Bulletin of the Psychonomic Society
1982, Vol. 20(4), 187-189

Deviance probabilities: Determination of judgmental bias within Kendall's coefficient of concordance data

L. W. BUCKALEW and W. H. PEARSON
Air Force Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, Ohio 45433

Kendall's coefficient of concordance is reviewed, with particular concern for nonsignificance. A statistic is presented that allows determination of whether one judge's rankings come from the same population as the other judges' rankings. This is Σ/D/, the sum of absolute differences in ranks that N objects receive from any two judges. Frequency distributions of Σ/D/, computed for N = 3-10 objects, were used to obtain the probability of a given /D/'s being greater than a certain constant under the null hypothesis. This cumulative probability density function for incremental Σ/D/s is tabled. Obtained probabilities that one judge's ranking came from the same population as all other judges' rankings may be calculated from Bayes' theorem. The statistic Σ/D/, associated probabilities, and comparison of one judge's rankings with collective rankings are based on the raw data of Kendall's W and operationally allow identification of judgmental bias.

Psychological research efforts often employ comparative rating or appraisal techniques, particularly in psychophysics, industrial, applied, and social psychology, and human engineering. Such techniques typically involve ordinal measurements that rely on nonparametric analyses. Acknowledging the power, necessity, and appropriateness of these analyses, it is not uncommon to experience frustration with the comparatively restrictive interpretation potentials afforded. For nominal data analysis involving chi square, Buckalew and Pearson (1981, 1982) offered extensions of traditional techniques allowing exploratory but more specific interpretation. The purpose of this note is to offer a technique whereby the data of a popular and utilitarian ordinal statistic may be used to facilitate more specific interpretation and increase the information available.

Kendall's coefficient of concordance (W) measures the relation among several ($k > 2$) rankings of N objects, events, or individuals (Ferguson, 1980; Siegel, 1956). It expresses the degree of association among rankings such that a high value of W may be interpreted as meaning that the observers or judges applied essentially similar standards in ranking the N objects. A low value of W suggests dissimilarity in rankings, also interpreted in terms of the substrate of standards. In essence, W constitutes a measure of interjudge or intertest reliability.

W may only assume positive values, with perfect agreement among judges expressed as 1 and maximum disagreement as 0. As noted by Ferguson (1980), W does not take negative values because with more than two judges, complete disagreement cannot occur. Conceptually, W is related to the mean of rank-order correlation coefficients computed between all possible pairs of ranks. Critical values of W depend on the number of sets of ranks (judges) and the number of ranks in each set (objects). For situations in which $N > 7$, W may be converted to a chi-square value to test its significance, and such texts as Siegel (1956) offer tabled values of W required for significance when $N \leqslant 7$.

Traditional interpretation of W is restricted to a collective consideration of rankings (judges). While this may be desirable in many analytic situations and is sufficient for interpretation of a significant W, the utilitarian and informational value of a nonsignificant W is minimal. Given that lack of sufficient agreement among judges exists, consideration of causation may be desirable. There is the possibility that, with identification of a biased judge or observer and removal thereof, the remaining rankings would be more concordant. Biased judgment of a single observer, relative to the collective judgments of all other observers, is the interpretive substrate of consideration.

What is offered is an analytic technique, using data of the traditional computation of W, to identify a source of nonsignificance (i.e., potential bias). The initial conceptual question is do the ratings of two or more judges come from the same population of ratings of N objects and, if not, how different are they? If judges are responding similarly to the same characteristics, the sum of absolute differences (Σ/D/) between the ratings of any two judges should be minimal, preferably 0. As ratings' similarity decreases, Σ/D/ becomes larger. In effect, Σ/D/ is assumed to be a metric for interjudge differences.

Given this framework, it becomes meaningful to know the likelihood of a Σ/D/. Assuming that one set

Requests for reprints may be addressed to either L. W. Buckalew, Department of Psychology, Box 200, Alabama A & M University, Normal, Alabama 35762, or W. H. Pearson, AFAMRL/HEG, Wright-Patterson Air Force Base, Ohio 45433.

