Manuscript as submitted to and published in
Communication Methods and Measures 1,1: 77-89, 2007.

Answering the Call for a Standard Reliability Measure for Coding Data

Andrew F. Hayes
School of Communication
The Ohio State University
hayes.338@osu.edu

Klaus Krippendorff
Annenberg School for Communication
University of Pennsylvania
kkrippendorff@asc.upenn.edu

Abstract

In content analysis and similar methods, data are typically generated by trained human observers who record or transcribe textual, pictorial, or audible matter in terms suitable for analysis. Conclusions from such data can be trusted only after demonstrating their reliability. Unfortunately, the content analysis literature is full of proposals for so-called reliability coefficients, leaving investigators easily confused not knowing which to choose. After describing the criteria for a good measure of reliability, we propose Krippendorff's alpha as the standard reliability measure. It is general in that it can be used regardless of the number of observers, levels of measurement, sample sizes, and presence or absence of missing data. To facilitate the adoption of this recommendation, we describe a freely available macro written for SPSS and SAS to calculate Krippendorff's alpha and illustrate its use with a simple example.

Answering the Call for a Standard Reliability Measure for Coding Data

Introduction

Much communication research is based on data generated by human beings asked to make some kind of judgment. In content analysis, for example, people (for generality, henceforth referred to as "observers") are employed in the systematic interpretation of textual, visual, or audible matter, such as newspaper editorials, television news, advertisements, public speeches, and other verbal or nonverbal units of analysis. Generating such data may take the form of judgments of *kind* (in which category does this unit belong?), *magnitude* (how prominent is an attribute within a unit?), or *frequency* (how often does something occur). When relying on human observers, researchers must worry about the quality of the data—specifically, their reliability. Are the data being made and subsequently used in analyses and decision making the result of irreproducible human idiosyncracies or do they reflect properties of the phenomena (units of analysis) of interest on which others could agree as well? The answer to this question can affirm or deny the usability of one's data, and respectable scholarly journals typically require quantitative evidence for the reliability of the data underlying published research.

Among the kinds of reliability—stability, reproducibility, and accuracy—reproducibility arguably is the strongest and most feasible kind to test (Krippendorff, 2004a). It amounts to evaluating whether a coding instrument, serving as common instructions to different observers of the same set of phenomena, yields the same data within a tolerable margin of error. The key to reliability is the agreement observed among independent observers. The more observers agree on the data they generate, and the larger the sample of units they describe, the more comfortable we can be that their data are exchangeable with data provided by other sets of observers (c.f. Hayes, 2005), reproducible, and trustworthy.

Choosing an index of reliability is complicated by the number of indices that have been proposed. For example, Popping (1988) compares an astounding 43 measures for nominal data, mostly applicable to reliability data generated by only two observers. Furthermore, these indices respond to rather different properties in the data, some related to reliability, others not. Understanding these properties is not a trivial matter but essential to correctly interpreting the meaning of such indices (Krippendorff, 2004b). This complexity combined with the lack of consensus among communication researchers on which measures are appropriate lead Lombard, Snyder-Duch, and Bracken (2002, 2004) to call for a reliability standard that can span the variable nature of available data. Unfortunately, they do not respond to their own call in a way that will help researchers choose a measure, and they leave the reader no better off with respect to knowing which measure of reliability should be the measure of choice. In the rest of this paper, we answer that call by suggesting that of the existing measures, Krippendorff's alpha (Krippendorff, 1970, 2004a) is best suited as a standard. It generalizes across scales of measurement, can be used with any number of observers with or without missing data, and it satisfies all of the important criteria for a good measure of reliability. Unfortunately, Krippendorff's alpha cannot be calculated by popular statistical packages used in the social sciences. To remedy this situation, we describe a macro written for SPSS and SAS that computes Krippendorff's alpha, thereby allowing for the widespread implementation of our recommendation.

Criteria for a Good Measure of Reliability

Before any statistic can measure reliability, it must be applied to reliability data properly obtained. Specifically, the units of analysis whose properties are to be recorded or described must be independent of each other and the data generating process—informed by instructions that are common to all observers who identify, categorize, or describe the units—must be repeated by different observers working independently of each other. Furthermore, the set of units used in the reliability data should be a random sample (or at least approximating one) from the universe of data whose reliability is in question, and the observers employed should be common enough to be found elsewhere.

