Available online at www.sciencedirect.com

ScienceDirect

Theriogenology 73 (2010) 1167–1179

Theriogenology

www.theriojournal.com

Invited Review

Method agreement analysis: A review of correct methodology

P.F. Watson [a,*], A. Petrie [b]

[a] *The Royal Veterinary College, London, United Kingdom*
[b] *The UCL Eastman Dental Institute, London, United Kingdom*

Abstract

The correct approach to analyzing method agreement is discussed. Whether we are considering agreement between two measurements on the same samples (repeatability) or two individuals using identical methodology on identical samples (reproducibility) or comparing two methods, appropriate procedures are described, and worked examples are shown. The correct approaches for both categorical and numerical variables are explained. More complex analyses involving a comparison of more than two pairs of data are mentioned and guidance for these analyses given. Simple formulae for calculating the approximate sample size needed for agreement analysis are also given. Examples of good practice from the reproduction literature are cited, and common errors of methodology are indicated.

Keywords: Agreement analysis; Reliability; Repeatability; Reproducibility; Sample size calculation

Contents

* Corresponding author. Tel.: +44 0 1920 466941.
E-mail address: pwatson@rvc.ac.uk (P.F. Watson).

0093-691X
doi:10.1016/j.theriogenology.2010.01.003

1. Introduction

A common question encountered in reproductive biology is whether or not the measurement of a variable by two different methods, or by two different operators using the same method, or by one operator repeating the measurement at two different times, produces essentially the same result. We are concerned both with accuracy (the way in which an observed value of a quantity agrees with the true value) and precision (a measure of the extent to which repeated observations conform). Examples might be the measurement of blood hormone concentrations or the use of two different techniques for determining pregnancy status. It is important to use appropriate statistical methods to address a question such as this.

For many years, it was common to use one of several incorrect methods to answer this question with the consequence of unsatisfactory or sometimes misleading conclusions. In this article, we will illustrate and highlight the correct approaches to address the problem of assessing the consistency of the measuring process using some examples drawn from the literature. An overview of the procedures discussed is given according to type of variable in Table 1.

1.1. Measurement variability and measurement error

When we measure a biological variable in a number of individuals or repeatedly within an individual (either within a short time or over a longer period), the data always exhibit, to a greater or lesser extent, a scatter of values. Inter-individual variation (between individuals) as well as intra-individual variation (within individual) is thus likely to be evident. Much of this variability is due to variation in associated factors (e.g., genetic, social, or environmental factors). For example, if these individuals differ in terms of their reproductive status, age, weight or gender, blood hormone measurements may be expected to vary. Similarly, if we take repeated measurements from an individual at different times of the day, they may well vary. This variability is termed *measurement variability*. In contrast, *measurement error* is defined as that which arises because the observed (or ''measured'') values and true values of a variable differ (note that although we refer to the ''true'' measurement here, it is rarely possible to obtain this value). Two kinds of measurement error can occur:

- *Random:* The observed values may be sometimes higher or lower than the true values, but on average they tend to balance out. For example, the measurement may be read on a scale to the nearest division. Although random error is governed by chance, the degree of error can be influenced by external factors (e.g., a balance may exhibit greater random variability when sited in a drafty location).
- *Systematic:* The observed values have a tendency to be consistently high (or low) because of some extraneous factor, known or unknown, affecting the measurements in the same way (e.g., because of an instrument that has not been calibrated correctly or an observer consistently overestimating the values). This kind of error, which concerns the overall accuracy of the observations, results in biased results if one set of results represents the true values. The error must be eliminated or minimized by attention to issues such as training of personnel, standardization of conditions of measurement, and proper calibration and maintenance of instruments (i.e., verification by comparison with a known standard).

Although this explanation of error has centered on laboratory measurements, the same concepts apply even if interest is focused on other forms of measurement,

Table 1
Summary of procedures for agreement analysis.

Number of methods to compare	Variable		Procedure
2	Categorical	2 categories	Cohen's kappa McNemar's test
		>2 ordered categories	Weighted kappa
	Numerical		Intraclass correlation coefficient Lin's concordance correlation coefficient Bland and Altman diagram Paired *t*-test British Standards reproducibility coefficient
>2	Consult an appropriate advanced text or a statistician		

such as an individual's assumed pregnancy status as assessed by a stockman's questionnaire. When establishing a measurement technique, we must consider both measurement variability and measurement error.

2. Reliability

In any quantitative biological study, we need to evaluate the consistency of the measuring process. A sample used for the reliability study should reflect that used for the investigative study. It is usual to carry out a reliability study as part of a larger investigative study. We want to know whether or not a particular method is stable enough to be of value. Will a second measurement in the same subject by the same observer under identical conditions be the same? In the previous section, it was emphasized that variation in measurement is inevitable, but *how much* variation is acceptable? In such circumstances, what is of interest is the evaluation of what is termed *repeatability* of the measurements. The intention is to establish a measure of the *within-observer agreement*.

Providing that the repeatability of a given procedure or observer is satisfactory, it is possible to assess what is commonly termed *reproducibility*. We are interested here to establish whether two persons using the same method of measurement obtain the same result or whether two techniques used to measure a particular variable, under identical circumstances, produce essentially the same result. In these circumstances, reproducibility is assessing the *between-method* or *between-observer agreement*. Understandably, if the repeatability has not been found to be acceptable, the reproducibility assessment under these circumstances will be unreliable.

Repeatability and reproducibility are measures of *reliability* and may be assessed in the same way. For simplicity, therefore, in this review we illustrate the statistical approach to measuring agreement by considering only one of these measures for a given situation, namely reproducibility for categorical data and repeatability for numerical data

It has been shown that very little advantage is gained from using more than three results per subject, and it is more efficient to compare only pairs or triplicates of results on a greater number of subjects rather than using a greater number of results on fewer subjects [1]. As a consequence, and because the statistical approach is much simpler using pairs of results, we restrict the comparison to two methods, with one member of every pair of results from each method in a reproducibility study and duplicate results in a repeatability study. In more complex circumstances, components of variance obtained from appropriate analyses of variance are commonly used as the tools for assessing agreement. (References to these techniques are given in Section 2.3 devoted to more complex analysis.)

