Using Standard and Asymmetric Confidence Intervals

Utilisation d'intervalles de confiance standards et asymétriques

Christopher J. Lee

KEY WORDS

CONFIDENCE INTERVAL

MEASUREMENT ERROR

CONSISTENCY COEFFICIENT

AGREEMENT COEFFICIENT

Abstract

All assessment procedures are affected by measurement errors that alter the score obtained by a client. The particular score obtained by a client is one score within a hypothetical family of scores, each score in this family differing from the others as a result of measurement error. A confidence interval describes this family of scores by placing an error band around a client score. This paper describes the calculation of a consistency coefficient, an agreement coefficient, and the role of these coefficients in calculation of a confidence interval. The working example is an inter-rater situation in which ratings of speech intelligibility are made by two speech-language pathologists. However, a confidence interval can be used in a variety of other situations in which the comparability of scores is an issue. Spreadsheet software is shown to be a practical method of performing these calculations.

Abrégé

Toutes les procédures d'évaluation sont affectées par des erreurs de mesure qui changent le résultat obtenu par le client. Le résultat particulier obtenu par le client n'est qu'un résultat parmi une famille hypothétique de résultats; chaque résultat, dans cette famille, diffère des autres à cause d'une erreur de mesure. Un intervalle de confiance décrit cette famille de résultats en plaçant une marge d'erreur autour du résultat du client. Cet article décrit le calcul d'un coefficient de consistance et d'un coefficient d'accord ainsi que le rôle de ces coefficients dans le calcul d'un intervalle de confiance. L'exemple fourni est une situation inter évaluateurs où deux orthophonistes évaluent l'intelligibilité de la parole. Toutefois, un intervalle de confiance peut être utilisé dans plusieurs autres situations où la comparabilité des résultats peut être problématique. Un logiciel de tableurs s'avère être une méthode pratique pour effectuer ces calculs.

Christopher Lee, PhD,
School of Health Studies,
Labatt Health Sciences Building,
University of Western Ontario,
London, ON
CANADA

The inconstancy of things in the world is an enduring issue. As reported by Plato, in the dialogue *Cratylus*, the philosopher Heraclitus said that "all things pass and nothing stays, and comparing existing things to the flow of a river, he says you could not step twice into the same river." In the context of clinical assessment, the issue of inconstancy is a matter of measurement error. All assessment procedures are affected by measurement errors that alter the score obtained by a client. The particular score obtained by a client is one score within a hypothetical family of scores, each score in this large family differing from the others as a result of measurement error. Even with diligent effort to hold assessment conditions constant, it is unlikely that the same score would be observed when a client is tested a second time.

This paper shows a confidence interval to be a practical method of describing the effect of measurement error by placing an error band around the score obtained by a client. In a situation where a standardized assessment is used, the psychometric information needed to construct a confidence interval may be supplied by the publisher of the assessment, or it might be found in a research article. In situations where this information is not available, however, it is necessary to estimate the level of measurement error before calculating a confidence interval. The first section of the paper describes a consistency coefficient and an agreement coefficient as two indices of the level of measurement error, and it shows the calculation of these coefficients using spreadsheet software. The second section of the paper explains the use of these coefficients in calculation of a confidence interval. It shows the use of a standard symmetric confidence interval to describe the effect of unsystematic error on the score obtained by a client, and it introduces an adaptation of the standard confidence interval to describe the combined effect of unsystematic error and systematic bias on the score obtained by a client.

The working example in this paper is an inter-rater situation in which ratings of speech intelligibility are made by two speech-language pathologists. However, as noted at the end of the paper, the method can also be used in an intra-rater situation where a client is rated on two occasions by the same therapist, or a retest situation where two versions of a standardized assessment are used.

