PSYCHOMETRIKA—VOL. 44, NO. 4.
DECEMBER, 1979

RAMIFICATIONS OF A POPULATION MODEL FOR κ AS A COEFFICIENT OF RELIABILITY

HELENA CHMURA KRAEMER

STANFORD UNIVERSITY

Coefficient κ is generally defined in terms of procedures of computation rather than in terms of a population. Here a population definition is proposed. On this basis, the interpretation of κ as a measure of diagnostic reliability in characterizing an individual, and the effect of reliability, as measured by κ, on estimation bias, precision, and test power are examined. Factors influencing the magnitude of κ are identified. Strategies to improve reliability are proposed, including that of combining multiple unreliable diagnoses.

Key words: kappa, reliability, Spearman-Brown formula.

Unreliability of observations is a matter of import both in clinical contexts, in which an observation is to be used to assess and make decisions about an individual subject, and in a research context, in which observations are to be used to assess and make decisions about populations. The classical coefficient of reliability for quantitative data, the intraclass correlation coefficient, ρ, is defined on the basis of a population model of an observation as an additive function of information and independent error [e.g., Cronbach, 1972]. On the basis of this model, ρ is interpretable as a measure of agreement between multiple observations of a single subject, and as a measure of the quality of an observation in characterizing a subject. As such, it is a reflection of the clinical value of such an observation. At the same time, the effect of unreliability on the precision of estimators of population parameters or on power of statistical test procedures can be explicitly and quantitatively related to ρ. Thus ρ is readily used as a basis of research design considerations.

Because ρ is explicitly defined in terms of certain population parameters, it is possible to identify factors which influence its magnitude. On this basis, strategies to increase reliability of a single measurement may be evolved. Finally, it is possible to compute the reliability of the average of multiple independent observations per subject from the reliability of a single such measurement [Spearman, 1910; Brown, 1910]. If the reliability of a single measurement is not sufficient for clinical and research applications, one may base on such considerations a decision to use multiple observations effectively.

The coefficient kappa, κ, [Cohen, 1960; Everitt, 1968; Fleiss, 1973; Fleiss, et. al., 1969; Light, 1971; Hubert, 1977] is proposed for categorical response data where such response is the assignment of the subject to one of k mutually exclusive and exhaustive categories. In contrast to ρ, it is generally defined not in terms of a population model, but in terms of the procedures used to compute the statistic, as that parameter of which the statistic is a consistent estimator [Fleiss, 1975]. By its definition, it is readily interpretable as a measure of agreement between multiple observations of a single subject. Its value as a measure of the quality of an observation in clinical or research contexts is not clear. Furthermore, it is

This investigation was supported in part by the National Institute of Mental Health Specialized Research Center Grant # MH-30854.

Requests for reprints should be sent to the author, Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA 94305.

0033-3123/79/1200-2818$00.75/0

difficult to evolve specific strategies to improve the reliability of measurement without a definition of the population characteristic κ represents.

A population model for κ will be proposed analogous to that for ρ but appropriate to categorical response measures. On the basis of this model, it will be demonstrated that κ, like ρ, is interpretable as a measure of the quality of individual response, and therefore a measure of the clinical value of such an observation. A few common research designs will be examined as an illustration of procedures for evaluating the effects of unreliability as measured by κ on the precision of estimators and the power of tests. Kappa, like ρ, directly indicates the loss of precision or power of statistical procedures associated with observation error.

Finally the problem of improving reliability of a single observation or improving reliability by use of multiple observations is considered.

Population Models

In the classical model for reliability of interval data [Walker & Lev, 1953; Cronbach, 1972], an observation x_i is described as:

$$x_i = \xi_i + \epsilon_i$$

where ξ_i is Subject i's characteristic response, with

$$\mathscr{E}(\xi_i) = \mu, \ \text{var}(\xi_i) = \sigma_\xi^2,$$

and ϵ_i, the error of measurement, assumed to be independent of ξ_i. If multiple independent observations are drawn for Subject i from the population which a single observation is meant to represent (multiple simultaneous observations by different observers, or multiple test forms of instruments of measurement, or multiple observations over a span of time by one observer, etc.), errors are assumed to be independent with $\mathscr{E}(\epsilon_i) = 0$ and $\text{var}(\epsilon_i) = \sigma_\epsilon^2$. The reliability of the observations is defined as