Given such reliability data, a good index of reliability should have the following properties:

(1) It should assess the *agreement* between two or more observers who describe each of the units of analysis separately from each other. For more than two observers, this measure should be (a) independent of the number of observers employed and (b) invariant to the permutation and selective participation of observers. Under these two conditions, agreement would not be biased by the individual identities and number of observers who happen to generate the data.

(2) It should be grounded in the distribution of the categories or scale points actually used by the observers. Specifically, the index should not be confounded by the number of categories or scale points made available for coding. This assures that reliability is not biased by the difference between what the authors of the coding instructions imagined the data may be like, and what the data turned out to be.

(3) It should constitute a numerical scale between at least two points with sensible reliability interpretations. By convention, perfect agreement is set to 1.000 or 100%. The absence of agreement, typically indicated by 0.000 (and not necessarily constituting the endpoint of the reliability scale), should represent a situation in which the units of analysis bear no statistical relationship to how they end up being identified, coded, or described. These two points enable an index to be interpreted as the degree to which the data can be relied upon in subsequent analyses.

(4) It should be appropriate to the level of measurement of the data. This demands that the information contained between the categories or in the metric underlying the reliability data is fully utilized, neither spuriously added nor ignored. When applying a statistic to several kinds of data, it must maintain its mathematical structure, except for responding to the properties of the metric involved, and only these. This enables comparisons across different metrics, as required for a reliability standard.

(5) Its sampling behavior should be known or at least computable.

With these criteria in mind, we discuss the adequacy of several measures of reliability that enjoy some use by researchers in communication and related fields. Given space constraints, we do not provide mathematical details of the measures, instead referring the interested reader to the original sources for this information.

Percent agreement is simply the proportion of units with matching descriptions on which two observers agree. This measure is easily calculated but flawed in nearly all important respects. It

satisfies (1), but only because of its limitation to two observers. As agreement is the more difficult to achieve the more categories are involved, it fails (2). While 100% is an unambiguous indicator of reliability, 0% is not. Zero percent can arise only when observers disagree on every unit being judged. Such a phenomenon would be unlikely unless observers are working together, which violates the condition that reliability data be generated by observers working independent of each other. Thus, without a meaningful reliability scale, all deviations from 100% agreement become unintelligible, failing (3). It satisfies (4), but only for nominal data. Incidentally, the above applies also to Osgood's (1959) index of reliability, christened C.R. by Holsti (1969), which is essentially a percent agreement measure.

Bennett et al's S (Bennett, Alpert, & Goldstein, 1954). This statistic has been reinvented with minor variations at least five times as Guilford's *G* (Holley & Guilford, 1964), the R.E. (random error) coefficient (Maxwell, 1970), *C* (Janson & Vegelius, 1979), κ_n (Brennan & Prediger, 1981), and the intercoder reliability coefficient $I_{r.}$ (Perreault & Leigh, 1989). *S* responds to the failure of percent agreement to satisfy (2), correcting it for the number of categories available for coding. However, *S* is inflated by the number of unused categories that the author of the instrument had imagined and by rarely used categories in the data, thus failing (2) and (3) as well. Because *S* corrects percent agreement, it is limited to two observers and nominal data.

Scott's pi (π) (Scott, 1955) was the first coefficient to fully satisfy (2) and (3). It corrects percent agreement, much like *S* does, but by the agreement that is expected when the units are statistically unrelated to their descriptions, "by chance," thus constituting a scale with valid reliability interpretations. But this correction does not overcome the limitations of percent agreement to two coders and nominal data.

Cohen's kappa (κ). Cohen (1960) intended to improve on Scott's π but created a hybrid index instead. Kappa corrects percent agreement, just as do *S* and π, but by what can be expected when the two observers are statistically independent of each other. Kappa violates (1) by not allowing observers to be freely permutable or interchangeable and it violates (3) by defining its zero point as would be appropriate in correlation or association statistics. Kappa, by accepting the two observers' proclivity to use available categories idiosyncratically as baseline, fails to keep κ tied to the data whose reliability is in question. This has the effect of punishing observers for agreeing on the frequency distribution of categories used to describe the given phenomena (Brennan & Prediger, 1981, Zwick, 1988) and allowing systematic disagreements, which are evidence of unreliability, to inflate the value of κ (Krippendorff, 2004a, 2004b). Kappa retains the limitations of percent agreement to two observers and nominal data. Cohen (1960) and Fleiss et al. (2003) discuss approximations to (5). However, its applicability in other empirical domains notwithstanding, κ is simply incommensurate with situations in which the reliability of data is the issue.

