Chapter 6
Statistical Assessment of Agreement

6.1 General Introduction to Agreement

Researchers have become increasingly aware of the problem of assessing agreement since more than one and a half century in the past. There are numerous examples that illustrate these situations, and here we list some of them. In clinical and medical measurement comparison of a newly developed measurement method with an established one, it is often desired to check whether they agree sufficiently and accurately enough for the new to replace the old. The new method of measurement is most often cheaper, quicker, and suboptimal; however, it needs a thorough and careful examination to see if it can effectively replace the old one. In criminal trials, a group of jurors are used and sentencing depends on the complete agreement among the jurors. Hotels receive five-star recognition only after several experts and designated visitors agree on the services and facilities rendered by the hotels. The medals and ranking in sport games are based on the ratings provided by several judges.

It has now become generally accepted that measurements of agreement are needed to assess the acceptability of new or generic process, methodology, and formulation in both science and non-science fields of laboratory performance, instrument or assay validation, method comparisons, statistical process control, goodness of fit, and individual bioequivalence. Examples include the agreement of laboratory measurements collected through various laboratory instruments, the agreement of a newly developed method with gold standard method, the agreement of manufacturing process measurements with specifications, the agreement of observed values with predicted values, and the agreement in bioavailability of a new or generic formulation with a commonly used formulation. By the way, measuring agreement has been used very often to designate the level of agreement between different data-generating sources, commonly referred to as observers or raters. A rater could be a chemist, a psychologist, a radiologist, a clinician, a nurse, a rating system, a diagnosis, a treatment, an instrument, a method, a process, a technique or a formula, to mention a few. Elementary to advanced statistical methods have been used over time to assess the level of

This chapter draws material from co-published work of one of the authors: 'Some further aspects of assessment of agreement involving bivariate normal responses,' published in Int'l Jour. of Statistical Sciences, Vol. 13, 2013, pp. 1–19. Portions have been used here with permission from the Author Dr. Ganesh Dutta as well as Publisher.

© Springer Nature Singapore Pte Ltd. 2018

S. P. Mukherjee et al., *Statistical Methods in Social Science Research*,
https://doi.org/10.1007/978-981-13-2146-7_6

agreement between different data-generating sources referred to above as observers or raters.

Cohen's Kappa statistic (1960) and weighted Kappa (1968) are the most popular indices for measuring agreement when the responses are nominal. Weighted Kappa statistic has been proposed by Landis and Koch (1977), and it is appropriate for assessing agreement when the categories of response are ordinal. Several authors have proposed guidelines for the interpretation of kappa statistic. Vide, for example, Landis and Koch (1977), Fleiss (1981), Bland and Altman (1986), and Kraemer et al. (2002). A comprehensive review paper is also worth reporting (Banerjee et al. 1999). Recently, some studies have been undertaken to critically examine certain aspects of Cohen's Kappa. These relate to its attaining the negatively extreme value and its standardization. See Pornpis et al. (2006).

Extensions have also been made to allow for more than two raters, more than two possible ratings, ordinal data and continuous data. In addition, many other applications of kappa statistic in a variety of different contexts can be found in the literature. A reference book in this area is by Eye and Mun (2005). Another book dealing with both categorical and continuous measurements for multiple raters and multiple ratings is by Shoukri (2004).

Lin (1989) introduced the concordance correlation coefficient (CCC) for measuring agreement which is more appropriate when the data are measured on a continuous scale. A weighted CCC was proposed by Chinchilli et al. (1996) for repeated measurement designs and a generalized CCC for continuous and categorical data was introduced by King and Chinchilli (2001). Lin (2000) also introduced total deviation index (TDI) for measuring individual agreement with applications in laboratory performance and bioequivalence. Further to this, Lin et al. (2002) proposed methods for checking the agreement in terms of coverage probability(CP) when the two measurements are quantitative in nature.

When the study of agreement involves three or more raters on a continuous scale, there are different approaches to follow. Two most recent references are (i) Barnhart et al. (2007) and (ii) Lin et al. The authors broadly follow (i) ANOVA and (ii) modeling approach to examine the extent of agreement. The approach proposed and studied in Lin et al. (2002) has been extended in Hedayat et al. (2009) for the case of multiple raters.

We will touch upon some of the techniques developed for study of agreement involving both types of data.

6.2 Cohen's Kappa Coefficient and Its Generalizations: An Exemplary Use

There are many instances of applications of the basic technique for assessing agreement between two raters, in case the subjects are rated according to a binary feature, to be designated as Yes and No. Cohen's Kappa (1960) was suggested in the agreement literature with this specific purpose. Generalizations and extensions to other

contexts were brought in from time to time. We will discuss at length one study carried out in a hospital in Bangkok [Pornpis et al. (2006)].