$$\rho = \frac{\sigma_\xi^2}{\sigma_\xi^2 + \sigma_\epsilon^2}.$$

In contrast to this theoretical approach to the definition of ρ, κ is generally defined in terms of the computation procedure used for its estimation. If one draws two or more independent observations per subject from the population of observations to which a single observation is meant to generalize, then

$$\kappa = \frac{p_0 - p_c}{1 - p_c}$$

where p_0 is a measure of observed agreement and p_c is a measure of the agreement expected by chance [Cohen, 1960; Fleiss, 1973; Fleiss, et al., 1969; Light, 1971; Hubert, 1977]. It is clear that κ is defined to be a measure of agreement between multiple observations per subject, as is ρ.

Let us propose the following population model. The entire population of possible observations for subject i which a single observation is meant to represent is described by a vector $\overline{\Pi}_i = (\Pi_{i1}, \Pi_{i2}, \cdots, \Pi_{ik})$ where Π_{ij} is the proportion of observations classifying subject i's response in category j, $j = 1, 2, \cdots, k$, $\sum_j \Pi_{ij} = 1$. Multiple observations drawn with replacement from this population have a multinomial distribution. If one draws a random observation from this hypothetical population, the probability that subject i is classified in category j is Π_{ij}. Let

$$P_j = \mathscr{E}(\Pi_{ij}),\ (P'_j = 1 - P_j),$$
$$\sigma_j^2 = \text{variance } (\Pi_{ij}),$$
$$\beta_j = \mathscr{E}(\Pi_{ij} - P_j)^3,$$

where expectations are taken over the population of subjects. Then let

$$\kappa_j = \frac{\sigma_j^2}{P_j P'_j},$$

the kappa-coefficient for category of response j, and overall kappa be defined by:

$$\kappa = \frac{\sum_j \sigma_j^2}{\sum_j P_j P'_j}.$$

The definition is in accord with usual definitions of κ. For example, the probability of agreement between two independent observations for subject i is

$$\sum_j \Pi_{ij}^2$$

and hence the overall probability of agreement is

$$p_0 = \mathscr{E}\left(\sum_j \Pi_{ij}^2\right) = \sum_j (\sigma_j^2 + P_j^2),$$

and the agreement expected by chance is

$$p_c = \sum_j P_j^2.$$

Thus

$$\kappa = \frac{\sum_j \sigma_j^2}{1 - \sum P_j^2} = \frac{\sum_j \sigma_j^2}{\sum_j P_j P'_j} = \frac{\sum_j P_j P'_j \kappa_j}{\sum_j P_j P'_j}.$$

Overall κ then is a weighted sum of individual κ-coefficients obtained for one response category at a time, having value zero (or one) if and only if each individual κ_j has value zero (or one). Interpretation of the overall κ depends on interpretation of the meaning of individual κ-coefficients. Hence, hereafter, we will consider the problem of interpreting an individual κ-coefficient only, i.e., we consider a dichotomous response where Π_i is the risk of subject i being given a positive "diagnosis" (D+),

$$\mathscr{E}(\Pi_i) = P; \qquad \text{var}(\Pi_i) = \sigma^2; \qquad \mathscr{E}(\Pi_i - P)^3 = \beta; \qquad \kappa = \frac{\sigma^2}{PP'}.$$

In general, the nature of the distribution of Π_i is not known. However, interpretation and applications of κ do not require any identification of the specific distribution, beyond the requirement that the first three moments exist and that PP' exceeds zero. Consideration of tactics to identify sources of unreliability and therefore to improve reliability are facilitated by closer study of a few specific and important special cases.