Fleiss's K. Fleiss (1971) generalized Scott's π to many observers, calling it kappa. This confusion led Siegel and Castellan (1988) to rename it *K*. Just as Scott's π, *K* satisfies (1) but for two or more observers. By fulfilling (3), *K* avoids κ's aforementioned numerical biases but remains limited to nominal data.

Cronbach's (1951) *alpha* (α_C) is a statistic for interval-level data that responds to the consistency of observers when numerical judgments are applied to a set of units. It is called a reliability

coefficient but does not measure agreement as required by (1). Instead, it quantifies the consistency by which observers judge units on an interval scale without being sensitive to how much the observers actually agree in their judgments. Without defining scale points with valid reliability interpretations—at least not regarding reproducibility as described in the foregoing—α_C fails to satisfy (3) and is therefore unsuitable to assess reliability of judgments. It is appropriate as a measure of the reliability of an aggregate measure across observers, such as the arithmetic mean judgment, but it does not directly index the extent to which observers actually agree in their judgments.

Krippendorff's (1970, 2004a) *alpha* (α) satisfies all of the above conditions and we propose it as the standard reliability statistic for content analysis and similar data making efforts. Regarding (1), α counts pairs of categories or scale points that observers have assigned to individual units, treating observers as freely permutable and being unaffected by their numbers. This dispels the common belief that reliability is the more difficult the more observers are involved. Regarding (2), α is exclusively rooted in the data generated by all observers. Regarding (3), α defines the two reliability scale points as 1.000 for perfect reliability and 0.000 for the absence of reliability, i.e., as if categories or scale points were statistically unrelated to the units they describe—not to be confused with the statistical independence of observers (as in κ). Regarding (4), α measures agreements for nominal, ordinal, interval, and ratio data, rendering the reliabilities for such data fully comparable across different metrics. Regarding (5), below we avoid assuming approximations and, instead, bootstrap the distribution of α from the given reliability data.

Krippendorff's α defines a large family of reliability coefficients and embraces several known ones. It accomplishes this by calculating disagreements instead of correcting percent-agreements, avoiding its above-mentioned limitations. In its two-observer nominal data version, α is asymptotically equal to Scott's π. In its two-observer ordinal data version, α is identical to Spearman's rank correlation coefficient ρ (rho) (without ties in ranks). In its two-observer interval data version, α equals Pearson et al.'s (1901**)** intraclass-correlation coefficient. Its extension to many observers is stated in analysis of variance terms (Krippendorff, 1970). Thus, α is in good company. Alpha also allows for reliability data with missing categories or scale points, a frequent reality that none of the reviewed measures has been able to cope with.

SPSS and SAS Implementation: The KALPHA Macro

There is little point to proposing a standard in the absence of computational support from existing software. Unfortunately, with a few exceptions, Krippendorff's α is not available in the majority of statistical software packages widely used by researchers, such as SPSS and SAS.[1] To facilitate the adoption of our recommendation, we have produced a macro (KALPHA) for these computing platforms that computes Krippendorff's α. A macro is a set of commands that, when executed, produces a new shortcut command that accepts arguments the user specifies to make the command set produce the desired output. Space constraints preclude the publication of the macro in this article. Interested readers are referred to http://www.comm.ohio-state.edu/ahayes/macros.htm where a copy of the macro can be downloaded and instructions for

[1] Kang, Kara, Laskey, & Seaton (1993) provide a SAS macro that calculates several reliability measures, including Krippendorff's alpha for nominal, ordinal, and interval data. However, their macro does not allow for missing data and it does not generate statistics useful for statistical inference, as our macro does.

its use are provided. In the rest of this article, we describe the functionality of the macro and work through a single example.