A Consistency Coefficient and an Agreement Coefficient

There are two basic forms of measurement error. Unsystematic error is a form of measurement error that increases or deceases individual scores by an unpredictable amount. In comparison, systematic bias is a form of measurement error that changes every score in the same direction and by a predictable amount. A consistency coefficient indicates the degree to which client scores are affected by unsystematic error, and an agreement coefficient indicates the degree to which client scores are affected by a combination of unsystematic error and systematic bias. Both coefficients range in value from zero to one, with higher values indicating that scores are less affected by measurement error.

A consistency coefficient and an agreement coefficient can be calculated based on means and variances obtained using spreadsheet software. For example, Figure 1 shows an Excel spreadsheet with ratings of speech intelligibility made by two speech-language pathologists (Anya and Beata). These ratings were made using a ten-point scale, with higher ratings indicating better intelligibility. Both therapists rated the same sample of six clients (labeled A to F). Anya's ratings are listed in rows 2 to 7 of column B, and Beata's ratings are listed in rows 2 to 7 of column C. A difference score was calculated for each client by subtracting Anya's rating from Beata's rating; these differences are listed in rows 2 to 7 of column D.

In a spreadsheet, means are calculated using the AVERAGE( ) function. In Figure 1, for example, the mean of Anya's ratings was calculated by typing = AVERAGE(B2:B7) into cell B9; the mean of Beata's ratings was calculated by typing = AVERAGE(C2:C7) into cell C9; and, the mean difference was calculated typing = AVERAGE(D2:D7) into cell D9, or more simply, by typing = C9–B9 into cell D9. As shown in the spreadsheet, the mean of Anya's ratings is 4.50, the mean of Beata's rating is 6.00, and the mean difference is 1.50.

A variance, designated by the symbol s^2, is a measure of the variability of the ratings. In a spreadsheet, a variance is calculated using the VAR( ) function. The variance of Anya's ratings was calculated by typing = VAR(B2:B7) into cell B10; the variance of Beata's ratings was calculated by typing = VAR(C2:C7) into cell C10; and, the variance of the difference was calculated by typing = VAR(D2:D7) into cell D10. As shown in the spreadsheet, the variance of Anya's ratings is 4.70, the variance of Beata's rating is 4.40, and the variance of the difference is 0.30.

Consistency Coefficient

In a situation with two raters, the variance of the difference is our estimate of the variance due to unsystematic error. A consistency coefficient (ICC_C) is calculated by dividing the variance of the difference (s^2_{diff}) by the sum of the variances of the two ratings ($s^2_1 + s^2_2$), and then subtracting the result from one. In our working

example, the consistency coefficient is calculated to be 0.969, as follows.

$$ICC_C = 1 - \frac{s^2_{diff}}{(s^2_1 + s^2_2)} = 1 - \frac{0.30}{(4.70 + 4.40)} = 0.967$$

In the spreadsheet (Figure 1), this calculation was made by typing =1−D10/(B10+C10) into cell D12.

Agreement Coefficient

In the calculation of an agreement coefficient, it is necessary to estimate the variance due to systematic bias. In a situation with two raters, the variance due to systematic bias (s^2_{bias}) is estimated by dividing the variance of the difference by the number of clients (n), and subtracting the result from the square of the mean difference. In the example, the variance of the difference is 0.30, the number of clients is 6, the mean difference is 1.50, and the variance due to systematic bias is calculated to be 2.20. This calculation was made in the spreadsheet (Figure 1) by typing =D9^2−D10/6 into cell D11. In algebraic notation, the calculation would be written as follows.

$$s^2_{bias} = M^2_{diff} - \frac{s^2_{diff}}{n} = 1.50^2 - \frac{0.30}{6} = 2.20$$

The agreement coefficient (ICC_A) is calculated by dividing the sum of the variances due to unsystematic error and systematic bias by the sum of the variances of the ratings and systematic bias, and then subtracting the result from one. In the spreadsheet (Figure 1), the agreement coefficient was calculated to be 0.779 by typing =1−(D10+D11)/(B10+C10+D11) into cell D13. In algebraic form, it would be written:

$$ICC_A = 1 - \frac{(s^2_{diff} + s^2_{bias})}{(s^2_1 + s^2_2 + s^2_{bias})} = 1 - \frac{(0.30 + 2.20)}{(4.70 + 4.40 + 2.20)} = 0.779$$