Case 1: True Dichotomy

Suppose the population were a true dichotomy. A proportion P^* of subjects are in a subpopulation having $\Pi_i = p$; a proportion $P^{*\prime} = 1 - P^*$ having $\Pi_i = q' = 1 - q$. The parameters P^*, p and q are closely related to the parameters prevalence, sensitivity, and specificity as used in assessing the quality of dichotomous medical diagnostic procedures [Galen & Gambino, 1975]. In this case $P = P^*p + P^{*\prime}q'$, i.e., observed prevalence P is generally a biased estimator of true prevalence P^*. Also,

$$\mathscr{E}(\Pi_i^2) = P^*p^2 + P^{*\prime}q'^2$$

and hence

$$\begin{aligned}\operatorname{var}(\Pi_i) &= P^*p^2 + P^{*\prime}q'^2 - (P^*p + P^{*\prime}q')^2,\\ &= P^*P^{*\prime}p^2 + P^*P^{*\prime}q'^2 - 2P^*P^{*\prime}pq',\\ &= P^*P^{*\prime}(p + q - 1)^2,\end{aligned}$$

and

$$\kappa = \frac{P^*P^{*\prime}}{PP'}(p + q - 1)^2.$$

It is clear that $\kappa = 1$ if and only if p, $q = 1$, i.e., if there are no misclassification errors. Otherwise, κ reflects both prevalence (P^*) and classification errors. At $P^* = 0$, κ is 0, increasing thereafter to a maximum value of

$$\kappa_{\max} = \{(pq)^{1/2} - (p'q')^{1/2}\}^2$$

when $P^* = \sigma_q/(\sigma_q + \sigma_p)$, where $\sigma_p^2 = p(1 - p)$, $\sigma_q^2 = q(1 - q)$. Thereafter κ decreases to 0 at $P^* = 1$.

Case 2: Dichotomization of a Continuum

In Case 1 a dichotomous observation is used for a population of subjects which is indeed a dichotomy. However, it often happens that such an observation is used when the population is not, in fact, a dichotomy. For example, a diagnosis of hypertension indicates that blood pressure (an interval response) exceeds a criterion value: a positive diagnosis is attributed to patient i if

$$x_i \geq \mu + h\sigma_x,$$

where we will assume that x_i satisfies the assumptions underlying ρ. Therefore,

$$P = \Pr\{x_i \geq \mu + h\sigma_x\} = 1 - \Phi(h),$$

where Φ is the cumulative distribution function of the standard normal distribution. The relationship between κ for the diagnosis and ρ of the underlying measure is [Kirk, 1973]:

$$\kappa = \frac{1}{2\pi PP'}\int_0^{\rho} \frac{1}{(1 - x^2)^{1/2}} \exp\left(\frac{-h^2}{1 + x}\right) dx.$$

When $0 < \rho < 1$,

$$0 < \kappa \leq \frac{2}{\pi}\arcsin\rho < \rho,$$

with equality only if $h = 0$, $P = 0.5$. The divergence between κ and ρ increases as P approaches 0 to 1.

Case 3: Confounded Population and Error Characteristics

In both Cases 1 and 2 a distinction is made between characteristics of the population (prevalence P in Case 1, σ_ξ^2 in Case 2) and decision-making errors (p, q in Case 1, σ_ϵ^2 in Case 2). More generally, population characteristics and decision-making errors may be confounded. For example, the distribution of Π_i in the population may be described by a distribution of the beta type:

$$\frac{\Pi^{l-1}(1-\Pi)^{m-1}}{B(l, m)}, \qquad 0 \le \Pi \le 1, \infty > l, m > 0.$$

Then

$$P = \frac{l}{l+m},$$

$$\kappa = \frac{1}{l+m+1}.$$

In this case, κ cannot actually achieve the values of 1 or 0, and it is not possible to differentiate population characteristics from error characteristics.

Theoretically one may identify the appropriate model by obtaining a large number of independent observations per subject and examining the distribution of $\hat{\Pi}_i$, the observed proportion of positive diagnoses for each subject. In Case 1, the distribution will be bimodal with modes centered at p and q'; in Case 3, it will be approximately of the beta type. In Case 2,

$$1 - \hat{\Pi}_i \approx \Phi(\mu - \xi_i + h\sigma_x)$$

and hence $\Phi^{-1}(1 - \hat{\Pi}_i)$ is approximately normally distributed with mean $h\sigma_x$ and variance σ_ξ^2.