The nature of the data determines the statistical approach to assessing reliability; we need to consider whether the variable of interest is categorical (e.g., poor/average/good) or numerical (e.g., serum hormone concentration in nanograms per milliliter).

2.1. Categorical variables

Suppose two methods are employed to assess the pregnancy status of cows after artificial insemination, and it is of interest to evaluate how well they agree, (e.g., Silva et al. [2]). One method is regarded as the gold-standard test and it is hoped that the other test, which is quicker, cheaper, or otherwise more efficient, may replace the gold-standard test. Cohen's kappa is commonly used to provide a measure of agreement in these circumstances. The results are presented in a two-way contingency table of frequencies with the rows and columns indicating the categories of response for each method (see Table 2).

2.1.1. Binary outcome

This section relates to data that come from studies where the response is binary (e.g., positive/negative, diseased/disease-free, above/below a threshold level).

2.1.1.1. Is there a systematic effect? The first question to be answered is whether there is a systematic difference between the results obtained from each method. For a binary response, this may be assessed by performing McNemar's test, a modification of the ordinary chi-square test that takes the paired nature of

Table 2
Contingency table of frequencies for bovine pregnancy determination using assessment of pregnancy associated glycoprotein at 27 d after timed AI (ELISA test) and transrectal ultrasound (gold-standard test).

		Gold-standard test		Total
		Pregnant	Nonpregnant	
ELISA test	Pregnant	596	61	657
	Nonpregnant	29	987	1016
	Total	625	1048	1673

Data from Silva E, Sterry RA, Kolb D, Mathialagan N, McGrath MF, Ballam JM, Fricke PM. Accuracy of a pregnancy-associated glycoprotein ELISA to determine pregnancy status of lactating dairy cows twenty-seven days after timed artificial insemination J Dairy Sci 2007;90:4612–4622.

the responses into account. A statistically significant result (generally if $P < 0.05$) shows that there is evidence of a *systematic difference* between the proportion of ''positive'' responses from the two methods (see, e.g., [3]). If one method provides the ''true values'' (i.e., it is regarded as the gold-standard method), the absence of a systematic difference implies that there is no *bias*. However, a non-significant result indicates only that there is no evidence of a systematic effect. A systematic effect may yet exist, but the power of the test may be inadequate to determine it. The *power* of a test is the ability of a test to detect as statistically significant a real difference, and is influenced by a number of factors. For binary data, when it is of interest to compare two proportions, these factors are the sample size (the power is smaller with a smaller sample size), the significance level (the cutoff for the P value such that any values of P below it indicate statistical significance), and the minimum difference in the proportions that the investigators believe represents an important difference. When the data are numerical, power is also influenced by the variation in the data. Power is generally expressed in percentage terms so, for example, an 80% power implies that there is an 80% chance of detecting as statistically significant a specified difference of a given magnitude between two proportions or two means.

2.1.1.2. Cohen's kappa. When a contingency table of the results of two methods is drawn up (Table 2), the frequencies of the agreement between the two methods are shown along the diagonal of the table. The corresponding frequencies *expected* if the categorizations were made randomly can be calculated; each is the relevant row total multiplied by the relevant column total, and this product is divided by the overall total. Expected frequencies are components of the chi-squared test statistic, which investigates a statistical hypothesis that there is no association between two factors. However, when two methods are being compared because they are believed to produce similar results, this chi-squared test is not relevant. We are interested, here, in *the degree of agreement*. This may be measured by Cohen's kappa (κ), which is given by:

$$\kappa = \frac{\text{Observed agreement} - \text{Chance agreement}}{\text{Maximum agreement} - \text{Chance agreement}}$$

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

It represents the chance-corrected proportional agreement, where:

- n = total observed frequency (e.g., total number of subjects, = 1673 in Table 2)
- O_D = sum of observed frequencies *along the diagonal*
- E_D = sum of expected frequencies along the diagonal, and
- $p_0 = O_D/n$
- $p_E = E_D/n$
- 1 in the denominator represents maximum agreement.

Perfect agreement is evident when Cohen's kappa equals 1; a value of Cohen's kappa equal to zero suggests that the agreement is no better than that which would be obtained by chance alone. Although there is no formal scale, the following levels of agreement are often considered appropriate for judging the extent of the agreement [4]. Agreement is

- Poor if $\kappa < 0.00$
- Slight if $0.00 \leq \kappa \leq 0.20$
- Fair if $0.21 \leq \kappa \leq 0.40$
- Moderate if $0.41 \leq \kappa \leq 0.60$
- Substantial if $0.61 \leq \kappa \leq 0.80$
- Almost perfect if $\kappa > 0.80$.

The approximate standard error of kappa is given by:

$$SE(\kappa) = \sqrt{\frac{p_0(1 - p_0)}{n(1 - p_E)^2}}$$

and the 95% confidence interval for the population value of kappa may be estimated by $\kappa \pm 1.96$ SE(κ).

As an example, Silva et al. [2] compared an early pregnancy enzyme-linked immunosorbent assay (ELISA) test for pregnancy associated glcoprotein (PAG) on blood samples collected from lactating diary cows at Day 27 after timed AI with transrectal ultrasound (TU) diagnosis of pregnancy at the same stage. In the case of disagreement between the two results, the TU was repeated at 32 d, and this result was taken as definitive. This final TU outcome was considered the gold standard or reference result (corresponding as closely as possible to the true result). The results on 1673 cows are shown in Table 2.

The estimated proportion pregnant by TU = 625/1673 = 0.374 and that by the ELISA test = 657/1673 = 0.393. A McNemar's test comparing the proportions pregnant by the two methods gives a chi-square test statistic = 10.7 on one degree of freedom, $P = 0.001$. When the methods suggested different pregnancy outcomes, PAG overestimates pregnancy compared with TU, indicating biased results in these circumstances. Nevertheless, the value of kappa = 0.886 (95% CI 0.863 to 0.909) suggests

that there was almost perfect agreement between the 1673 pairs of results, after taking chance agreement into account.

It should be noted that kappa is dependent on the number of categories of response in that its value is generally greater if there are fewer categories; kappa tends to be relatively high when there are only two categories, as in this example. Kappa is also dependent on the prevalence of the condition. We should thus be careful when comparing kappa values from different studies when the prevalences vary.