In statistical terms, a consistency coefficient and an agreement coefficient are examples of intraclass correlation coefficients (*ICC*). The term *intraclass correlation* refers to a family of coefficients used to describe the relationship between measurements of the same characteristic, such as between two ratings of speech intelligibility. In comparison, the more familiar Pearson correlation coefficient, routinely taught in introductory statistics courses, can be thought of as an "inter-class" correlation (McGraw & Wong, 1996) because it is used to describe the relationship between measurements of two different characteristics, such as between a rating of speech intelligibility and an index of stroke severity. I have used the abbreviation ICC_C to designate a consistency coefficient and the abbreviation ICC_A to designate an agreement coefficient. Other authors have used different notional schemes. For comparison, ICC_C in this paper corresponds to ICC(C,1) in McGraw and Wong (1996) or ICC(3,1) in Shrout and Fleiss (1979), and ICC_A in this paper corresponds to ICC(A,1) in McGraw and Wong (1996) or ICC(2,1) in Shrout and Fleiss (1979).

Lastly, it should be noted that the formulas and calculations shown above involve two raters. The same general method is involved in a situation with more than two raters, but these calculations require different formulas, and they are usually left to specialized software. In SPSS, for instance, consistency and agreement coefficients are optional statistics available in the Reliability Analysis procedure listed under the Scale heading in the Analysis menu. In the Reliability Analysis dialog, specify the Model as "alpha". Then, select "statistics" to open a secondary dialog listing optional statistics, and select "intraclass correlation coefficient", set Model to "two-way mixed", and specify Type as either "consistency" or "absolute agreement". Run the analysis, and in the resulting SPSS output, the intraclass correlation listed for "single measures" corresponds to the calculations shown above.

Confidence Interval

In our working example, the speech intelligibility of a small sample of clients was rated by two speech-language pathologists. This was done for the purpose of calculating a consistency coefficient and an agreement coefficient. Presumably, there are numerous other clients, now and in the future, for whom speech intelligibility is rated. But, rather than having two therapists rate each and every client, it is more practical to obtain a rating of speech intelligibility from one therapist, and then estimate the extent to which the rating of another therapist is likely to differ. This practical aim is served by calculating a confidence interval.

As noted earlier, any particular rating is one score within a hypothetical family of scores, each score in this family differing from the others as a result of measurement error. A confidence interval describes this family of scores by placing an error band around the score obtained by a client.

The width of a confidence interval is equal to the upper bound of the interval minus the lower bound of the interval. The width of the interval depends on the desired level of confidence and the amount of measurement error. The level of confidence is established using a z score. A z score of 1.96 is used to obtain a 95 percent confidence interval, and a z score of 1.64 is used to obtain a 90 percent

confidence interval. The amount of measurement error is expressed in terms of a *standard error*. As shown below, a standard symmetric confidence interval is calculated using a *consistency standard error*. An asymmetric confidence interval is calculated using a consistency standard error and an *agreement standard error*.

Standard Symmetric Confidence Interval

A standard symmetric confidence uses a consistency standard error to represent the effect of unsystematic error. A consistency standard error (SE_C) is calculated using the variances of the two ratings and the consistency coefficient, as follows.

$$SE_C = \sqrt{(s_1^2 + s_2^2)(1 - ICC_C)}$$

In our working example, the variance of Anya's ratings is 4.70, the variance of Beata's ratings is 4.40, the consistency coefficient is 0.967, and the consistency standard error is calculated to be 0.548.

$$SE_C = \sqrt{(4.70 + 4.40)(1 - 0.967)} = 0.548$$

In the spreadsheet (Figure 1), this calculation was made by typing =SQRT((B10+C10)*(1−D12)) into cell D14.

