however, the interval begins to be defined by a gradually diminishing percent of the total sample. With 9 coefficients we can say with 95% confidence that the median of the population is no less than the 1st coefficient in the series and no greater than the 8th, an interval which contains roughly 90% of our sample. With a large sample, such as the one presented in Figure 1, the interval is found to be between the 321st coefficient (.892) and the 374th, (.906) an interval which contains only 8% of the coefficients in that sample.

With a sample size of 6 through 8 the table gives an upper limit only, due to the fact that the interval which estimates the median with 95% confidence shrinks rather slowly at first so that only one coefficient can be dropped at a time, and the scientific convention of conservatism dictates that it should be at the top of the interval since we would usually like to see the highest possible median estimated. For sample sizes of 6 through 8, then, we say simply that the population median is probably less than our largest score. Conover (1971) calls this a "one-sided interval".

Notice also that the range of the coefficients determines the sample size needed. If, for instance we choose the .01 level and we have a sample size of 12, the table tells us that the median coefficient is probably (99% confidence) less than our highest coefficient and greater than our lowest. If the highest is .99 and our lowest is .97, the estimate would satisfy most people. If, on the other hand, the lowest coefficient is .47, then our estimated median would be greater than .47 and less than .99, which doesn't tell us very much at all. As we increase the sample size, however, we are allowed to exclude that .47 from the interval that estimates the median and, assuming that .47 is an unusual, low score, the lower limit of the interval might rise dramatically and end up estimating an acceptable median as the second (with an N of 15) or tihrd (N = 18) coefficient in the ascending series becomes the lower end of the interval. In general, the larger the sample the proportionately smaller the interval, and, of course, the better the estimate.

The widespread reporting of confidence intervals for reability coefficients can give the researcher a way of judging the size of the sample of coefficients that must be drawn to satisfy his or her own criterion for a reliable data set, and help curtail the use of very small samples of coefficients which are poor estimators of precision and accuracy.

REFERENCES

Blalock, H. M. *Social Statistics*. New York: McGraw-Hill, 1960.

Conover, W. J. *Practical Nonparametric Statistics*. New York: Wiley, 1971.

Documenta Geigy. *Scientific Tables, 7th edition*. Basle, Switzerland: Ciba-Geigy Limited, 1970.

Edgington, E. S. *Statistical Inference: the Distribution-Free Approach*. New York: McGraw-Hill, 1969.

Meighan, M. J., Burgess, R. L., & Lovitt, T. C. Matching and exchange in long-term conversational rhythms. Unpublished manuscript, University of Kansas Medical Center, 1977.

Stallings, W. M., & Gillmore, G. M. A note on "accuracy" and "precision". *Journal of Educational Measurement*, 1971, *8*, 127-129.

Topping, J. *Errors of Observation and Their Treatment*. London: Latimer Trend and Company Limited, 1962.