In practice however, it may be difficult to obtain a large enough number of independent assessments per subject to distinguish between, for example, Case 2 and Case 3 with l and m each exceeding 1, or between Case 1 and Case 3 with l or $m < 1$. As noted above, this does not comprise the interpretation of the population kappa as it affects clinical decision making or research design considerations. Except as these special cases are used to illustrate certain points, the principles which follow are valid whatever the nature of the distribution, provided only the first three moments exist and $PP' > 0$.

The Clinical Significance of κ

For an interval response satisfying the model underlying ρ,

$$\mathscr{E}(\xi_i|x_i) = (1-\rho)\mu + \rho x_i,$$

$$\operatorname{var}(\xi_i|x_i) = (1-\rho)\sigma_x^2.$$

The difference in characteristic response $(\xi_i - \xi_j)$ between two subjects can be predicted from the observed response difference $x_i - x_j$ to an extent reflected by ρ:

$$\mathscr{E}(\xi_i - \xi_j|x_i, x_j) = \rho(x_i - x_j),$$

$$\operatorname{var}(\xi_i - \xi_j|x_i, x_j) = 2(1-\rho)\sigma_x^2.$$

Similarly for a dichotomous response under the model proposed for κ,

$$\mathscr{E}(\Pi_i|D+) = \frac{\mathscr{E}(\Pi_i^2)}{\mathscr{E}(\Pi_i)} = \frac{\sigma^2 + P^2}{P} = P(1 - \kappa) + \kappa,$$

and

$$\mathscr{E}(\Pi_i|D-) = \frac{\mathscr{E}[\Pi_i(1 - \Pi_i)]}{\mathscr{E}(1 - \Pi_i)} = P(1 - \kappa).$$

The difference in expected value of the risk of positive diagnosis between a subject classified $D+$ and one classified $D-$ is κ. Also

$$\mathscr{E}(\Pi_i^2|D+) = \frac{\mathscr{E}(\Pi_i^3)}{P} = \frac{\beta = 3P\sigma^2 + P^3}{P},$$

$$\mathscr{E}(\Pi_i^2|D-) = \frac{\mathscr{E}[\Pi_i^2(1 - \Pi_i)]}{P'} = \frac{\sigma^2 + P^2P' - \beta - 3P\sigma^2}{P'},$$

and hence

$$\operatorname{var}(\Pi_i|D+) = PP'\kappa - P'^2\kappa^2 + \frac{\beta}{P},$$

$$\operatorname{var}(\Pi_i|D-) = PP'\kappa - P^2\kappa^2 - \frac{\beta}{P'}.$$

Unlike ρ, the precision of a single observation as an estimate of characteristic response (i.e., risk Π_i) depends on what that observation is. If $\kappa = 1$, both variances are equal to zero; if $\kappa = 0$, (since $\beta = 0$) once again both variances are zero.

Thus, in general, both ρ and κ are measures of the sensitivity of the system of observation to existing intersubject differences in the population. Either insensitivity of the system of observation or nonexistent differences result in a lack of reliability.

The Research Significance of κ

Unreliability of observation attenuates the precision of estimators, and reduces the power of statistical tests of hypotheses [Cochran, 1968]. The correlation coefficient, for example, between two unreliable observations, each satisfying the assumptions underlying ρ and having coefficients of reliability ρ_1 and ρ_2 (both greater than zero) is, let us say, τ_0. This correlation coefficient is related to that between the two true subject characteristics, τ, by the relationship

$$\tau_0 = \tau[(\rho_1\rho_2)^{1/2}].$$

This is the phenomenon known as attenuation of correlation, i.e., one always underestimates the magnitude of the true correlation [Walker & Lev, 1953]. If a test of the null hypothesis $\tau = 0$ is based on such unreliable data, the significance level of the test is correct since τ is zero if and only if τ_0 is zero. The power of this test depends approximately on the value of $(N)^{1/2}\tau[(\rho_1\rho_2)^{1/2}]$. Thus to achieve the power of test obtainable with reliable observations and N subjects, one would require approximately $N/\rho_1\rho_2$ subjects with unreliable observations [Kraemer, 1975]. Similar situations pertain when tests are based on unreliable categorizations.

