DOCUMENT RESUME

ED 079 322 TM 002 920

AUTHOR Swan, William W.
TITLE Four Indices for Investigating Inter-Observer Accuracy of Observational Instruments.
INSTITUTION Rutland Center, Athens, Ga.
SPONS AGENCY Bureau of Education for the Handicapped (DHEW/OE), Washington, D.C.
PUB DATE 73
NOTE 16p.; Paper presented at a joint session of the Annual Meeting of the American Educational Research Association and the National Council for Measurement in Education (New Orleans, Louisiana, February 25-March 1, 1973)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Data Analysis; *Mathematical Applications; *Measurement Instruments; *Observation; Speeches; *Statistical Analysis

ABSTRACT

Four indices for investigating inter-observer accuracy in observational instruments (contingency coefficient, Scott's pi, Bernstein's coefficient, and percent agreement) are reviewed concerning their assumptions, formulation, and tables indicating numerical functioning. Three of the four indices (excluding the contingency coefficient) are compared by computing each for four sets of observational data. It was found that Bernstein's coefficient had the highest median and the smallest range, percent agreement the second highest median and the second smallest range, and Scott's pi the lowest median and the largest range. It is hoped that authors will employ this information in their practical application and interpretation of these indices. (Author)

Four Indices for Investigating Inter-Observer Accuracy of Observational Instruments[1]

William W. Swan[2]

Rutland Center, Athens, Georgia*

Review of Indices

Four indices for investigating inter-observer accuracy (agreement between a criterion observer and another observer or two observers) of observational instruments have been selected. The contingency coefficient (C) is often used for determining the relationship between two nominal variables and is based on the results of a chi-square test of independence. The second coefficient is percent of inter-observer agreement (P); this coefficient is an input for the following two coefficients. The third, Bernstein's (1968) coefficient (P_b), has been chosen as the assumptions provided in its derivation are generally applicable for examining inter-observer accuracy of instruments. And the fourth, Scott's (1955) pi, was chosen because of its traditional usage as an accuracy index for observational data.

The formulation for each of the coefficients is presented in Table 1. The calculation of the contingency coefficient is based on the results of a chi-square test of independence (See Table 1). The two assumptions for X^2, and thus the contingency coefficient, are that X^2 be used with nominal or classification data and that the categories for X^2 be mutually exclusive. Assuming the use of nominal data with expected cell entries greater than five, the contingency coefficient is also restricted by the size of the array (number of rows and/or columns). The computation of C_{max} and subsequent comparison of C to C_{max} ($C/C_{(max)}$) gives a corrected estimate of

* Presented to joint Division D, AERA, and NCME critique session at Annual AERA Meeting, New Orleans, 1973

relationship of the two classification variables based on the size of the array. Garrett (1967, p. 395) provides information as to the relationship of the contingency coefficient and the product-moment correlation coefficient. The value of C ranges from 0 to 1.00.

Insert Table 1 about here

Percent agreement assumes the use of mutually exclusive categories, as does the contingency coefficient; the calculation for percent agreement is presented in Table 1. Two sets of criteria for determining percent of agreement between two observers were employed: 1. Same category at same time (C); 2. Same category at same time in same who-to-whom column (E). The range of values for percent agreement is from 0 to 100%.

An abbreviated form of the derivation of Bernstein's (1968) coefficient is presented in Table 2. The range of possible values for Bernstein's coefficient is presented in Table 3 for .05 intervals of P beginning at P = .51 (Bernstein assumes that P is no less than .51, thus any value of P less than .51 results in P_b equal to .00).

Insert Tables 2 & 3 about here

The calculation of Scott's (1955) pi, often cited in the literature as an observational instrument inter-observer accuracy coefficient, is presented in Table 1. P_e is dependent on the number of categories employed.

TABLE 1

Formulation of Coefficients

1. Contingency Coefficients:

$$C = \sqrt{\frac{X^2}{N+X^2}}$$

$$X^2 = \sum \frac{(O-E)^2}{E}$$ O = Observed Frequencies
E = Expected Frequencies

N = Total Number of Observations &

$$C_{max} = \sqrt{\frac{k-1}{k}}$$ where k = # of arrays, either columns or rows

2. Percent of Agreement:

$$P = \frac{\text{Number of Agreements}}{\text{Total Number of Possible Agreements}}$$

P_E = Exact percent of agreement, i.e., two observers recording the same category at same time in same who-to-whom column is the criterion for number of agreements.