Rajavithi Hospital, Bangkok, Thailand houses Thai Screening for Diabetic Retinopathy Study Group in its Department of Ophthalmology. Three MD doctors Dr. Paisan Ruamviboonsuk, Dr. Khemawan Teerasuwanajak, and Dr. Kanokwan Yuttitham carried out a revealing diagnostic study in this specialist eye hospital having [in-house and confined to hospital beds] 600+ diabetic patients. All the patients were under treatment for diabetic retinopathy of different degrees of severity. The study was based on randomly selected 400/600+ diabetic patients and from each selected patient, one good single-field digital fundus image was taken with signed consent and with due approval by Ethical Committee on Research with Human Subjects.

The purpose was to extract information from each image on three major features:

(i) Diabetic Retinopathy Severity [6 options]:

No Retinopathy/Mild/Moderate NPDR/Severe NPDR/PDR/Ungradable;

(ii) Macular Edema [2 options]: Presence/Absence/Ungradable;

(iii) Referral to Ophthalmologists [2 options]: Referrals / Non-Referrals / Uncertain.

These features were extracted by (i) Retina Specialists [3], (ii) General Ophthalmologists [3], (iii) Photographers [3] and (iv) Nurses [3]—all engaged in their respective meaningful professions within the hospital. It thus transpires that altogether 12 raters collected data on each of the 3 features mentioned above and from each of the 400 images so collected. Therefore, the study group was loaded with massive amount of data.

The objective of the research study was to examine the extent of agreement within and between different Expert Groups and to provide adequate interpretation of the results. It is believed that all the 12 experts/raters examined the images independently of one another.

As noted from the above, items (ii) and (iii) deal mostly with binary response [Presence versus Absence or Referral versus Non-Referral] data while item (i) deals with multi-response categorical data. We will slightly modify item response for (ii) to give it a shape of binary response data. It is revealed that the first two Retina Specialists RS1 and RS2 independently counted the respective Presence–Absence responses [in respect of the Feature: Macular Edema] as: 337 versus 40 and 344 versus 33. This indeed showed remarkable agreement among them upfront [89 versus 11 percent and 91–9 percent]! It was too good to be acceptable. The study group wondered about the validity of the findings and contacted Dr Montip Tiensuwan, Statistics Faculty, Department of Mathematics, Mahidol University, Bangkok. Dr Tiensuwan had already studied the literature on Statistical Assessment of Agreement and worked with one of the authors of this article [Sinha]. Her collaboration with the Hospital Study Group was successful, and it eventually resulted in a good journal publication. We will now elaborate on the major findings of their study.

It is clear that each image was inspected by each of the three RSs, and hence, it is possible to examine the scope of agreement more closely before deciding on its extent. As is stated above, RS1 and RS2 largely agreed on classification of patients into Presence–Absence Categories w.r.t. Macular Edema. But this only reflected what

is called marginal nature of binary classification. We are also in a position to check case by case the nature of agreement or otherwise of RS1 and RS2. For example, when pairwise ratings given by $RS1$ and $RS2$ are considered for each of the 377 patients, we find that

$$[(Y, Y) : 326/377; (Y, N) : 11/377; (N, Y) : 18/400; (N, N) : 22/377]$$

- the 'marginal' totals being $[RS1(Y) : 337/377; RS2(Y) : 344/377]$, as was specified above. It transpires that there are altogether $29/377 = 89$ percent cases of disagreement between the two raters. In effect, therefore, $RS1$ and $RS2$ are in very good agreement. And this Cohen designated as observed agreement, denoted by θ_0. According to Cohen, this is only half of the story and it could as well be due to what he assigned as chancy agreement! The idea is that two so-called experts could purely agree by chance—by making assessments independently. Using elementary probability formula, he computed the contribution from chancy agreement as:

$$\theta_e = P[Y, Y] + P[N, N] = P[Y, .]P[., Y] + P[N, .]P[., N]$$

by referring to the 'marginal probabilities.' According to this formula, for the above data set, chancy agreement, denoted by θ_e is computed as 82.50 percent! Cohen then suggested 'chance-corrected' agreement index as

$$\kappa = \frac{\theta_0 - \theta_e}{1 - \theta_e}.$$

Computation yields $\kappa = 56$ percent which suggests a moderate level of agreement only. Likewise, it is a routine task to compute κ coefficient between $RS1$ and $RS3$ or, between $RS2$ and $RS3$. It may be noted that the κ coefficients do not obey any transitivity law.