TABLE 1

Calculation of Diagnostic Probabilities Given True Probabilities and Sensitivity (p_1, p_2) and Specificity (q_1, q_2) of the Two Diagnoses

Response Diagnosis 1	 Diagnosis 2	True State Probability	Observation Probability
Positive	Positive	P_{++}	$Q_{++} = P_{++}p_1p_2 + P_{+-}p_1q_2' + P_{-+}q_1'p_2 + P_{--}q_1'q_2'$
Positive	Negative	P_{+-}	$Q_{+-} = P_{++}p_1p_2' + P_{+-}p_1q_2 + P_{-+}q_1'p_2' + P_{--}q_1'q_2$
Negative	Positive	P_{-+}	$Q_{-+} = P_{++}p_1'p_2 + P_{+-}p_1'q_2' + P_{-+}q_1p_2 + P_{--}q_1q_2'$
Negative	Negative	P_{--}	$Q_{--} = P_{++}p_1'p_2' + P_{+-}p_1'q_2 + P_{-+}q_1p_2' + P_{--}q_1q_2$

The χ^2-test:

If, for example, the population could be described as a true double dichotomy (Case 1) with response probabilities as in Table 1, the power of the χ^2-test of independence with one degree of freedom depends on the non-centrality parameter of the χ^2-statistic approximately:

$$\lambda = \frac{N(P_{++}P_{--} - P_{+-}P_{-+})^2}{P_1P_1'P_2P_2'}$$

where P_1 is true prevalence of the first diagnosis and P_2 the true prevalence of the second. When the misclassification probabilities of the first diagnosis are (p_1', q_1') of the second diagnosis are (p_2', q_2') what one observes is determined by the true response pair of the subject and the misclassification probabilities (Table 1). The non-centrality parameter of this χ^2-test is approximately:

$$\lambda_0 = \frac{N(Q_{++}Q_{--} - Q_{+-}Q_{-+})^2}{Q_1Q_1'Q_2Q_2'}$$

where Q_1 and Q_2 are observed prevalences of positive diagnoses. It may be computed that:

$$\lambda_0 = \lambda\kappa_1\kappa_2,$$

where κ_1 and κ_2 are the reliability coefficients of the two diagnostic decisions. Once again, as in the case of the correlation coefficient between unreliable measures, to achieve the power obtainable with N subjects and a totally reliable classification system, one would need approximately $N/\kappa_1\kappa_2$ subjects with systems of observation having reliability coefficients κ_1 and κ_2.

t-Test:

If one used an unreliable diagnostic procedure to identify two subgroups of subjects (e.g., schizophrenic vs. non-schizophrenic patients) in order to compare the groups as to a certain quantitative response (e.g., neuroendocrine level) using a t-test procedure, the power of the test would be attenuated both by the unreliability of the diagnosis (κ) and that of the response measure (ρ).

Let us suppose that a subject's characteristic response (ξ_i) is linearly related to his risk of positive diagnosis (Π_i):

$$\xi_i = \alpha + \Delta\Pi_i + \eta_i,$$

where η_i measures the prediction error and

$$\mathscr{E}(\eta_i) = 0, \text{ var}(\nu_i) = \sigma_\xi^2.$$

If one had a totally reliable diagnostic procedure, schizophrenic patients would have $\Pi_i = 1$, others $\Pi_i = 0$, and the true response differential would be Δ, the within group variance of characteristic responses, σ_ξ^2. The power of the t-test would depend on degrees of freedom ($2(N-1)$) and on the noncentrality parameter of the t-distribution:

$$\lambda = \frac{N}{2}\delta \qquad \text{where} \qquad \delta = \frac{\Delta}{\sigma_\xi}\cdot$$

In the real situation, however, one cannot observe either ξ_i or Π_i, but observes $x_i = \xi_i + \epsilon_i$ in the positive (D+) and negative (D−) diagnosis groups. Then:

$$\mathscr{E}(x_i|D+) = \alpha + \Delta\mathscr{E}(p_i|D+),$$
$$\mathscr{E}(x_i|D-) = \alpha + \Delta\mathscr{E}(p_i|D-).$$