P_C = Column-time percent of agreement, i.e., two observers recording the same category at same time is the criterion for number of agreements.

3. Bernstein's P_b:

$$P_b = \frac{1 + \sqrt{2A - 1}}{2}$$

A = P_E or P_C as defined above.

4. Scott's <u>pi</u>:

$$pi = \frac{P - P_e}{1 - P_e}$$

P = P_E or P_C as defined above.

$$P_e = \sum_{i=1}^{k} P_i^2$$, where k is the total number of categories used and P_i is the proportion of the entire sample which falls in the i^{th} category.

TABLE 2

Derivation of Bernstein's Coefficient*

Definitions:

Px = probability that coder X will correctly code a given item
Py = probability that coder Y will correctly code a given item

A = Ratio (percent) of agreement in the set of paired codes derived from matching the codings.

Now: $Qx = 1\text{-}Px$ $\qquad$ $Qy = 1\text{-}Py$

Assumptions:

1. Px and Py are constant and independent.
2. The number of categories is constant.

The probabilities associated with the possible outcomes for X and Y are given by: $(Px + Qx)\ (Py + Qy) = PxPy + QxPy + QyPx + QxQy$

or

Outcomes	Prob.	Nature of Agreement and Disagreement
X and Y correct	PxPy	X and Y agree
X correct, Y incorrect	PxQy	X and Y disagree
X incorrect, Y correct	QxPy	X and Y disagree
X and Y correct	QxQy	X and Y agree on the same incorrect code or X and Y disagree, but both select an incorrect code

We can now see that: $A = PxPy + QxQyK$

where K = fraction of events in the set associated with the probability QxQy, for which X and Y have selected the same incorrect code. K can be estimated by a variety of assumptions.

Now: $\qquad A = PxPy + QxQyK$ can be written as

$$A = PxPy + (1\text{-}Px)\ (1\text{-}Py)K$$

If the two coders X and Y are properly trained, it is reasonable to assume that $Px = Py = P$

$$A = P^2 + (1\text{-}P)^2K \quad \text{or} \quad P^2 + K\text{-}2PK + P^2K = A$$

Solution of this quadratic gives

$$P = \frac{K \pm \sqrt{A\ (1+K)\text{-}K}}{1 + K}$$

TABLE 2 (continued)

The restriction of $A \geq 1/2$ and $P > 1/2$ seems reasonable in situations in which percents of agreement are employed (for example, the investigation of inter-observer accuracy of observational instruments). With these restrictions and with $0 \leq K \leq 1$, the smaller quadratic root

$$P = \frac{K - \sqrt{A\,(1 + K) - K}}{1 + K}$$

is excluded since the largest possible $P = \frac{K}{1 + K}$ which is less than or equal to 1/2, is attained only when

$$A = \frac{K}{1 + K} \leq 1/2$$

Using the larger quadratic root

$$P = \frac{K + \sqrt{A\,(1 + K) - K}}{1 + K}$$ the values of P can be calculated.

The extreme values of $K = 1$ and $K = 0$ lead to slightly different values from each other until A is as low as .70.

$K = 1$** and thus

$$P = \frac{1 + \sqrt{2A - 1}}{2}$$

which is the formulation of Bernstein's (1968) coefficient used in this paper.

---

*This is abstrated from Bernstein (1968). The complete derivation may be obtained from the previous reference.