Other recent examples of studies illustrating kappa analysis for binary variables are Waldner et al. [5], Mainar-Haime and Barberán [3], and Ambrose et al. [6].

2.1.1.3. Validity. ***(A) Sensitivity and specificity, positive and negative predictive values.*** Suppose, as in the previous example, the response of interest is dichotomous or binary (i.e., falls into one of two categories), and one of the two methods that have been assessed for reproducibility is the gold-standard test for a particular state or condition. If Cohen's kappa shows that reproducibility is acceptable, the novel test is a reasonable alternative for experimental or diagnostic purposes. To properly evaluate the novel test, we should also assess the validity of the test to discriminate between the two outcomes. To this end, we evaluate some additional indices; the sensitivity, specificity, and positive and negative predictive values of the test.

Table 3 shows the observed frequencies in general terms. Then for the novel test:

Sensitivity = proportion of subjects with the condition who are correctly identified by the test = $a/(a + c)$.
Specificity = proportion of subjects without the condition who are correctly identified by the test = $d/(b + d)$.
Positive predictive value (PPV) = proportion of subjects with a positive test result who have the condition = $a/(a + b)$.
Negative predictive value (NPV) = proportion of subjects with a negative test result who do not have the condition = $d/(c + d)$.
Prevalence = proportion of subjects who have the condition = $(a + c)/n$.

Table 3
Contingency table showing the observed frequencies when the gold-standard test is compared with an alternative test.

		Gold-standard test		Total
		+	−	
Alternative test	+	a	b	$a + b$
	−	c	d	$c+ d$
Total		$a + c$	$b + d$	$n = a + b + c + d$

(+): Test is positive for the condition; (−) test is negative for the condition.

The formulae provide estimates of the true proportions in the population and, as such, confidence intervals should be calculated for these measures to provide an indication of the precision of the estimates. In each case, the 95% confidence interval for the relevant proportion, π, estimated by p, is approximated by $p \pm 1.96\sqrt{(p[1 - p]/n)}$. Usually, the measures are multiplied by 100 and expressed as percentages.

Using the data in Table 2 obtained from Silva et al. [2], the following estimates (with 95% CI) for the PAG test are obtained:

Sensitivity = 596/625 = 0.954 or 95.4% (95% CI, 93.7% to 97.0%)
Specificity = 987/1048 = 0.942 or 94.2% (95% CI, 92.8% to 95.6%)
PPV = 596/657 = 0.907 or 90.7% (95% CI, 88.5% to 92.9%)
NPV = 987/1016 = 0.971 or 97.1% (95% CI, 96.1% to 98.2%)
Prevalence in the population = 625/1673 = 0.374 or 37.4% (95% CI, 35.5% to 39.3%).

The sensitivity and specificity are properties of the test. A perfect test has sensitivity = specificity = 100%. However, in practice, sensitivity is gained at the expense of specificity, and vice versa. The choice of test (i.e., one that tends toward a high sensitivity or high specificity) depends on what condition we are anxious to detect, together with the importance of either a false-positive or false-negative test result.

Knowledge of the sensitivity and specificity of a particular test, however, does not help the investigator decide on how likely it is that a particular individual has or does not have the condition of interest, once that individual's test result is known. This information is provided by the predictive values. It should be noted that predictive values are dependent on the prevalence of the condition in the population being studied. In situations where the condition is common, the positive predictive value will be much higher than in populations where the condition is infrequent. Conversely, the negative predictive value will be lower. From the early pregnancy ELISA test results, Silva et al. [2] concluded that with a negative predictive value of 97%, few cows would be needlessly aborted if a resynchronization

protocol using prostaglandin $F_{2\alpha}$ were adopted following the test.

We have shown how to estimate the predictive values of a test using the information gleaned from the contingency table. Generally, these data are not available to the investigator faced with diagnosing a particular individual. A simple approach in this situation is to use *Fagan's nomogram*, Fig. 1. (An interactive version of this nomogram can be found on the Web site of the Centre for Evidence Based Medicine, http://www.cebm.net/index.aspx?o=1161, but for our example we use Fig. 1.) The likelihood ratio of a particular test result must first be determined in order to proceed. The *likelihood ratio for a positive test result* (LR_+) is the ratio of the chance of a positive result if the individual has the condition to that if the individual does not have the condition. It is equal to the sensitivity divided by (1 – specificity). If the investigator has some idea of how likely it is that the individual has the condition before the test result is available (commonly this pre-test probability is taken as the prevalence of the condition in the population), then all she or he has to do is connect this pre-test probability in Fagan's nomogram to the likelihood ratio of the test and extend the line to where it meets the right-hand axis. The cut-point of this axis provides an estimate of the posttest probability, the chance that the individual has the condition if that individual has a positive test result.

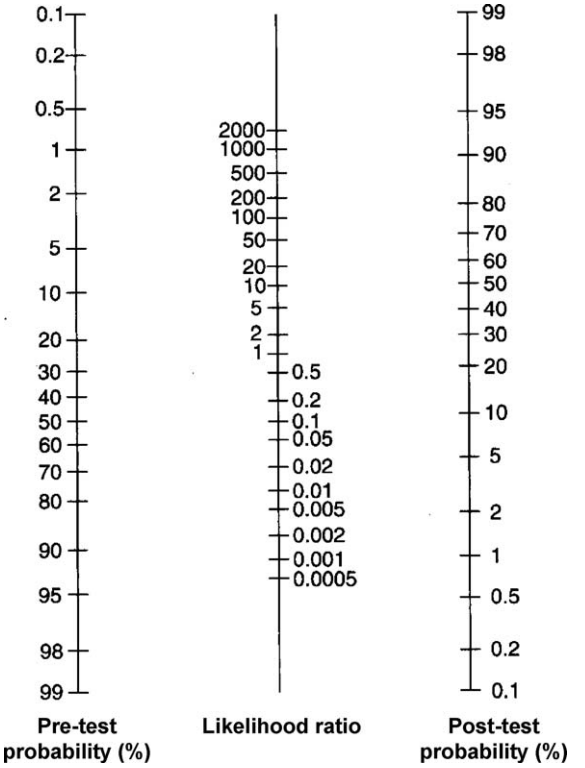

Fig. 1. Fagan's nomogram. (Adapted from Sackett Dl, Richardson WS, Rosenberg W, Haynes RB. Evidence-based Medicine: How to Practice and Teach EBM, 2nd Edn., Churchill-Livingstone, 1977, with permission.)