The data for this example come from five observers who were asked to evaluate the local news coverage given to the challenger running against an incumbent for a political office in one of several political races. The observers were given 40 newspaper articles describing the race published by the largest circulation newspaper in the incumbent's district within one week prior to the election. These 40 articles were randomly selected from the pool of all articles published during that one week period. Observers rated whether the tone of the article suggested the challenger was a sure loser (0), somewhat competitive (1), competitive (2), or a likely winner (3). After training, observers read and judged the articles independently. The data were entered into SPSS such that each article was represented with a row in the data file, and each observer's evaluation of the articles was located in the columns, with the columns labeled "obs1", "obs2" "obs3", "obs4", and "obs5". Thus, in this example, the data occupied a 40 (articles) × 5 (observers) matrix, with each cell in the matrix containing a 0, 1, 2, 3 or a period character (".") for missing judgments (see Table 1). To compute α, the following command was executed in SPSS after activating the KALPHA macro:

KALPHA judges = obs1 obs2 obs3 obs4 obs5/level = 2/detail = 1/boot = 10000.

The "/level" subcommand provides information to the macro about the level of measurement of the judgments (1 = nominal, 2 = ordinal, 3 = interval, 4 = ratio). The "/detail" subcommand tells the macro to print (1) or suppress the printing (0) of the computational details such as the observed and expected coincidence matrices and the delta matrix (see Krippendorff, 2004a).

The output from the macro is displayed in Figure 1. As can be seen, Krippendorff's ordinal α is 0.7598, a modest degree of reliability. In the observed coincidence matrix, the disagreements between observers cluster around the diagonal containing perfect matches. A coincidence matrix should not be confused with the more familiar contingency matrix, which is used to represent data for two observers and define other estimators of reliability. The expected coincidence matrix can be interpreted as what would be expected under conditions of "chance", as if observers responded independent of the properties of the units they were asked to describe. The delta matrix visualizes how α weights the coincidences according to the level of measurement of the data. For technical details on the construction of the observed and expected coincidence and delta matrices, see Krippendorff (2004a)[2].

The macro output also provides statistics pertinent to statistical inference. The obtained value of α is subject to random sampling variability—specifically, variability attributable to the selection of units in the reliability data and the variability of their judgments. It is generally of interest to infer the true value of alpha, α_{true}— if the measure were applied to the whole universe of units rather than to the sub sample used to estimate data reliability— and the likelihood of α_{true} being below the minimally acceptable α-value.[3] To answer these two questions, one needs to know the sampling distribution of α. Unfortunately, the sampling distribution of α is not

[2] Computational details can also be obtained by downloading the document at http://www.asc.upenn.edu/usr/krippendorff/webreliability2.pdf or http://www.comm.ohio-state.edu/ahayes/macros.htm

[3] Investigators often ask whether observed agreement is sufficiently above "chance." However, this null-hypothesis is irrelevant where data reliability is considered. For data to be trustworthy, their reliability must not significantly deviate from perfect reliability, which calls for a rather different test. The question is whether observed unreliabilities are still tolerable for data to be relied upon in subsequent analyses.

known and existing approximations are rough at best. However, this sampling distribution can be empirically generated by bootstrapping (see e.g., Mooney & Duval, 2003, or Efron & Tibshirani, 1998, for a discussion of the rationale for bootstrapping in statistical inference). This is accomplished by acknowledging that the units of bootstrapping for reliability are the *pairs of judgments* associated with particular units. With 5 observers judging 40 units, there are 200 judgments possible and 400 pairs of judgments (10 pairs per unit). But in the 40 × 5 matrix of judgments in Table 1 there only 159 judgments because not all observers judged all 40 units, resulting in only 239 pairs of judgments. The bootstrap sampling distribution of alpha is generated by taking a random sample of 239 pairs of judgments from the available pairs, weighted by how many observers judged a given unit. Alpha is computed in this "resample" of 239 pairs, and this process is repeated very many times, producing the bootstrap sampling distribution of α.[4] We use this sampling distribution to produce 95% confidence intervals for α_{true}, whose lower and upper limits are the values of α that define the 2.5th and 97.5th percentiles of the bootstrap distribution, respectively. Using the bootstrap sampling distribution, it is also possible to estimate the probability, q, that the reliability, α_{true}, is less than the required minimum value, α_{min} The larger the number of bootstrap samples taken, the more accurate the inferential statistics will be, but computational time will be longer. Fortunately, little additional precision is gained by setting the number of bootstrap samples larger than 10,000 or so.