This study became instantly famous because of the following special feature. For any group of 3 Experts [Retina Specialists/General Ophthalmologists/etc/etc], the purview of the study also captured Consensus Rating [CR] of the raters for each feature. Thus, for example, in respect of Macular Edema, there was a Consensus Rating given collectively by the 3 RSs as follows: [Presence: 355/400; Absence: 35/400; Ungradable: 10/400]. Subsequently, κ coefficient was computed for the RSs as against the CR[RS] one by one.

Also for that matter, we can compute κ values in respect of the feature (iii), by restricting to the 2×2 case of binary response, neglecting the uncertain category. We will skip the details.

So far as the feature in item (i) is concerned, we need to be careful in assessing the extent of agreement between any two raters [or between a rater of a category and the CR of the same category]. This is because we are now dealing with six categories of response in respect of the status of Diabetic Retinopathy [DR] as mentioned in (i). Cohen's original idea of computation of κ, based on θ_0 and θ_e, does not pose any difficulty anyway. First of all, we can visualize the response count data for a pair of experts as forming a table of order 6×6 with the percentage counts along the main

diagonal [say, f_{ii}/n for the ith category of response] serving as constituents of θ_0. By the same token, products of percentage counts [based on the notion of independence] such as $(f_{i,.}/n)(f_{.,i}/n)$ will add up to the computation for θ_e. Then the formula for κ can be routinely applied. This was done and sooner or later, it drew criticism! We will take up the data set for $RS1$ versus $RS2$ and examine the matter below.

We will follow the codes: Code I - No Retinopathy; Code II - Mild; Code III - Moderate NDPR; Code IV Severe NPDR; Code V PDR and Code VI: Ungradable.

Along the main diagonal, the percentage of observed agreement θ_0 amounts to 80.50 percent. Further, direct computation yields for $\theta_e = 48.60$ percent. Hence, $\kappa = 62$ percent a very moderate level of agreement. The criticism has been based on the following arguments: Pairwise categories

$$[(Code\ I, Code\ II), (Code\ II, Code\ I), (Code\ II, Code\ III), (Code\ III, Code\ II) etcetc]$$

represent what may be termed as 'narrowly missed' cases. Cohen's κ does not take cognizance of these narrowly missed cases/classes and attributes no credit whatsoever to the raters. It is argued that one should make a case of allowing for partial credits to be attributed to such and similar categories. In contrast to Cohen's original κ—termed henceforth as Unweighted κ—weights were assigned to all the categories and κ was modified to Weighted κ, written as $\kappa(W)$. It is computed along similar lines as

$$\kappa(W) = (\theta(W)_0 - \theta(W)_e)/(1 - \theta(W)_e)$$

where $f_{ij}W_{ij}$s are used in the computation of $\theta(W)_0$ and $f_{i.}f_{.j}W_{ij}$s are used in the computation of $\theta(W)_e$. The choice of the weight matrix $\mathbf{W} = ((\mathbf{W_{ij}}))$ has not been any smooth matter. Reasonable and acceptable choice of the weight matrix of dimension R have the elements $W_{ij} = 1 - (i - j)^2/(R - 1)^2$.

Weighted κ statistics were calculated for pairs of raters, including comparison against the CR in respect of all the three features listed in (i), (ii), and (iii). The results are shown in the Appendix.

This study also covered another important aspect of comparison of expertise across different specialist groups. In the published literature, there are formulae available to account for this kind of comparison. Applied to this case, a measure of composite performance of 3 Retina Specialists/3 Ophthalmologists/3 Technicians/3 Nurses for each of the 3 features was computed. For example, for DR, it was revealed that composite performance indices are

$$RS - 0.58;\ Oph. - 0.36;\ Tech. - 0.37,\ Nurses - 0.26.$$

Likewise, for Macular Edema, the values are: [0.58, 0.19, 0.38, 0.20] and for Referral, these are: [0.63, 0.24.0.30, 0.20].

It transpired that except for the Retina Specialists, no other categories of so-called experts showed any visible mode of agreement in any of the features. Of all 400 cases, 44 warranted Referral to Ophthalmologists due to Retinopathy Severity and 5 warranted Referral to Ophthalmologists due to uncertainty in diagnosis. Fourth Retina Specialist carried out dilated fundus examination of these 44 patients, and

substantial agreement [$\kappa = 0.68$] was noticed for DR severity examination confirmed Referral of 38/44 cases.

In conclusion, it is stated that Retina Specialists are all in active clinical practice and hence are most reliable for digital image interpretation of images. Individual Raters' background and experience play roles in digital image interpretation expertise. Unusually, high percentage of images were declared as ungradable by nonphysician raters, though only 5 out of 400 were declared as ungradable by consensus of the Retina Specialists Group. Lack of confidence of non-physicians, rather than true image ambiguity, is likely to be a realistic reason for this. For this study, other factors [blood pressure, blood sugar, cholesterol, etc.] had not been taken into account.