The process of determining the post-test probability of the condition using Fagan's nomogram is based on a *Bayesian* approach to statistics. This relies on specifying the probability of a particular outcome *before* the study has been conducted; consequently, it is called the prior probability. This, of course, assumes that there is some background experience of the condition on which to base a prior probability estimate. As mentioned previously, this is commonly taken as the prevalence of the condition in the population. Then, in a Bayesian analysis, the results from the study are used to update (improve) this prior probability to provide what is known as the posterior probability of the outcome. In the context of a diagnostic test, the pre-test probability of the condition is the prior probability, the likelihood ratio contains the relevant information from the sample data, and the post-test probability is the posterior probability.

For the data in Table 2:

LR_+ = sensitivity/(1 – specificity) = 0.954/(1 – 0.942) = 16.4

Prevalence = pre-test probability = 37.4%.

From Fagan's nomogram, the post-test probability (probability that a positive test is correct) is approximately 91%. (It can be shown that the 95% confidence interval is from 89% to 93%.) Thus if a cow tests positive for pregnancy using the ELISA test, she has a 91% chance of actually being pregnant. Note that this is the same value as was obtained for the PPV from the contingency table. Furthermore, the likelihood ratio for a negative test result, LR_- = (1 – sensitivity)/specificity = (1 – 0.954)/0.942 = 0.049, and use of Fagan's nomogram indicates that the post-test probability of a cow being pregnant after testing negative for pregnancy using the ELISA test is approximately 3% (95% CI, 2% to 4%). So the chance of a cow not being pregnant after a negative ELISA test is approximately 97%, which is equal to the estimated NPV from the contingency table.

(B) The receiver operating characteristic (ROC) curve. Sometimes we rely on a numerical or ordinal measurement rather than a binary outcome (e.g., positive or negative) to diagnose a condition. In such cases, there is often no simple cutoff above or below which the condition is present (i.e., there is no threshold for the condition). So we need to set a cutoff value for

the measurement that provides the greatest chance of detecting the condition. We can set the cutoff as the upper or lower limit of the reference interval, and the sensitivity, specificity, and predictive values can be calculated for this threshold. By raising or lowering the cutoff value, we can calculate a series of sensitivities, specificities, and predictive values and choose that cutoff which produces the optimal set.

One approach to determining the optimal cutoff for a diagnostic test is to draw the receiver operating characteristic (ROC) curve. The ROC curve is obtained by plotting the sensitivity (i.e., the probability of a true positive) against (1 – specificity) [i.e., the probability of a false positive] for each cutoff, and connecting the points so obtained by lines. The resulting curve then relates to a comparison of the probabilities of a positive test result in those with and without the condition. The diagonal, representing the 45-degree line through the origin, indicates that the test is no better than chance at discriminating between subjects with and without the condition. The ROC curve for a useful test will lie to the left of the diagonal of the graph. Generally, the best cutoff value for discriminating between subjects with and without the condition corresponds with that point on the curve which is nearest the top left-hand corner of the graph. However, there may be circumstances where the importance of either false positives or false negatives is overriding, and thus a different cutoff value may be chosen.

The area under the ROC curve (sometimes called the AUROC) can be used to compare the overall accuracy of different tests for the same condition. It can be calculated manually or is given by the c statistic, the probability that a randomly chosen subject from the group with the condition has a higher predicted probability of testing positive than a randomly chosen subject from the group without the condition. The test giving the higher c statistic has the better chance of discriminating between the two possible outcomes. When $c = 1$, the test is perfectly accurate, and $c = 0.5$ indicates the test is no better than chance alone at discriminating between the two outcomes.

The use of an ROC curve was demonstrated by Martinez-Pastor et al. [7]. These authors considered two tests for sperm DNA fragmentation, the sperm chromatin dispersion (SCD) test and the sperm chromatin structure assay (SCSA) before and after an oxidative stress for 6 h. SCSA was then expressed as the percentage of sperm with damaged chromatin (%DFI). The SCD test failed to distinguish between the control and oxidized samples (Fig. 2, lower ROC curve with blue circles), whereas %DFI was strongly discriminating between the control and oxidized samples (Fig. 2, higher ROC curve with red circles). Note that the AUROC values indicated the discriminating power of the test.

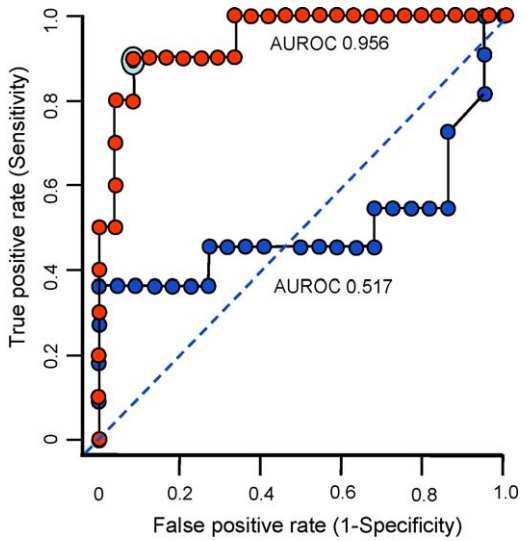

Fig. 2. Receive operating characteristic curves illustrating a poorly discriminating test (blue circles, AUROC 0.517) and a test that has good discriminating power (red circles, AUROC 0.956; green circle indicates the point of greatest discrimination). (Redrawn from Martínez-Pastor F, del Rocío Fernández-Santos M, Domínguez-Rebolledo ÁE, Esteso MC, Garde JJ. DNA status on thawed semen from fighting bull: a comparison between the SCD and the SCSA tests. Reprod Domest Anim 2008;44:424–431.)

2.1.2. *Greater than two ordered categories: Weighted kappa*

For ordinal data (i.e., when there are three or more categories of response and they are ordered), it is helpful to provide a measure that gives consideration not only to the agreement between the pairs of results but also to the extent to which there is disagreement between them. Clearly, if the two methods differ in their responses for a particular subject by two categories, there is greater disagreement than if the methods differ by only one category. To take the extent to which there is disagreement into account, we can calculate a weighted kappa [8], which is a modification of the kappa described in Section 2.1.1.2. We assign weights to the frequencies in the nondiagonal cells of the contingency table according to their distance from the diagonal, with the magnitude of the weight diminishing the further the cell is from the diagonal. The weighted kappa value is generally calculated automatically using specialist statistical software.