Thus the observed response differential is

$$\Delta_0 = \Delta\kappa,$$

i.e., the unreliability of the diagnosis attenuates response differential between groups. Also

$$\begin{aligned}\mathscr{E}(x_i^2|D+) &= \mathscr{E}((\alpha + \Delta p_i + \eta_i + \epsilon_i)^2|D+)\\ &= \alpha^2 + 2\alpha\Delta\mathscr{E}(p_i|D+) + \Delta^2 \text{ var}(p_i|D+) - \Delta^2\mathscr{E}^2(p_i|D+) + \sigma_\xi^2 + \sigma_\epsilon^2\end{aligned}$$

and thus

$$\begin{aligned}\text{var}(x_i|D+) &= \Delta^2 \text{ var}(p_i|D+) + \sigma_\xi^2 + \sigma_\epsilon^2,\\ &= \Delta^2 \text{ var}(p_i|D+) + \frac{\sigma_\xi^2}{\rho}\cdot\end{aligned}$$

Similarly

$$\begin{aligned}\text{var}(x_i|D-) &= \Delta^2 \text{ var}(p_i|D-) + \sigma_\xi^2 + \sigma_\epsilon^2,\\ &= \Delta^2 \text{ var}(p_i|D-) + \frac{\sigma_\xi^2}{\rho}\cdot\end{aligned}$$

The only effect of unreliability of the measure is to increase intragroup variance. The effect of unreliability of the diagnosis, however, is (i) to attenuate observed response differential, (ii) to increase intragroup variance even further, and (iii) to induce inequality of variance in the two groups. The latter effect might comprise the validity of the t-test procedure if group sizes were unequal [Scheffé, 1959]. The power would depend on the degrees of freedom [$2(N - 1)$ as before] and the non-centrality parameter

$$\lambda_0 = \left(\frac{N}{2}\right)^{1/2} \frac{\Delta\kappa}{\sigma_\xi\left[\left(\delta^2\,\overline{\text{var}} + \frac{1}{\rho}\right)^{1/2}\right]},$$

$$= \lambda\ \frac{\kappa[(\rho)^{1/2}]}{(\delta^2\rho\,\overline{\text{var}} + 1)^{1/2}},$$

where $\overline{\text{var}} = .5[\text{var}(p_i|D+) + \text{var}(p_i|D-)]$. Thus with reliable diagnosis and measure, $\lambda_0 = \lambda$ as before; with reliable diagnosis ($\kappa = 1$) and unreliable measure $\lambda_0 = (\rho)^{1/2}\lambda$; with unreliable diagnosis and reliable measure ($\rho = 1$)

$$\lambda_0 = \frac{\lambda\kappa}{(\delta^2\,\overline{\text{var}} + 1)^{1/2}}.$$

Once again, to achieve the power attainable with a reliable measure and a reliable diagnosis with N subjects, one would need a much larger sample with an unreliable measure and/or diagnosis, and the increase would be directly determined by the degrees of unreliability.

Factors Affecting Reliability

Reliability As It Reflects Errors Of Observation

Coefficient ρ is 1.0 if and only if $\sigma_\epsilon = 0$; similarly, coefficient κ is 1.0 if and only if each p_i is either 1 or 0. In both cases perfect reliability (ρ, $\kappa = 1$) occurs only if disagreement is impossible.

At the other extreme, a ρ or κ near zero may indicate deficiencies in the instrument of measurement or in the training of observers, i.e., error of measurement ($\sigma_\epsilon^2 \to \infty$). Thus the obvious first strategy to improve reliability is to standardize conditions of observation, to select the best available instruments of measurement (whether this be a scale, assay procedure, or pencil-and-paper test), to clarify and specify criteria, to select and train observers.