6.3 Assessment of Agreement in Case of Quantitative Responses

In this section, we focus on the feature of agreement involving data for two competing raters measured on a continuous scale. There are several usual approaches for evaluating agreement for such paired data such as Pearson correlation coefficient, regression analysis, paired t-tests, least-squares analysis for slope and intercept, within subject coefficient of variation, and intra-class correlation coefficient.

The concordance correlation coefficient (CCC) was first proposed by Lin (1989) for assessment of agreement in continuous data. It represents a breakthrough in assessing agreement between two raters for continuous data in that it appears to avoid all the shortcomings associated with usual approaches in some situations. In short, Lin (1989) expresses the degree of concordance between two variables X and Y by the Mean Squared Deviation (MSD), $E(X-Y)^2$ and defines the CCC as

$$\rho_c = 1 - \frac{\mathrm{E(Y-X)^2}}{\mathrm{E_{Indep}(Y-X)^2}} = \frac{2\sigma_{xy}}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{6.3.1}$$

where $\mathrm{E_{Indep}}(.)$ represents expectation under the assumption of independence of X and Y, $\mu_x = \mathrm{E(X)}$, $\mu_y = \mathrm{E(Y)}$, $\sigma_x^2 = \mathrm{Var(X)}$, $\sigma_y^2 = \mathrm{Var(Y)}$, and $\sigma_{xy} = \mathrm{Cov(X,Y)} = \rho\sigma_x\sigma_y$.

It is readily seen that ρ_c can be expressed as

$$\rho_c = \rho \times \frac{2\sigma_1\sigma_2}{(\mu_x - \mu_y)^2 + (\sigma_x^2 + \sigma_y^2)}$$

Further to this, it follows that

$$\rho_c = 1 \; if \; and \; only \; if \; [\rho = 1, \mu_x = \mu_y; \sigma_x = \sigma_y].$$

Lin (1989) estimates this CCC [ρ_c] with data by substituting the sample moments of bivariate sample into above formula to compute the sample counterpart of CCC (ρ_c). The CCC translates the MSD into a correlation coefficient that measures the

agreement along the identity line. It has the properties of a correlation coefficient in that it ranges between -1 and $+1$, with -1 indicating perfect reversed agreement ($Y = -X$), 0 indicating no agreement, and $+1$ indicating perfect agreement ($Y = X$). Lin et al. (2002) gave a review and comparison of various measures, including the CCC, of developments in this field by comparing the powers of the tests:

(1) $\mu_x = \mu_y$, (2) $\sigma_x = \sigma_y$, and (3) $\rho = \rho_0$, where ρ_0 is a given value, assumed to be substantially high.

Their calculation is illustrated using a real data example. This work was further extended in Hedayat et al. (2009) involving multiple raters. In another direction, Yimprayoon et al. (2006) extended the work of Lin et al. (2002) by combining the problems of testing for $\mu_x = \mu_y$, $\sigma_x = \sigma_y$, and $\rho \geq \rho_0$ into one overall testing problem under bivariate normal setup and then they presented the result based on simulation study.

An intuitively clear measurement of agreement is a measure that captures a large proportion of data within a predetermined boundary from the line of agreement, i.e., $X = Y$. In other words, we want the probability of the absolute value of $D = Y - X$ less than the specified boundary, k, to be large. This probability is termed in the literature as coverage probability (CP) (cf. (Lin et al. 2002)), and it is defined as

$$\mathrm{CP}(k) = \mathrm{P}[|D| < k], \tag{6.3.2}$$

where X and Y denote random variables representing paired observations for assessing the agreement. It is generally assumed that X and Y have a bivariate normal distribution with means μ_x and μ_y, variances σ_x^2 and σ_y^2 and correlation coefficient ρ so that the covariance of X and Y is $\sigma_{xy} = \rho\sigma_x\sigma_y$.

The multiparameter hypothesis involving (6.3.1), (6.3.2), and (6.3.3) displayed above is too demanding for agreement between the two raters. Therefore, a more appropriate and plausible null hypothesis can be formulated as

$$\mathrm{H}_0 : \ |\mu_x - \mu_y| \geq \varepsilon_0, \ \frac{\sigma_x}{\sigma_y} \ or \ \frac{\sigma_y}{\sigma_x} \geq \eta_0, \ \rho \leq \rho_0 \tag{6.3.3}$$

where ε_0 is close to zero and η_0 and ρ_0 are close to unity—all are assumed to be specified. A large sample test [known as Likelihood Ratio Test] of this hypothesis has been worked out in Dutta and Sinha (2013).