As an example, Oskram et al. [9] in a study of testicular maturation in male pigs used a standard

Table 4
Contingency table of frequencies showing the comparison of histologic assessment and DNA flow cytometry for the evaluation of testicular tissue in entire male pigs.

		Histology			Total
		Immature	Transitional	Mature	
DNA flow cytometry	Immature	6	2	0	8
	Transitional	4	17	3	24
	Mature	0	4	19	23
	Total	10	23	22	55

Source: Oskam IC, Ropstad E, Andersen Berg K, Fredriksen B, Larsen S, Dahl E, Andresen Ø. Testicular germ cell development in relation to 5a-androstenone levels in pubertal entire male pigs. Theriogenology 2008;69:967–976. (Table reproduced with permission).

histologic classification of testicular histology to compare with a flow cytometric classification of cellular quantity of nuclear DNA. Their results are reproduced in Table 4.

Number of observed agreements: 42 (76.4% of the observations)
Number of agreements expected by chance: 20.7 (37.6% of the observations)
Weighted $\kappa = 0.688$ (95% CI, 0.536 to 0.840), which is slightly greater than the unweighted kappa = 0.621 (95% CI, 0.441 to 0.801).

The weighted kappa generally gives a better indication of the agreement but can only be used with data that are ranked on an ordinal scale and contain at least three categories. It is very similar to the *intraclass correlation coefficient,* which may be used when the variable of interest is numerical (see Section 2.2.3.3).

2.2. Numerical variables

The kappa statistic is an *inappropriate* measure of the agreement between pairs of readings when the variable of interest is numerical (e.g., serum hormone concentration in nanograms per milliliter). Again, the correct approach to be adopted in these circumstances can be used both to evaluate repeatability and reproducibility. For example, we might want to assess the reproducibility of two ways of measuring a numerical outcome variable by comparing their results when a measurement is made by each method on n subjects. The example we use to illustrate the techniques is one of repeatability: it uses data that compare the follicular diameter before ovulation in two consecutive spontaneous cycles in 20 mares (full results given in [10]). The mean (and SD) follicular diameter of the 20 mares in Cycles 1 and 2, respectively, were 46.03 mm (6.36 mm) and 46.33 mm (6.01 mm).

2.2.1. Is there a systematic effect?

To determine whether there is a systematic difference between the two methods in a reproducibility study or duplicate observations in a repeatability study, we calculate the difference between each of the n pairs of measurements. We can generally use a paired t-test to test the null hypothesis that the true mean difference is zero, although if the differences between the pairs do not approximate a Normal distribution, we should use a non-parametric test such as the Wilcoxon signed ranks test or the sign test. (Most introductory statistical texts have some information on non-parametric tests, but a dedicated text is that by Siegel and Castellan [11]). If the mean of these differences is zero, then it may be concluded that there is no systematic difference between the pairs of results (i.e., *on average*, the results are reproducible or repeatable, as relevant). A significant result suggests that there is a systematic difference, but a non-significant result indicates only that there is no evidence of a systematic effect. As with a categorical variable, if one method in a reproducibility study is regarded as the gold standard, the presence of a systematic difference implies that there is bias. Using the pairs of values of follicular diameter (mm) from 20 mares, we find that we obtain an estimated mean difference (Cycle 2 – Cycle 1) of 0.30 mm (95% CI, 1.09 mm to 1.69 mm), with the differences being approximately Normally distributed. The paired t-test statistic is 0.45 on 19 degrees of freedom, giving P = 0.66. Hence, there is no evidence to reject the null hypothesis that the true mean difference is zero. This indicates that there is no evidence of a systematic difference between the follicular diameter measurements in the two cycles.

2.2.2. The Bland and Altman diagram

A display of the differences between the pairs of readings may offer an insight into the pattern (and extent) of the agreement. The *Bland and Altman diagram* [12] is such a display; the difference between a pair is plotted on the vertical axis of the diagram against the mean of the pair on the horizontal axis. Fig. 3 shows the Bland and Altman plot of the follicular diameter data obtained from 20 mares in two repeated cycles. If a random scatter of points is observed, a single measure of repeatability is acceptable. To determine such a measure, we first estimate the standard deviation of the differences (s_d). Assuming a Normal distribution of differences, approximately 95% of the differences in the

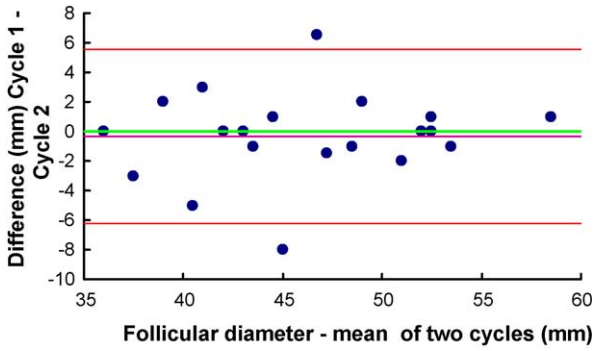

Fig. 3. Bland and Altman diagram showing the plot of the difference between the diameters (mm) of the equine follicle just prior to ovulation in two consecutive cycles of the mare against the mean of the pair (n = 20). Red lines show limits of agreement, and the purple line shows the mean value of the differences. The green line is the zero line used to assess the discrepancy of the observed mean difference from zero. (Data from Ref. 10, courtesy of Dr. Cuervo-Arango.)

