Rehabilitating the Agreement Rate as the Best Measure of Interjudge Replicability

Much data used in consumer research is created by subjective classification of text or other qualitative material. The extent to which two or more judges, acting independently, agree in classifying the qualitative material is regarded as a measure of the quality of the classification scheme and the resulting data. This presentation discusses how the agreement rate between judges is the best measure of the extent to which the data show interjudge replicability.

John E. Kushman, University of Delaware[1]

Interjudge Replicability and the Quality of Data

The quality of data is one of the most important issues in any research endeavor. Judges often create data for consumer research by classifying text or other qualitative material into categories. For instance, advertisements may be classified as informational or emotive. Consumer comments may be classified from highly unfavorable to highly favorable. The extent to which different judges, acting independently, agree on the classifications is widely regarded as an indication of the "quality" of the classification process and the data. Exactly what types of validity or reliability constitute quality of the process and the data is not important in this discussion. The issue addressed here is finding a measure of how much interjudge replicability the classifications show and testing whether the replicability is enough to regard the data as a good basis for analysis. Only nominal data are considered here, and the presentation is confined to two judges. The same analysis and technique can be extended to ordinal data and to multiple judges.

A Good Measure of Replicability

The following describes a good measure of replicability: (1) A measure that is easier to compute is better; (2) A measure for which tests of statistical significance are easier to compute is better; (3) A measure with valid statistical tests for "enough" evidence of replicability is better than one for which no statistical test is available or one where the available tests are based on assumptions of doubtful validity; (4) A measure which requires less complicated theoretical or conceptual justification is better (Ockham's Razor); (5) A measure which uses more familiar statistical concepts and calculations is better; (6) A measure that accounts for the agreement between judges that would be expected by chance is better than one that does not.

The properties above characterize a measure of replicability that is intuitively meaningful, readily accessible to many users and readily understandable to many readers. A good measure of replicability also should rely on an interpretation of "chance" that is intuitively related to the issue of quality of the classification process and rigorously related to tests for enough improvement over chance. The following definition of chance provides the required properties:

For each judge there is a uniform probability distribution of classification across the k categories. This is the most intuitive and simplest definition of chance in the absence of any exogenous information indicating a known distribution for a judge. If the classification process had no quality, all judges would presumably be reduced to assigning units to categories randomly with equal likelihood of any category. With k categories, the proportion of observations on which agreement is expected by chance is 1/k, or

$$p_c = 1/k. \quad (1)$$

The Agreement Rate: Problems and Rehabilitation

If there are 100 units of observation, and each of two judges classifies each unit into one of four categories, there are 200 classifications. Each unit has a category from judge one and a category from judge two. The observed agreement rate, p_o, is the proportion of units where the judges choose the same category.

If the judges arrive at the same category for 75 of the 100 units, p_o =.75. The agreement rate is easy to calculate and its interpretation is intuitive.

The agreement rate alone, however, meets none of the other criteria for a good measure of replicability. Alone, p_o implies no null hypothesis to define "enough" evidence of replicability. Alone, p_o does not account for the agreement among judges that would be expected by chance. For instance, if there are two categories, 50 percent agreement would be expected if each judge were flipping a coin. In this case, 75 percent agreement is not impressive. If there were four categories, expected agreement by chance would be 25 percent. Against this expectation, 75 percent agreement is more impressive evidence for replicability. If ρ is the true population agreement rate between judges, the binomial distribution can be used to test the hypothesis $\rho \leq 1/k$. This is a test of the null hypothesis that the population agreement rate is less than or equal to the agreement expected by chance. The consensus among researchers, however, is that this is a trivial null hypothesis.

In response to the deficiencies of the agreement rate, many measures of interjudge replicability have been developed. All of them, however, lack the intuitive meaning of the agreement rate. They all have problems of complexity, missing tests for enough replicability or tests with doubtful validity.

To rehabilitate the agreement rate, the first step is to define enough agreement over chance. If the agreement expected by chance is 1/k, the potential improvement over chance is (1-(1/k)). Let β be any subjectively chosen fraction such that $0 \leq \beta \leq 1$, and the required agreement is

$$p_c+ = p_c + \beta(1-(1/k)). \tag{2}$$

If β = .5, enough agreement is defined as that expected by chance plus 50% of the potential improvement over chance. The null hypothesis that accounts for an acceptable level of agreement over that expected by chance is

$$\rho \leq p_c+. \tag{3}$$

Judge one classifies each of N units into one of the k categories. Acting independently, judge two classifies each unit. According to the null hypothesis, the probability of agreement remains constant at p_c+ over all units. Thus, agreement follows a Bernoulli process with probability of success on any unit p_c+. This null hypothesis can be tested using the binomial distribution. Using the binomial distribution for N trials with probability of success p_c+, the researcher finds the probability of the observed agreement rate, p_o or higher. This probability is compared to a significance level selected *ex ante*, denoted α. If the probability is less than or equal α, the researcher rejects the null hypothesis and concludes there is sufficient agreement to justify confidence in the data.

For instance, for N=100, p_o = .75, k=2, β = .5, and α = .05, p_c+ = .75, and the binomial probability is 0.553. Seventy-five percent observed agreement is not enough to conclude the population agreement rate captures at least half the potential improvement over chance. However, if the classification scheme has four categories (k = 4), p_c+ = .625, and the binomial probability of 75% or better agreement is 0.006. In the case of four categories, observing 75% agreement is sufficient to reject the null hypothesis. The evidence supports a population agreement rate at least that expected by chance plus half the potential improvement over chance.

A researcher can force a finding of sufficient agreement over that expected by chance. Adding artificial categories to the classification scheme would increase k and reduce the observed agreement that would reject the null hypothesis for given N, α, and β. Such manipulation would be apparent in the definition of the artificial categories and the low frequencies with which those categories would be chosen.

Conclusion

Data created by subjective classifications of text and other subjective items into categories are often used in consumer research. An important question is whether the classification methods and the resulting data reflect sufficient agreement among judges to be used with confidence. The agreement rate between judges is the most obvious and intuitive measure of interjudge replicability, but the agreement rate alone has many deficiencies. The procedure presented here achieves all the desirable characteristics of a measure of interjudge replicability. It provides a test for enough replicability, while keeping the intuition of the agreement rate and requiring only easy and familiar computations.

Endnote

[1] Professor, Consumer Economics