0090-5054/82/100187-03$00.55/0

of ratings is as likely as another, one can obtain a probability density function of $\Sigma/D/$ for a given N. From this distribution, the probability of obtaining a $\Sigma/D/$ equal to or greater than a certain number can be ascertained. Operationally, the probability that ratings of one judge came from the same population as those of all other judges may now be calculated. Using Bayes' theorem (Uspensky, 1937), one may calculate the cumulative probability, incremented by all judges' ratings in turn, that one set of ratings is from the same population as the collective other sets.

For any given N, or number of things to be rated, the total number of permutations, or arrangements in order, is N factorial (N!). If each possible permutation, without replacement, duplication, or omission, or a ratings arrangement is printed out, each rank in the sequence could be subtracted from its positional counterpart in a standard arrangement (1,2,3, . . . N) of ranks. The absolute differences may be summed to provide $\Sigma/D/$. The values of $\Sigma/D/$s for all possible sequences of ranks may be plotted in a frequency distribution reflecting the number of $\Sigma/D/$s of each value. The cumulative proportion of the $\Sigma/D/$s less than or equal to each increasing value of $\Sigma/D/$ (a probability distribution function) was computed and is presented as Table 1.[1]

The probability of evidencing a particular $\Sigma/D/$ or greater between any two judges' rankings, given N objects ranked, may be obtained by subtracting the tabled value for that $\Sigma/D/ - 2$ for a given N from the quantity 1.000. For example, using values of Table 1, it may be seen that the probability of obtaining a $\Sigma/D/ \geqslant 10$ between two judges, when $N = 5$, would be $1.000 - .633 = .367$, and that of obtaining a $\Sigma/D/ \geqslant 12$, when $N = 5$, is $1.000 - .833 = .167$. In this fashion, the probability of evidencing any particular $\Sigma/D/$ resulting from a comparison of any two judges' rankings for situations in which $N \leqslant 10$ may be readily obtained from the values provided in Table 1. These probabilities may then be interpreted in terms of whether any given pair of judges, considering their rankings, came from the same population.

Table 1
Cumulative Probability Density Function for $\Sigma/D/$s

	Number of Objects Ranked							
$\Sigma/D/$	3	4	5	6	7	8	9	10
0	.167	.042	.008	.001	.000	.000	.000	.000
2	.500	.167	.042	.008	.001	.000	.000	.000
4	1.000	.458	.142	.033	.006	.001	.000	.000
6		.833	.342	.097	.021	.004	.000	.000
8		1.000	.633	.226	.059	.012	.002	.000
10			.833	.417	.131	.031	.006	.000
12			1.000	.622	.248	.069	.015	.003
14				.811	.396	.132	.033	.006
16				.955	.571	.224	.064	.014
18				1.000	.737	.344	.112	.028
20					.879	.486	.183	.051
22					.950	.632	.274	.086
24					1.000	.767	.386	.136
26						.871	.508	.203
28						.942	.633	.286
30						.986	.745	.383
32						1.000	.843	.488
34							.914	.596
36							.963	.699
38							.986	.790
40							1.000	.866
42								.923
44								.961
46								.983
48								.996
50								1.000

Table 2
Hypothetical Rankings by Three Judges

	Objects						
Judge	1	2	3	4	5	$\Sigma/D/$	ρH_0
A	3	1	2	5	4	12	.167
B	3	2	1	5	4	10	.367
C	2	4	5	1	3		

Table 1 provides probability values only for comparison of the rankings of one judge with those of another judge, through consideration of the statistic $\Sigma/D/$. This procedure may be extended to allow comparison of a single judge's rankings with those of all other judges collectively. Given three judges (A, B, and C), Bayes' theorem may be applied to ascertain the probability that Judge C's rankings belong to the same population as Judge A's and Judge B's. The null hypothesis (H_0) is that Judge C comes from the same population as Judges A and B. The alternate hypothesis (H_1) is that he does not. Table 2 gives the three judges' rankings for $N = 5$ objects. $\Sigma/D/$ is computed for a comparison of Judge C with Judge A and Judge C with Judge B. For each $\Sigma/D/$, probability values are obtained using Table 1. These constitute probabilities of $\Sigma/D/$ expected under H_1, with 1.000 minus each of these reflecting the probability of these data ($\Sigma/D/$) under H_0.