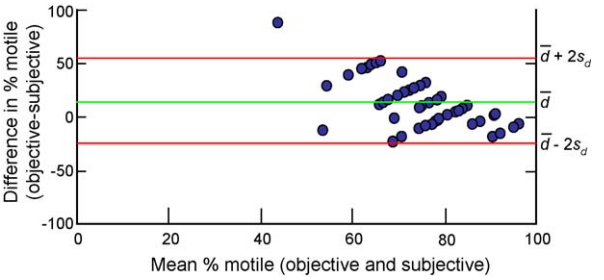

Fig. 4. Bland and Altman plot showing limits of agreement between two methods of measuring sperm motility, an objective Hamilton-Thorne computer-based semen analyzer and a subjective visual assessment, of samples of boar semen. $\bar{d}$ is the mean difference, and s_d is the standard deviation of the differences between pairs of measurements. (Redrawn and modified from Vyt P,Maes D, Rijsselaere T, Dejonkheere E, Castryck F, Van Soom A. Motility assessment of porcine spermatozoa: a comparison of methods. Reprod Dom Anim 2004;39:447.)

population are expected to lie between $\bar{d} \pm 2s_d$, where $\bar{d}$ is the mean of the observed differences. The upper and lower limits of this interval, usually displayed on the Bland and Altman diagram, provide the *limits of agreement*; from them, we can decide (subjectively) whether the agreement between pairs of readings in a given situation is acceptable (see Fig. 3). For the mare data, the standard deviation of the differences is estimated as 2.97 mm and the 95% limits of agreement by –6.12 mm and 5.52 mm. The limits of agreement are shown as red lines in Fig. 3. The purple line is the line corresponding with the mean difference of –0.30 mm (it is negative in the diagram, indicating that on average the diameter measurements from the second cycle are greater than those of the first cycle).

Furthermore, the *British Standards Institution repeatability/reproducibility coefficient* ($2\ s_d$) may be used as a single measure of agreement. It indicates the maximum likely difference between a pair of readings. The British Standards repeatability coefficient for the mare data is $2 \times 2.97 = 5.94$ mm, which the investigators found represented acceptable repeatability.

It should be noted that if the extent of agreement between the pairs depends on the magnitude of the measurement, a single measure of agreement is inappropriate. This would be evident on inspecting the Bland and Altman diagram if a funnel effect were observed. In such a situation, the variation in the differences is larger (say) for smaller mean values and decreases as the mean values become larger.

No funnel effect is observed in Fig. 3, but an example of its occurrence is shown in Fig. 4 (e.g., Vyt et al. [13]). These authors compared boar semen motility scores using a Hamilton-Thorne computer-based semen analyzer (HTR) with subjective microscope scoring from two experienced individuals. Fig. 4 shows the Bland and Altman diagram comparing the HTR with results from the first of the two individuals, in which the differences get smaller with the higher percentages. Note also that the mean difference departs substantially from zero indicating that the automated system gives systematically higher values for percentage motility.

In this situation, where a funnel effect is observed, the problem must be reassessed. An appropriate transformation of the raw data may resolve the issue, so that when the process is repeated on the transformed observations, the required conditions are satisfied. Otherwise, we should not calculate a *single* measure of reproducibility.

The Bland and Altman diagram can also be used to detect outliers. Outliers are occasional extreme readings departing from the main body of the data, possibly caused by errors of measurement.

2.2.3. Indices of reliability

There are a number of different indices of reliability that may be calculated for numerical data, all giving comparable results. It is important that values of a particular index of reliability are not compared using different data sets as the indices are influenced by the character of the data, such as its variability (tending to increase as the observations become more variable). Note that both of the indices recommended in this article, Lin's concordance correlation coefficient and the ICC, are independent of the actual scale of measurement and of the size of error that is considered experimentally or clinically acceptable.

2.2.3.1. Inappropriate use of the Pearson correlation coefficient. Paired observations from two different occasions or from two different observers/methods are often inappropriately evaluated for agreement using the Pearson correlation coefficient between the pairs (e.g., [10,13–16]). This is an *incorrect* measure of reproducibility or repeatability. Whether the data fall on a straight line in a scatter diagram, when the observation from one member of a pair is plotted against that from the other, is not in question; it would be entirely unsurprising if the data from two methods, two observers, or duplicate readings were related, given that this is what we are hoping to verify. This is shown in Fig. 5 for two different situations for the comparison of two methods of measurement. In one case, all the points lie on a straight line that does not pass through the origin, so there is strong correlation with $r = 1$ but no agreement between the pairs of data (in this case, there is a clear systematic effect with one member of the pair [Method 1] always having a greater response than the other [Method 2]). In the other situation, there is considerable scatter around the best fitting line and a poor correlation ($r = 0.5$), but there is no evidence of a systematic effect (bias) so that, on average, the methods agree. Neither of these outcomes helps in assessing the agreement between the two data sets. What we need to establish is whether the paired data conform to a line of equality (i.e., the 45-degree line through the origin when the two scales are the same). This will not be established by testing the null hypothesis that the true Pearson correlation coefficient is zero.

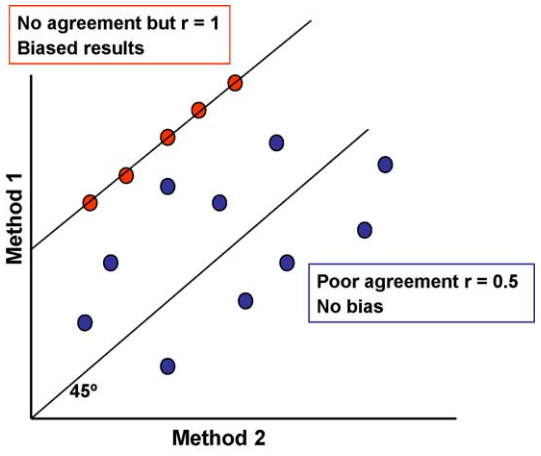

Fig. 5. Diagram showing two kinds of association between the results of Method 1 and those of Method 2. The red circles on the upper line demonstrate perfect correlation but no agreement. The blue circles around the lower line demonstrate poor correlation but no systematic difference between the two methods.

2.2.3.2. Lin's concordance correlation coefficient. Lin's concordance correlation coefficient [17] may be calculated as an index of reliability. An understanding of Lin's concordance correlation coefficient is obtained if the line of best fit to the data comparing two methods is shown in a scatterplot when the results from one method are plotted against the other. The Pearson correlation coefficient provides a measure that describes the extent to which the points in the scatter diagram conform to the best fitting line. Lin's coefficient modifies the Pearson correlation coefficient by assessing not only how close the data are about the line of best fit but also how far that line is from the 45-degree line through the origin, this 45-degree line representing perfect agreement. Lin's coefficient is 1 when all the points lie exactly on the 45-degree line drawn through the origin and diminishes as the points depart from this line and as the line of best fit departs from the 45-degree line.