**K chosen equal to 0 gives slightly different results.

TABLE 3

Value of Bernstein's P_b[a]

Percent Agreement	P_b
100%	1.00
95%	.97
90%	.95
85%	.92
80%	.89
75%	.85
70%	.82
65%	.77
60%	.72
55%	.65
51%	.57

[a]Assumes Percent of Agreement is greater than or equal to .51

The error term (P_e) for pi increases as the number of categories used during any one observation period decreases. Those categories used most often get a disproportionately higher weighting in the error term because of the nature of squaring decimals, i.e. . . . $.10^2 = .01$, $.40^2 = .16$. The possible values for pi are presented in Table 4 for intervals of .05 for both P and P_e. Values of P_e are located on the top horizontal margin of the matrix, and values of P are located on the left vertical margin of the matrix. A generally accepted lower limit for accuracy is approximately .70. The heavy line in Table 4 indicates that portion of the matrix which contains positive values of pi greater than or equal to .70. The least value of P which provides a pi value greater than .70 is P = .75. And in this case, P_e equals .15 or less. For example, assuming one has 10 categories in his coding system, no particular category could be employed 40% of the time and few categories could be employed 20% of the time, the remainder being employed 10% or less, if one wanted to obtain a pi equal to .70 or greater. This additionally assumes that the percent of agreement is .75 or greater.

Insert Table 4 about here

Comparison of Indices

The contingency coefficient was not used for the comparisons. Garrett (1967, p. 258) states that the expected value of entries in the cells of the contingency table should be five or greater. In this case, using the observational instrument, one effectively is dealing with a 26 x 26 contingency table (26 total categories of the observational instrument used in this case with one dimension for each of two observers, each observation being composed of approximately twenty recordings). Thus, one would appro-

TABLE 4

Coefficient Pi[1]

Matrix of Possible Coefficients at Increments of .05 for P_o and P_e

P \ Pe	1.00	.95	.90	.85	.80	.75	.70	.65	.60	.55	.50	.45	.40	.35	.30	.25	.20	.15	.10	.05
1.00	0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.95	*	0	.50	.67	.75	.80	.83	.86	.88	.89	.90	.91	.92	.92	.93	.93	.94	.94	.94	.95
.90	*	-1.00	0	.33	.50	.60	.67	.71	.75	.78	.80	.82	.83	.85	.86	.87	.88	.88	.89	.90
.85	*	-2.00	-.50	0	.25	.40	.50	.57	.63	.67	.70	.73	.75	.77	.79	.80	.81	.82	.83	.84
.80	*	-3.00	-1.00	-.33	0	.20	.33	.43	.50	.56	.60	.64	.67	.69	.71	.73	.75	.77	.78	.79
.75	*	-4.00	-1.50	-.67	-.25	0	.17	.29	.38	.44	.50	.55	.58	.62	.64	.67	.69	.71	.72	.74
.70	*	-5.00	-2.00	-1.00	-.50	-.20	0	.14	.25	.33	.40	.46	.50	.54	.57	.60	.63	.65	.67	.68
.65	*	-6.00	-2.50	-1.33	-.75	-.40	-.17	0	.13	.22	.30	.36	.42	.46	.50	.53	.56	.59	.61	.63
.60	*	-7.00	-3.00	-1.67	-1.00	-.60	-.33	-.14	0	.11	.20	.27	.33	.39	.43	.47	.50	.53	.56	.58
.55	*	-8.00	-3.50	-2.00	-1.25	-.80	-.50	-.29	-.13	0	.10	.18	.25	.31	.36	.40	.44	.47	.50	.53
.50	*	-9.00	-4.00	-2.33	-1.50	-1.00	-.67	-.43	-.25	-.11	0	.09	.17	.23	.29	.33	.38	.41	.44	.47
.45	*	-10.00	-4.50	-2.67	-1.75	-1.20	-.83	-.57	-.38	-.22	-.10	0	.08	.15	.21	.27	.31	.35	.39	.42
.40	*	-11.00	-5.00	-3.00	-2.00	-1.40	-1.00	-.71	-.50	-.33	-.20	-.09	0	.08	.14	.20	.25	.29	.33	.37
.35	*	-12.00	-5.50	-3.33	-2.25	-1.60	-1.17	-.86	-.63	-.44	-.30	-.18	-.08	0	.07	.13	.19	.24	.28	.32
.30	*	-13.00	-6.00	-3.67	-2.50	-1.80	-1.33	-1.00	-.75	-.56	-.40	-.27	-.17	-.08	0	.07	.13	.18	.22	.26
.25	*	-14.00	-6.50	-4.00	-2.75	-2.00	-1.50	-1.14	-.88	-.67	-.50	-.36	-.25	-.15	-.07	0	.06	.12	.17	.21
.20	*	-15.00	-7.00	-4.33	-3.00	-2.20	-1.67	-1.29	-1.00	-.78	-.60	-.46	-.33	-.23	-.14	-.07	0	.06	.11	.16
.15	*	-16.00	-7.50	-4.67	-3.25	-2.40	-1.83	-1.43	-1.13	-.89	-.70	-.55	-.42	-.31	-.21	-.13	-.06	0	.06	.11
.10	*	-17.00	-8.00	-5.00	-3.50	-2.60	-2.00	-1.57	-1.25	-1.00	-.80	-.64	-.50	-.39	-.29	-.20	-.13	-.06	0	.05
.05	*	-18.00	-8.50	-5.33	-3.75	-2.80	-2.17	-1.71	-1.38	-1.11	-.90	-.73	-.58	-.46	-.36	-.27	-.19	-.12	-.06	0