A standard confidence interval is symmetric around a client score (X). It ranges from a lower bound of z times SE_C below the score to an upper bound of z times SE_C above the score. The confidence interval extends equally above and below the score because unsystematic error is equally likely to increase a score or decrease a score.

$$X - z\,SE_C \sim X + z\,SE_C$$

Using a z score of 1.96 and a consistency standard error of 0.548, a 95 percent symmetric confidence interval is found to range from a lower bound of 1.1 (= 1.96 × 0.548) below the obtained rating to an upper bound of 1.1 above the obtained rating.

$$X - 1.1 \sim X + 1.1$$

Now, consider a new client, named Zappora, who is assessed by Anya and given a speech intelligibility rating of 5. Our confidence interval indicates that Zappora is likely to obtain a rating that falls between 3.9 (= 5 – 1.1) and 6.1 (= 5 + 1.1) if she were assessed by Beata. In using this confidence interval, it is assumed that Zappora's actual level of speech intelligibility has not changed, and the two ratings differ entirely as a result of unsystematic error in the rating procedure.

Asymmetric Confidence Interval

An asymmetric confidence interval is an adaptation of the standard symmetric confidence interval to address the combined effect of unsystematic error and systematic bias. Systematic bias is a directional form of measurement error. The effect of *positive systematic bias* is to move the upper bound of a confidence interval in an upward direction, whereas the effect of *negative systematic bias* is to move the lower bound of a confidence interval in a downward direction.

Systematic bias is indicated when the estimated variance due to systematic bias (cell D11) is greater than zero. Whether the direction of this systematic bias is positive or negative depends on which rater's score is estimated by the confidence interval. The direction is *positive* when a confidence interval is used to estimate a score made by the rater who, on average, gives *higher* ratings, whereas the direction is *negative* when a confidence interval is used to estimate a score made by the rater who, on average, gives *lower* ratings. In our working example, Beata's ratings are higher, on average, than Anya's ratings. Thus, a confidence interval that is used to estimate a rating made by Beata is subject to positive systematic bias, whereas a confidence interval that is used to estimate a rating made by Anya is subject to negative systematic bias.

An asymmetric confidence interval is calculated by using an *agreement standard error* to widen the biased side of the confidence interval. An agreement standard error (SE_A) is calculated using the variances of the two ratings and the agreement coefficient, as follows.

$$SE_A = \sqrt{(s_1^2 + s_2^2)(1 - ICC_A)}$$

In a situation with positive systematic bias, the upper bound is calculated using an agreement standard error, while the lower bound is calculated using a consistency standard error. Thus, a positive asymmetric confidence interval ranges from a lower bound of z times SE_C below the score to an upper bound of z times SE_A above the score.

$$X - z\,SE_C \sim X + z\,SE_A$$

In a situation with negative systematic bias, the lower bound is calculated using an agreement standard error, while the upper bound is calculated using a consistency standard error. Thus, a negative asymmetric confidence interval ranges from a lower bound of z times SE_A below the score to an upper bound of z times SE_C above the score.

$$X - z\,SE_A \sim X + z\,SE_C$$

In our working example, the variance of Anya's ratings is 4.70, the variance of Beata's ratings is 4.40, the agreement coefficient is 0.779, and the agreement standard error is equal to 1.419.

$$SE_A = \sqrt{(4.70 + 4.40)(1 - 0.779)} = 1.419$$

In the spreadsheet (Figure 1), this calculation was made by typing =SQRT((B10+C10)*(1−D13)) into cell D15.

Once again consider the case of Zappora, who is assessed by Anya and given a speech intelligibility rating of 5. This is a situation where a positive asymmetric confidence interval could be employed because Beata's ratings are higher, on average, Anya's ratings. Using a z score of 1.96, a consistency standard error of 0.548, and an agreement standard error of 1.419, the 95 percent positive asymmetric confidence interval is found have a lower bound of 1.1 (= 1.96 × 0.548) below a rating made by Anya to an upper bound of 2.8 (= 1.96 × 1.419) above a rating made by Anya.

$$X - 1.1 \sim X + 2.8$$

On this basis, Zappora is likely to obtain a rating that falls between 3.9 (= 5 – 1.1) and 7.8 (= 5 + 2.8) if she were rated by Beata.