The macro implements bootstrapping, with the number of bootstrap samples determined by the optional "/boot" subcommand (if this option is left out of the command, bootstrapping is disabled and the statistics described below are not printed). The macro saves the bootstrap estimates of α into a file, and this file can be used to generate a visual representation of the distribution of α (see Figure 2). In this example, 10,000 bootstrap samples were taken to produce confidence intervals at the 95% level and estimates of q for several α_{min} (here α_{min}= 0.9, 0.8, 0.7, 0.67, 0.6, 0.5). As can be seen in Figure 1, the confidence interval for α_{true} is 0.7078 to 0.8078, meaning that if the population of units were coded, reliability would likely be somewhere between these two values. The probability, q, of failing to achieve α_{min} is displayed for five values of α_{min}. As can be seen, q = 0.0125 for α_{min} = 0.70, and q = 0.9473 for α_{min} = 0.80. In other words, if this research problem demands that reliability must not be below α_{min} = 0.800, our reliability data would suggest serious trouble. If the reliability standard were relaxed to α_{min} = 0.700, the risk of accepting the data as reliable when they are not is quite low, q = 0.0125.

It is worth pointing out that the obtained alpha is always higher than its nominal version when the observers' disagreements conform to the metric of the chosen α. In our example, disagreements, which can be examined in the coincidence matrix, are not randomly distributed in the off-diagonal as would be expected if data were nominal. In fact, they cluster around the diagonal and, moreover, most ranks on which coders disagree are just one rank apart. If the data were treated as nominal data, α = 0.4765, whereas the ordinal α is 0.7598. By the same token, the interval α is 0.7574 and the ratio α is 0.6621, attesting to the fact that the majority of observed disagreements confirm the observers' ordinal conceptions.

[4] A document describing the bootstrapping algorithm can be found at http://www.comm.ohio-state.edu/ahayes/macros.htm

Conclusion

In this article, we outlined the reasons why Krippendorff's α should be the standard measure of reliability in content analysis and similar data making efforts. Although this statistic has been known for over 30 years and used in simple cases, its lack of implementation in popular statistical packages has undoubtedly prevented the full utilization of its capabilities. We hope that the macro described here will encourage researchers to take advantage Krippendorff's α rather than one of the dozens of other measures, none of which is as well suited to the task.

References

Bennett, E. M., Alpert, R., & Goldstein, A. C. (1954). Communications through limited response questioning. *Public Opinion Quarterly, 18*, 303-308.

Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement, 41*, 687-699.

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

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrica, 16*, 297-334.

Efron, B., & Tibshirani, R. J. (1998). *An introduction to the bootstrap*. Boca Raton, FL: CRC Press.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*, 378-382.

Fleiss, J. L., Levin, B., & Paik, M.C. (2003). *Statistical methods for rates and proportions (*3rd Ed.*)*. New York: John Wiley & Sons.

Hayes, A. F. (2005). *Statistical methods for communication science*. Mahwah, NJ: Lawrence Erlbaum Associates.

Holley, W., & Guilford, J. P. (1964). A note on the G-index of agreement. *Educational and Psychological Measurement, 24*, 749-754.

Holsti, O. R. (1969). *Content analysis for the social sciences and humanities*. Reading, MA: Addison-Wesley.

Janson, S., & Vegelius, J. (1979). On generalizations of the G index and the phi coefficient to nominal scales. *Multivariate Behavioral Research, 14*, 255-269.

Kang, N., Kara, A., Laskey, H. A., & Seaton, F. B. (1993). A SAS Macro for calculating intercoder agreement in content analysis. *Journal of Advertising, 12*, 17-28.

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement, 30*, 61–70.

Krippendorff, K. (2004a). *Content analysis: An introduction to its* methodology (2nd Ed.). Thousand Oaks, CA: Sage Publications.

Krippendorff, K. (2004b). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, *30*, 411-433.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research, 28*, 587-604.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2004). A call for standardization in content analysis reliability. *Human Communication Research, 30*, 434-437.

Maxwell, A. E. (1970). Comparing the classification of subjects by two independent judges. *British Journal of Psychiatry, 116*, 651-655.

Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping: A nonparametric approach to statistical inference*. Thousand Oaks, CA: Sage.

Osgood, C. E. (1959). The representational model and relevant research methods. In I. de Sola Pool (Ed.), *Trends in content analysis* (pp. 33-88). Urbana: University of Illinois Press.