Note.-- * = undefined

[1] Appreciation is expressed to R. Gregory Litaker for his obtaining these calculations.

priately place twenty recordings in a 26 x 26 matrix. Collapsing levels of the classification are not possible as it would be extremely difficult, at best, to interpret the results of such a procedure. Additionally, one would need to collapse to a 2 x 2 table to obtain an expected value of five entries per cell, i.e., four cells with five entries each = 20. It is not reasonably possible then to obtain an expected entry of five per cell and thus not possible to obtain an estimate of the relationship of inter-observer agreement.

The data used for this comparison were obtained by employing Systematic Who-to-Whom Analysis Notation (Swan, 1971), which is an observational instrument based on the overt behavioral components of the representative objectives of Developmental Therapy (Wood, 1972), a treatment approach for emotionally disturbed children. The instrument is composed of eight major and sixteen minor categories (a total of 24 categories) based on various subsets of the Developmental Therapy objectives. The basic outline of the system is shown in Table 5. One category is recorded every three seconds in the appropriate who-to-whom column of the who-to-whom observation sheet and each observation period is approximately one minute in length.

Insert Table 5 about here

Four sets of observational data were obtained for the comparison of the indices. Each set is from a SWAN criterion training session (composed of video-tapes). For each set of tapes there are three coefficients, (P, P_b, *pi*) each computed for the two sets of criteria for observer agree-

TABLE 5

Systematic Who-to-Whom Analysis Notation

(SWAN)

1. OBSERVERS	O
In response to child's name being called	ON
Observes one who is talking	OT
2. PHYSICAL CONTACT	C
Inappropriate	C-
Receives	CR
3. FOLLOWS DIRECTIONS	F
Does not follow directions	F-
4. WORKS	W
Works, but not appropriately sitting	W-
5. VERBALIZES	V
Inappropriate	V-
Non-understandable	VN
I-statement	VI
Group rules	VG
In response	VR
6. PHYSICAL ACTIVITY	A
Inappropriate	A-
Parallel play	P+
Play	P
7. RESPONDING ACTIVITY	RA
8. NON-DIRECTED ACTIVITY	N
Removal from view	/
Removal from view by teacher	//

ment E and C (See Table 1), and each computed for each pair of observers. Thus, there are six coefficients each with a median and a range as shown in Tables 6, 7, 8, and 9.

Insert Tables 6, 7, 8, and 9 here

Discussion

The medians of all three coefficients for all four sets of data for the C condition are slightly higher than or equal to those for the E condition. This is expected as the more stringent the criteria for agreement the less the number of agreements. The ranges of all three coefficients for all four sets of data for the C condition are slightly smaller than or equal to those for the E condition and this is expected as per the same rationale.

For both conditions, and all four sets of data (except for one case), Bernstein's coefficient has the highest median and the smallest range; and _pi_ has the lowest median and the lowest range. The exception to this statement occurs in Table 9 where the ranges of Bernstein's coefficient and _pi_ are identical. This occurs because percent of agreement for one case of inter-observer agreement was less than 51% and this results in P_b equal to .00. The relationship for the medians holds for this case. Slight variations exist between the sets of data with respect to the size of the differences between the medians and the ranges.