Lastly, consider one other client, Ambrose, who is assessed by Beata and given a speech intelligibility rating of 6. This is a situation where a negative asymmetric confidence could be employed because Anya's ratings are lower, on average, than Beata's ratings. Using a z score of 1.96, a consistency standard error of 0.548, and an agreement standard error of 1.419, the 95 percent negative asymmetric confidence interval is found to range from a lower bound of 2.8 (= 1.96 × 1.419) below a rating made by Beata to an upper bound of 1.1 (= 1.96 × 0.548) above a rating made by Beata.

$$X - 2.8 \sim X + 1.1$$

In the case of Ambrose, the asymmetric confidence interval indicates that he is likely to obtain a rating that falls between 3.2 (= 6 – 2.8) and 7.1 (= 6 + 1.1) if he were rated by Anya.

In closing, it is important to note that it is mathematically possible for the value of SE_A to be less than the value of SE_C. But, in principle, the value of SE_A must be greater than, or equal to, the value of SE_C because the former represents unsystematic error plus systematic bias whereas the latter represents unsystematic error alone. In a situation where SE_A is less than SE_C, it should be assumed that there is no systematic bias and SE_C should be used to establish both bounds of a confidence interval.

Summary

All assessment procedures are affected by measurement errors that alter the score obtained by a client. The particular score obtained by a client is one score within a hypothetical family of scores, each score in this family differing from the others as a result of measurement error. A confidence interval describes this family of scores by placing an error band around the score obtained by a client. In a standard symmetric confidence interval, the error band around a client score represents the effect of unsystematic error. An asymmetric confidence interval is a useful adjunct to the standard confidence interval because it describes the combined effect of unsystematic error and systematic bias.

Our working example has described an inter-rater situation in which ratings of speech intelligibility are made by two speech-language pathologists. However, the methods presented above can also be used in an intra-rater situation in which a client is rated on two occasions by one therapist, or a retest situation in which two versions of a standardized assessment are used. In an intra-rater situation, columns B and C of the spreadsheet (Figure 1) would contain a set of initial ratings and a set of subsequent ratings of the same clients by the same therapist. A confidence interval around an initial rating describes a range of subsequent ratings that are likely to occur as a result of measurement error. In a retest situation, columns B and C would list the scores obtained by testing each client twice, once using one version and once using the other version. A confidence interval placed around a score obtained with one version describes a range of scores obtainable with the other version as a result of measurement error. In these situations (inter-rater, intra-rater, retest) a score is expected to fall inside the confidence interval in the absence of any form of treatment. A client who receives treatment would be expected to fall outside the confidence interval because

her or his score is expected to differ as a result of treatment as well as measurement error. In sum, a confidence interval is a highly practical method of addressing the effect of measurement error in a variety of situations in which the comparability of scores is an issue.

References

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*, 30-46.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428.

Authors' Note

Correspondence concerning this article should be addressed to Christopher Lee, PhD, School of Health Studies, Labatt Health Sciences Building, University of Western Ontario, London, ON, Canada, N6A 5B9. Email: cjlee@uwo.ca.

	A	B	C	D
1	Client	Anya	Beata	Difference
2	A	2	3	1.0
3	B	2	4	2.0
4	C	4	6	2.0
5	D	6	7	1.0
6	E	6	8	2.0
7	F	7	8	1.0
8				
9	Mean	4.50	6.00	1.50
10	Variance	4.70	4.40	0.30
11	Variance (Bias)			2.20
12	ICC_C			0.967
13	ICC_A			0.779
14	SEc			0.548
15	SE_A			1.419

Figure 1. A spreadsheet with example data and calculated values used in determining a confidence interval.