Reliability As It Reflects Population Characteristics

A low magnitude of ρ will also result if intersubject variability (σ_ξ^2, σ^2) is small. Furthermore, when the error incorporates inconsistency of subjects' responses over time, as in test-retest reliability (as opposed to split-half or simultaneous observer reliability), the achievable lower bound of σ_ϵ^2, i.e., that which exists when all error of observation is removed, may exceed zero. Hence a low value of ρ or κ may be close to optimal achievable reliability given the nature of the population in terms of intersubject variance and/or inconsistency of subjects' responses. For example, in Case 1, if p and $q = .95$, when $P^* = .05$, $\kappa = .448$. It is difficult to identify any medical diagnostic procedure with sensitivity specificity exceeding these levels [Galen & Gambino, 1975]. What would be regarded as "poor" reliability (.448) using suggested standards [Landis & Koch, 1977] would here reflect, not error, but low prevalence, and may be near optimal for such a population.

Thus poor reliability may reflect the nature of the population, not of the observation procedure. In such cases one would need to compensate for unreliability by increasing sample size in research applications or by using multiple observations per subject in research or clinical applications.

Reliability As It Reflects the Choice of Measure

Finally, if the system of measurement does not reflect the true nature of the population, this too will result in a poor reliability. For example, in Case 2, when $\rho = .70$, which is usually considered acceptable test-retest reliability for a quantitative response measure, the *maximal* attainable value of κ is .494 (when there is a median split of the population). In this case the apparent poor reliability results not because of error or low prevalence, but because a dichotomous response measure is applied to a population not in fact a dichotomy. The resulting low coefficient of reliability reflects a loss of information.

Similarly if a truly dichotomous population is described by more than two response categories or by a quantitative measure, the reliability of the resulting system of observation will be diminished.

Exploration of possible alternative methods of describing observations may therefore be necessary to identify the most appropriate system of observations.

Use of Multiple Independent Observations

If no single observation can be identified with satisfactory reliability, it may yet be possible to utilize multiple independent observations per subject to yield one highly reliable observation for clinical or research use.

For example, if one were to use the average of r independent observations for Subject i, $\bar{x}_i$, each satisfying the model underlying $\rho(\rho \neq 0)$, the coefficient of reliability of $\bar{x}_i$ is [Spearman, 1910; Brown, 1910]

$$\rho_r = \frac{r\rho}{1 + (r-1)\rho}.$$

Thus if $\rho = .400$, the average of 3 such observations has reliability $\rho_3 = .667$, and 5 has $\rho_5 = .769$. How reliable an observation one can produce under this strategy is limited only by the practical consideration of how many independent observations per subject are feasible.

Similarly, one could use r independent dichotomous observations for subject i, using as the subject's response measure the proportion of positive diagnoses, $\hat{\Pi}_i$. Then for subject i, $\mathscr{E}(\hat{\Pi}_i) = \Pi_i$ and $\text{var}(\hat{\Pi}_i) = \Pi_i(1 - \Pi_i)/r$. This represents an interval response measure, but not one satisfying assumptions for ρ. However, let $x_i = 2\arcsin\,(\hat{\Pi}_i)^{1/2}$. It is known that [Walker & Lev, 1953] x_i (for large r) is approximately normally distributed with mean $\xi_i = 2\arcsin\,(\Pi_i)^{1/2}$ and variance $1/r$. This interval measure satisfies underlying assumptions of ρ and has reliability coefficient

$$K_r = \frac{r\sigma_\xi^2}{r\sigma_\xi^2 + 1}.$$

However

$$\sigma_\xi^2 = \text{var}\left(\frac{2\arcsin}{p_i}\right) \approx \frac{\sigma^2}{PP'} = \kappa,$$

i.e.,

$$K_r \approx \frac{r\kappa}{1 + r\kappa},$$

or for large r

$$K_r \approx \frac{r\kappa}{1+(r-1)\kappa}.$$

Thus the Spearman-Brown formula holds approximately for categorical responses as well. Again, how reliable a measure one can base on multiple use of a single measure of low but nonzero reliability, is limited only by practical feasibility.

In the case of multiple (k) categories, $\hat{\Pi}_i$ is a k-dimensional vector with elements equal to the proportion of the r observations classifying subject i into each of the k response categories. The covariance matrix of the vector $\hat{\Pi}_i$ is inversely proportional to r, the number of independent assessments per subject. While the Spearman-Brown formula has not, to our knowledge, been extended to multivariate responses of any kind, what is clear is that use of multiple observations per subject, whatever the nature of the response, will reduce unreliability by reducing the error in characterizing each subject. Therefore the principle of the Spearman-Brown approach extends even if the precise mathematical formula may not.