TABLE 6

Data Set I

Inter-Observer Reliability Coeffici ..[b]

	Percent Agreement		Bernstein's P_b		Scott's Pi	
	E	C	E	C	E	C
Range	55-95	65-95	.67-.97	.77-.97	.32-93	.44.-.93
Median	80	85	.89	.92	.66	.71

[b]Based on three observers reviewing seven tapes producing 21 estimates of each coefficient for each condition.

TABLE 7

Data Set II

Inter-Observer Accuracy Coefficients[b]

	Percent Agreement		Bernstein's P_b		Scott's Pi	
	E	C	E	C	E	C
Range	55-95	60-95	.57-.97	.72-.97	.00-.91	.10-.91
Median	75	76	.85	.86	.43	.48

[b]Based on three observers reviewing seven tapes producing 21 estimates of each coefficient for each condition.

TABLE 8

Data Set III

Inter-Observer Accuracy Coefficients[b]

	Percent Agreement		Bernstein's P_b		Scott's *Pi*	
	E	C	E	C	E	C
Range	71-100	75-100	.83-1.00	.85-1.00	.31-1.00	.54-1.00
Median	89	89	.94	.95	.76	.82

[b]Based on three observers reviewing 12 tapes producing 36 estimates of each coefficient for each condition.

TABLE 9

Data Set IV

Inter-Observer Accuracy Coefficients[b]

	Percent Agreement		Bernstein's P_b		Scott's *Pi*	
	E	C	E	C	E	C
Range	40-100	50-100	.00-1.00	.00-1.00	.00-1.00	.00-1.00
Median	85	85	.92	.92	.59	.63

[b]Based on three observers reviewing seven tapes producing 21 estimates of each coefficient for each condition.

Conclusions

The educational importance of this study concerns the practical application of these indices. If the assumptions are satisfied for a particular coefficient, the user must be aware of the nature of the coefficient and its behavior in order to interpret his results for the reader. It is particularly in the area of inter-observer accuracy that the user is often simply looking for some index, and the results are often presented without being interpreted for the reader. It is the writer's responsibility to interpret these values to his readers, either in terms of significance levels or in terms of the functioning of the index.

One would for example interpret a resulting inter-observer figure of .75 differently depending on whether it is a Bernstein's (1968) coefficient or a Scott's (1955) *pi*. If it is a Bernstein's (1968) coefficient, the .75 is not extremely large, while if it is a Scott's *pi*, the .75 is very large. If the sample size of recordings is large enough, and the .75 represents a calculated contingency coefficient, one would need to compare such to C_{max}.

Thus, those individuals who use a specific index should be aware of the variability of the specific index used and the functioning of that index and its range in order to enable them to interpret more clearly the degree of inter-observer accuracy with respect to the constraints implied by the index.

FOOTNOTES

1. The data were collected in part through a special project grant from the U.S. Office of Education, Bureau of Education for the Handicapped, under the Children's Early Education Assistance Act, P. O. 91-230, Part C, formerly 90-538.

2. Dr. Linwood E. Tisdel, Mr. R. Gregory Litaker, Mrs. Mary M. Beussee, and Mrs. Faye Swindle are acknowledged for their careful readings of an earlier draft of this manuscript.

REFERENCES

Bernstein, A. L. An estimate of the accuracy (objectivity) of nominal category coding. MOREL Monograph Series: No. 1, 1968. Detroit: Michigan-Ohio Regional Educational Laboratory.

Garrett, H. E. Statistics in psychology in education, David McKay Company, Inc., New York, 1967, p. 258, 395.

Scott, W. A. Reliability of content analysis: the case of nominal scale coding. Public Opinion Quarterly, 1955, 19, 321-325.

Swan, W. W. The development of an observational instrument based on the objectives of Developmental Therapy. Unpublished doctoral dissertation, College of Education, University of Georgia, Athens, 1971.

Wood, Mary M. The Rutland Center model for treating emotionally disturbed children, Rutland Center, Athens, Georgia, 1972. Chapter 4.