Pearson, K., et al. (1901). Mathematical contributions to the theory of evolution: IX. On the principle of homotyposis and its relation to heredity, to variability of the individual, and to that of race. Part I: Homotyposis in the vegetable kingdom. *Philosophical Transactions of the Royal Society, 197* (Series A): 285–379.

Perreault, W. D., & Leigh, L. E. (1989). Reliability of nominal data based on qualitative judgments. *Journal of Marketing Research, 26*, 135-148.

Popping, R. (1988). On agreement indices for nominal data. In William E. Saris & Irmtrat N. Gallhofer (Eds.), *Sociometric research: Data collection and scaling*. Cambridge, MA: MIT Press.

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

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences* (2nd Ed.). Boston, MA: McGraw-Hill.

Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin, 103*, 347-387.

Author Notes

Both authors contributed equally to this paper. Correspondence should be addressed to Andrew F. Hayes, School of Communication, The Ohio State University, 154 N. Oval Mall, 3016 Derby Hall, Columbus, OH, 43210, hayes.338@osu.edu, or Klaus Krippendorff, Annenberg School for Communication, University of Pennsylvania, 3620 Walnut Street, Philadelphia, PA, 19104-6220, kkrippendorff@asc.upenn.edu. Please direct all questions regarding the macro to Hayes.

Table 1. Example Reliability Data Set

Unit	obs1	obs2	obs3	obs4	obs5
1	1	1	2	.	2
2	1	1	0	1	.
3	2	3	3	3	.
4	.	0	0	.	0
5	0	0	0	.	0
6	0	0	0	.	0
7	1	0	2	.	1
8	1	.	2	0	.
9	2	2	2	.	2
10	2	1	1	1	.
11	.	1	0	0	.
12	0	0	0	0	.
13	1	2	2	2	.
14	3	3	2	2	3
15	1	1	1	.	1
16	1	1	1	.	1
17	2	1	2	.	2
18	1	2	3	3	.
19	1	1	0	1	.
20	0	0	0	.	0
21	0	0	1	1	.
22	0	0	.	0	0
23	2	3	3	3	.
24	0	0	0	0	.
25	1	2	.	2	2
26	0	1	1	1	.
27	0	0	0	1	0
28	1	2	1	2	.
29	1	1	2	2	.
30	1	1	2	.	2
31	1	1	0	.	0
32	2	1	2	1	.
33	2	2	.	2	2
34	3	2	2	2	.
35	2	2	2	.	2
36	2	2	3	.	2
37	2	2	2	.	2
38	2	2	.	1	2
39	2	2	2	2	.
40	1	1	1	.	1

Figure 1. Output from the SPSS KALPHA macro

```
kalpha judges = obs1 obs2 obs3 obs4 obs5/level = 2/detail = 1/boot = 10000.

Run MATRIX procedure:

Krippendorff's Alpha Reliability Estimate


             Alpha    LL95%CI    UL95%CI      Units   Observrs      Pairs
Ordinal      .7598      .7078      .8078    40.0000     5.0000   239.0000

Probability (q) of failure to achieve an alpha of at least alphamin:
   alphamin          q
      .9000     1.0000
      .8000      .9473
      .7000      .0125
      .6700      .0004
      .6000      .0000
      .5000      .0000

Number of bootstrap samples:
  10000

Judges used in these computations:
 obs1     obs2     obs3     obs4     obs5

=====================================================

Observed Coincidence Matrix
     32.33      8.83       .83       .00
      8.83     25.33     13.17       .67
       .83     13.17     35.83      6.17
       .00       .67      6.17      6.17

Expected Coincidence Matrix
     10.90     12.76     14.89      3.46
     12.76     14.28     17.01      3.95
     14.89     17.01     19.49      4.61
      3.46      3.95      4.61       .99

Delta Matrix
       .00   2025.00   9409.00  17292.25
   2025.00       .00   2704.00   7482.25
   9409.00   2704.00       .00   1190.25
  17292.25   7482.25   1190.25       .00

Rows and columns correspond to following unit values
       .00      1.00      2.00      3.00

------ END MATRIX -----
```

Figure 2. A graphical representation the sampling distribution of alpha from 10,000 bootstrap samples from the data in Table 1.