Summary

The coefficients of reliability ρ and κ are measures of the sensitivity of the system of observations to existing differences in the population. As such, both are useful as indicators of the clinical and research value of an observation. In the research context particularly, these measures of reliability can be used to indicate explicitly and quantitatively the degree of loss of precision or power of statistical procedures. However, in evaluating the magnitude of ρ or κ, one notes that the maximum magnitude of ρ or κ achievable by efforts to improve instrumentation or training of observers is determined by the nature of the subject population in terms of its homogeneity and intrinsic instability, and by the level of measurement selected.

However, other strategies to overcome the effects of low reliability are available. In research applications one may increase sample sizes to compensate for the effects of unreliability. In both clinical and research contexts, one can use the strategy of combining multiple observations per subject of low but nonzero reliability to yield one highly reliable observation.

Such evaluations and strategies depend on a conceptualization of the role κ plays in the population, since the concern is with applying the information which may be gained in a reliability study on one sample to consideration of use of the system of observations for future evaluations of patients or future uses in research.

REFERENCES

Brown, W. Some experimental results in the correlation of mental abilities. *British J. of Psychology*, 1910, *3*, 296–322.

Cochran, W. G. Errors of measurement in statistics. *Techmometrics*, 1968, *10*, 637–666.

Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 1960, *20*, 37–46.

Cronbach, L. J., Gleser, G. C., Nanda, H. & Rajaratnam, N. *The dependability of behavioral measurements*. New York: John Wiley & Sons, Inc., 1972.

Everitt, B. S. Moments of the statistics kappa and weighted kappa. *British Journal of Mathematical and Statistical Psychology*, 1968, *21*, 97–103.

Fleiss, J. L. *Statistical methods for rates and proportions*. New York: John Wiley & Sons, 1973.

Fleiss, J. L. Measuring agreement between two judges on the presence or absence of a trait. *Biometrics*, 1975, *31*, 651–659.

Fleiss, J. L., Cohen, J. & Everitt, B. S. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 1969, *72*, 323–327.

Galen, R. S. & Gambino, S. R. *Beyond normality: The predictive value and efficiency of medical diagnosis.* New York: John Wiley & Sons, 1975.

Helzer, J. E., Robins, L. N., Tarbleson, M., Woodruff, R. A., Reich, T. & Wish, E. D. Reliability of psychiatric diagnosis: I. A methodological review. *Archives of General Psychiatry,* 1977, *34,* 129–133.

Helzer, J. E., Clayton, P. J., Pambakian, R., Reich, T., Woodruff, R. A. & Reveley, M. A. Reliability of diagnostic classification. *Archives of General Psychiatry,* 1977, *34,* 136–141.

Hubert, L. Kappa revisited. *Psychological Bulletin,* 1977, *84,* 289–297.

Kirk, D. B. On the numerical approximation of the bivariate normal (tetrachoric) correlation coefficient. *Psychometrika,* 1973, *38,* 259–267.

Koran, L. M. The reliability of clinical methods, data and judgments. *N.E. Journal of Medicine,* 1975, *293,* 695–701.

Kraemer, H. C. On estimation and hypothesis testing problems for correlation coefficients. *Psychometrika,* 1975, *40,* 473–485.

Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics,* 1977, *33,* 159–174.

Light, R. J. Measures of agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin,* 1971, *76,* 365–377.

Scheffé, H. *The analysis of variance.* New York: John Wiley & Sons, Inc., 1959.

Spearman, C. Correlation calculated from faulty data. *British Journal of Psychology,* 1910, *3,* 271–295.

Sptizer, R. L. & Fleiss, J. L. A re-analysis of the reliability of psychiatric diagnosis. *British Journal of Psychiatry,* 1974, *125,* 341–347.

Walker, M. H. & Lev, J. *Statistical inference.* New York: Henry Holt & Company, 1953.

Manuscript received 1/3/79
Final version received 7/2/79