References and Suggested Readings

Anderson, S., & Hauck, W. W. (1990). Consideration of individual bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics*, *18*, 259–273.

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis*. New York: Wiley.

Banerjee, M., Capozzoli, M., McSweeney, L., & Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *The Canadian Journal of Statistics, 27*, 3–23.

Barnhar, H. X., Haber, M. J., & Lin, L. I. (2007). *An Overview On Assessing Agreement With Continuous Measurement*.

Bland, J. M., & Altman, D. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, *8*, 307–310.

Chinchilli, V. M., Martel, J. K., Kumanyika, S., & Lloyd, T. (1996). A weighted concordance correlation coefficient for repeated measures designs. *Biometrics, 52*, 341–353.

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

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*, 213–220.

Dutta, G., & Sinha, B. K. (2013). Some further aspects of assessment of agreement involving bivariate normal responses. *International Journal of Statistical Sciences*, *13*.

Fleiss, J. L. (1981). Statistical methods for rates and proportions. 2nd ed. (pp. 38–46). New York: John Wiley.

Hedayat, A. S., Lou, C., & Sinha, B. K. (2009). A statistical approach to assessment of agreement involving multiple raters. *Communications in Statistics - Theory & Methods*, *38*, 2899–2922.

Holder, D. J., & Hsuan, F. (1993). Moment-based criteria for determining bioequivalence. *Biometrika*, *80*, 835–846.

King, T. S., & Chinchilli, V. M. (2001). A generalized concordance correlation coefficient for continuous and categorical data. *Statistics in Medicine*. pp. 2131–2147.

Kraemer, H. C., Periyakoil, V. S., & Noda, A. (2002). Kappa coefficients in medical research. *Statistics in Medicine* (pp. 2109–2129).

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–74.

Lin, L., Hedayat, A. S., Sinha, B. K., & Yang, M. (2002). Statistical methods in assessing agreement: Models, issues, and tools. *Journal of the American Statistical Association*, *97*, 257–270.

Lin, L., Hedayat, A. S., & Wu, W. (2012). *Statistical tools for measuring agreement*. Berlin: Springer.

Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, *45*, 255–268.

Lin, L. I. (1992). Assay validation using the concordance correlation coefficient. *Biometrics*, *48*, 599–604.

Lin, L. I. (1997). Rejoinder to the letter to the editor by Atkinson and Nevill. *Biometrics*, *53*, 777–778.

Lin, L. I. (2000). Total deviation index for measuring individual agreement: With application in lab performance and bioequivalence. *Statistics in Medicine*, *19*, 255–270.

Lin, L. I., & Torbeck, L. D. (1998). Coefficient of accuracy and concordance correlation coefficient: New statistics for method comparison. *PDA Journal of Pharmaceutical Science and Technology*, *52*, 55–59.

Pornpis, Y., Tiensuwan, M., & Sinha, B. K. (2006). Cohen's kappa statistic: A critical appraisal and some modifications. *Calcutta Statistical Association Bulletin, 58*, 151–169.

Schall, R. (1995). Assessment of individual and population bioequivalence using the probability that bioavailabilities are similar. *Biometrics*, *51*, 615–626.

Schall, R., & Luus, H. G. (1993). On population and individual bioequivalence. *Statistics in Medicine*, *12*, 1109–1124.

Schall, R., & Williams, R. L. (1996). Towards a practical strategy for assessing individual bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics*, *24*, 133–149.

Sheiner, L. B. (1992). Bioequivalence revisited. *Statistics in Medicine*, *11*, 1777–1788.

Shoukri, M. M. (2004). *Measures of interobserver agreement*. Boca Raton: Chapman & Hall/CRC.

Von Eye, A., & Mun, E. Y. (2005). *Analyzing rater agreement: Manifest variable methods*.

Vonesh, E. F., & Chinchilli, V. M. (1997). *Linear and nonlinear models for the analysis of repeated measurements*. New York: Marcel Dekker.

Vonesh, E. F., Chinchilli, V. M., & Pu, K. (1996). Goodness-of-fit in generalised nonlinear mixed-effect models. *Biometrics*, *52*, 572–587.

Yimprayoon, P., Tiensuwan, M., & Sinha Bimal, K. (2006). Some statistical aspects of assessing agreement: Theory and applications (English summary). *Festschrift for Tarmo Pukkila on his 60th birthday* (pp. 327–346). Tampere: Department of Mathematics, Statistics and Philosophy, University of Tampere.