To this scenario may be applied a variation of Fisher's method of randomization, as described by Bradley (1960). Assume that several samples (A, B, and C) equally likely under H_0 have been drawn from a population. Values of a statistic ($\Sigma/D/$) sensitive to H_1 have been calculated for all possible random samples. The rejection region for H_0 is the most extreme of these values, which becomes more probable when H_1 is true. The probabilities for H_0 and H_1 are then computed. For the Judge C vs. A comparison, the probability of the $\Sigma/D/$ (12) under H_1 is .833 (tabled), and that under H_0 is .167 (1.000 – .833). For the Judge C vs. B comparison, the probability of the $\Sigma/D/$ (10) under H_1 is .633 (tabled), and that under H_0 is .367 (1.000 – .633). These probabilities may be used to obtain the probability that Judge C came from the same population as did Judges A and B, using the data of Judge A as a priori probabilities: $[(\rho H_0 A,C)\ (\rho H_0 B,C)] / [(\rho H_0 A,C)$

(ρH_0 B,C) + (ρH_1 A,C) (ρH_1 B,C)] = ρH_0 A,B,C. Applying the obtained H_0 and H_1 probabilities for the Judges A and C and Judges B and C comparisons yields: [(.167) (.367)] / [(.167) (.367) + (.833) (.633)] = .061/(.061 + .527) = .061/.588 = .104. Hence, the probability that Judge C came from the same population as Judges A and B is .10.

To add the data of a fourth judge (D) to the present data, the probabilities for Judges A and B together are used as the a priori probabilities: [(ρH_0 A,B,C) (ρH_0 D,C)] / [(ρH_0 A,B,C) (ρH_0 D,C) + (ρH_1 A,B,C) (ρH_1 D,C)] = ρH_0 A,B,C,D. Additional judges' data may be incremented in like manner.

This method, through consideration of Σ/D/s and associated probabilities as tabled, will identify judges who are rating on a different basis from their fellow judges, both on a paired comparison and on a collective basis. As an ordinal technique, the influence of data extremes has been avoided by using ranks, and samples have been made more nearly comparable. Given the simple calculations of Σ/D/s between two judges' rankings and the conversion of these data to probabilities provided in Table 1, the determination of judgmental bias is readily made. However, as may be noted in Table 1, the rejection region is truncated, and not too sensitive to H_1. This would suggest that sensitivity would increase with the amount of data (i.e., number of judges). The virtue of this technique is its simplicity in calculation and transformation, as other statistics could be used for the same purpose. For example, all the intercorrelations among judges' rankings could be computed and relationships so described. These correlations could then be tested for significance. Comparatively, this would be laborious, and the proposed technique is simple, direct, and yields a probability.

This technique provides probabilistic values for a single judge's deviance from any other judge or from the collective rankings. Its primary intended use is for situations in which the initial value of W was not appreciably less than that required for significance. No criterion is offered for designating a judge as biased, based on the deviance probabilities provided, nor is a criterion offered for the decision to remove a judge and recompute W. The significance level needed for rejecting H_0 (i.e., that the judges' rankings all came from the same population) is left to the discretion of the researcher.

REFERENCES

Bradley, J. V. *Distribution-free statistical tests* (WADD Tech. Rep. 60-661). Wright-Patterson AFB, Ohio: Aerospace Medical Division, August 1960.

Buckalew, L. W., & Pearson, W. H. Determination of critical observed frequencies in chi square. *Bulletin of the Psychonomic Society,* 1981, **18**, 289-290.

Buckalew, L. W., & Pearson, W. H. Critical factors in the chi square test of independence: A technique for exploratory data analysis. *Bulletin of the Psychonomic Society,* 1982, **19**, 225-226.

Ferguson, G. A. *Statistical analysis in psychology and education* (5th ed.). New York: McGraw-Hill, 1980.

Siegel, S. *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill, 1956.

Uspensky, J. V. *Introduction to mathematical probability.* New York: McGraw-Hill, 1937.

NOTE

[1] The computer program that generated these permutations and associated probabilities was written by and is available from W. H. Pearson.

(Received for publication August 6, 1982.)