Fig. 6 shows the follicular diameter data in 20 mares when the results from Cycle 1 are plotted against those of Cycle 2. The estimated regression line drawn through the midst of the points has a slope of 1.06 mm per mm (which is close to the slope of 1 mm per mm for the 45-degree line through the origin that would be obtained if there were perfect agreement) and a value of $r^2 = 0.785$ (this is a measure of goodness of fit about the line, indicating that just under 80% of the variation in one variable can be explained by its linear relationship with the other).

Lin's coefficient can be calculated as:

$$r_c = \frac{2rs_x s_y}{s_x^2 + s_y^2 + (\bar{x} - \bar{y})^2}$$

where r_c is the estimated Pearson correlation coefficient between the n pairs of results (x_i, y_i), and $\bar{x}$ and $\bar{y}$ are the

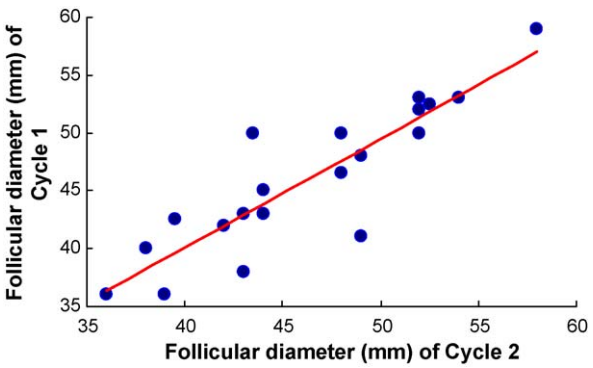

Fig. 6. A scatterplot of the diameter of the ovulating follicle just prior to ovulation from two consecutive cycles in 20 mares [10]. The line of best fit is drawn though the points.

sample means of x and y, respectively

$$s_x^2 = \frac{\sum(x_i - \bar{x})^2}{n} = \frac{n-1}{n} \text{ times the estimated variance of } x$$

$$s_y^2 = \frac{\sum(y_i - \bar{y})^2}{n} = \frac{n-1}{n} \text{ times the estimated variance of } y.$$

Using this formula for the follicular diameter data in 20 mares, where $r = 0.886$ and the estimated variances of the follicular diameter in the 20 mares in the first and second cycles, respectively, are 40.51 mm^2 and 36.09 mm^2, gives an estimated value for Lin's coefficient of 0.883. This value is close to the maximum value of 1, indicating that there is good repeatability between the two sets of results. It can be shown that the 95% confidence interval for Lin's correlation coefficient is (0.78 to 0.98).

Studies using Lin's concordance analysis can be found in Quist et al. [18]. Barlund et al. [19] illustrate the use of kappa, sensitivity and specificity analysis, and Lin's concordance analysis in a comprehensive study of five different methods (two with different cutoffs) to diagnose endometritis in cattle.

2.2.3.3. The intraclass correlation coefficient. The *intraclass correlation coefficient (ICC)* is another index of reliability that may be calculated to measure reproducibility and repeatability; it is almost identical to Lin's concordance correlation coefficient. The ICC takes a value from zero (implying no agreement) to 1 (perfect agreement). When measuring the agreement between pairs of observations, it represents the between-pair variance expressed as a proportion of the total variance of the observations (i.e., it is the proportion of the total variability in the observations that is due to the *differences between pairs*).

Providing there is no evidence of a systematic difference between the pairs, we may calculate the ICC as the Pearson correlation coefficient between the $2n$ pairs of observations obtained by including each pair twice, once when its values are as observed and once when they are interchanged. The estimated value of the Pearson correlation coefficient from the 40 pairs of follicular diameter values obtained in this way from the data we introduced in Section 2.2 and displayed in Fig. 6 [10] is 0.884 with 95% confidence interval (0.78 to 0.94). This value of the estimated ICC is almost identical to Lin's concordance correlation coefficient, which was estimated as 0.883.

If a systematic difference between the observations in a pair is to be taken into account, the ICC is calculated as:

$$\frac{s_a^2 - s_d^2}{s_a^2 + s_d^2 + \frac{2}{n}(n\bar{d}^2 - s_d^2)}$$

where the difference between and the sum of the observations in each of the n pairs is determined and

s_a^2 is the estimated variance of the n sums;
s_d^2 is the estimated variance of the n differences;
$\bar{d}$ is the estimated mean of the differences (an estimate of the systematic difference).

Using the follicular diameter data from two cycles in 20 mares, we find that $s_a^2 = 144.37\ mm^2$, $s_d^2 = 8.83\ mm^2$, and $\bar{d} = 0.30$ m. Hence, using the formula that takes the systematic effect into account, we obtain a virtually identical estimated ICC of 0.889. Examples of use of the ICC can be found in Waldner et al. [5].

2.3. More complex situations with categorical and numerical variables

Sometimes more complex problems when assessing agreement may arise. For example, there may be more than two replicates, or more than two observers, or each of a number of observers may have replicate observations. Details of the analysis of such problems may be found in Streiner and Norman [20]. Some other authors who deal with these more complex analyses are Dunn [21], Blackman [22], Shrouki [23], Bannerjee et al. [24], de Vet [25], and Fleiss et al. [26]. An example of such complex analysis of a binary variable is seen in David et al. [27] who compared increasingly complex nested models of breeding components by maximum likelihood to predict fertility of sheep in the French AI service. An example of repeatability estimates using analysis of variance calculations with numerical variables is to be seen in a study of the ''cost'' of reproduction in Zebra finches in which hematologic variables were investigated as an indicator of reproductive cost to the bird [28].

3. Sample size estimation for reliability calculations

There are a number of different approaches to estimating the optimal sample size for a calculation of a measure of agreement such as kappa or the intraclass

correlation coefficient (e.g., [1,29–31]). Some of these approaches are concerned with estimating the sample size when it is of interest to test the significance of the measure of agreement, and relevant tables for ease of use are available (e.g., [30]). However, as the significance of the measure from zero (the most common hypothesis test) or some other value is generally not an issue in an agreement study, we prefer the two approaches detailed in the following sections. Both rely on specifying the maximum acceptable width of the confidence interval for the measure of agreement. It can be shown [30] that for reliability values of 0.40 or greater, two or three observations per subject will minimize the total number of subjects required. For simplicity, the explanation of sample size calculations is therefore restricted to determining the sample size for a reliability study with pairs of measurements; for example, a reproducibility study comparing two methods of measurement or a repeatability study comparing duplicate measurements on each subject by one observer. For both calculations, if a different confidence interval is required (e.g., a 99% confidence interval), the 1.96 in the formulae provided in Sections 3.1 and 3.2 is replaced by the relevant percentage point of the Normal distribution (e.g., 2.58 for a 99% confidence interval). Sample size determination may be simplified by the use of tables (e.g., [32]) or appropriate statistical software.

3.1. Kappa for a binary outcome

When the outcome variable is binary (e.g., positive/negative), it can be shown [33] that if W is the maximum acceptable width of kappa's 95% confidence interval, π is the underlying true proportion of positives, and κ is the anticipated value of kappa, the optimal sample size (e.g., the number of pairs of measurements) is

$$4\frac{(1-\kappa)}{W^2}\left((1-\kappa)(1-2\kappa)+\frac{\kappa(2-\kappa)}{2\pi(1-\pi)}\right)1.96^2$$

Using the example of Silva et al. [2] described in Section 2.1.1.2, let us assume that they wanted to estimate their sample size to give a kappa = 0.8 with a confidence interval width of, say, 0.2, and that they believed that approximately 40% of dairy cows would become pregnant (i.e., the estimated true proportion of positives is 0.4). Substituting these values into the formula suggests that 123 diary cows should be used in the study. Clearly their actual trial size of 1673 far exceeded these expectations!

3.2. Intraclass correlation coefficient for a numerical outcome

If W_ρ is the acceptable width of the 95% confidence interval for the ICC for a numerical variable and ρ_1 is the anticipated value of the ICC, then the optimal number of pairs of measurements for the study [29] is

$$1+\frac{8(1.96)^2(1-\rho_1)^2(1+\rho_1)^2}{2W_\rho^2}.$$

For the study of Cuervo-Arango and Newcombe [10] on the measurement of follicular diameter in mares in two consecutive cycles (see Section 2.2), on an assumption of an anticipated ICC = 0.8 with an acceptable confidence interval width of, say, 0.25, we arrive at an optimal sample size of 33 mares. This exceeds the sample size of 20 mares that the authors actually used in the study.

4. Conclusion

In general, there has been a noteworthy improvement in standards of statistical data analysis in the past few years, perhaps coinciding with the ready availability of computer packages. Unfortunately, this improvement has not been evident in all areas of statistical methodology; in particular, the procedures to assess reliability and measure agreement are often overlooked or else time-warped, with researchers relying on inappropriate methods found in previously published material. To combat such failings, we have concentrated in this review article on relatively simple approaches to investigating reliability. We have outlined the appropriate techniques to ascertain the reliability of paired categorical or paired numerical data sets when assessing reproducibility or repeatability. We have provided worked examples to illustrate these techniques, and we have also offered references to studies that used these methods. More complex analyses are best dealt with under guidance, and we recommend that when the complexity exceeds the approaches we have covered here, professional statistical advice should be sought early in the planning of the study.

Acknowledgment

We thank Dr. J. Cuervo-Arango for making his data available to us.

References

[1] Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. Statist Med 1998;17:101–11.

[2] Silva E, Sterry RA, Kolb D, Mathialagan N, McGrath MF, Ballam JM, Fricke PM. Accuracy of a pregnancy-associated glycoprotein ELISA to determine pregnancy status of lactating dairy cows twenty-seven days after timed artificial insemination. J Dairy Sci 2007;90:4612–22.
[3] Mainar-Jaime RC, Barberán M. Evaluation of the diagnostic accuracy of the modified agglutination test (MAT) and an indirect ELISA for the detection of serum antibodies against *Toxoplasma gondii* in sheep through Bayesian approaches. Vet Parasitol 2007;148:122–9.
[4] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.
[5] Waldner CL, Cunningham G, Campbell JR. Agreement between three serological tests for *Neospora caninum* in beef cattle. J Vet Diagn Invest 2004;16:313–5.
[6] Ambrose DJ, Radke B, Pitney PA, Goonewardene LA. Evaluation of early conception factor lateral flow test to determine nonpregnancy in dairy cattle. Can Vet J 2007;48:831–5.
[7] Martínez-Pastor F, del Rocío Fernández-Santos M, Domínguez-Rebolledo ÁE, Esteso MC, Garde JJ. DNA status on thawed semen from fighting bull: a comparison between the SCD and the SCSA tests. Reprod Domest Anim 2008;44:424–31.
[8] Cohen J. Weighted kappa: nominal scale agreement with provision for scale disagreement or partial credit. Psychol Bull 1968;70:213–20.
[9] Oskam IC, Ropstad E, Andersen Berg K, Fredriksen B, Larsen S, Dahl E, Andresen Ø. Testicular germ cell development in relation to 5a-androstenone levels in pubertal entire male pigs. Theriogenology 2008;69:967–76.
[10] Cuervo-Arango J, Newcombe JR. Repeatability of preovulatory follicular diameter and uterine edema pattern in two consecutive cycles in the mare and how they are influenced by ovulation inductors. Theriogenology 2008;69:681–7.
[11] Siegel S, Castellan NJ. Nonparametric Statistics for the Behavioural Sciences, 2nd edition, McGraw-Hill; 1988.
[12] Bland JM, Altman DG. Statistical methods for assessing agreement between two pairs of clinical measurement. Lancet 1986;i:307–10.
[13] Vyt P, Maes D, Rijsselaere T, Dejonckheere E, Castryck F, Van Soom A. Motility assessment of porcine spermatozoa: a comparison of methods. Reprod Domest Anim 2004;39:447–53.
[14] Nagy P, Solti L, Kulcsár M, Reiczigel J, Huszenicza G, Abaváry K, Wölfling A. Progesterone determination in equine plasma using different immunoassays. Acta Vet Hung 1998;46:501–13.
[15] Christensen P, Hansen C, Liboriussen T, Lehn-Jensen H. Implementation of flow cytometry for quality control in four Danish bull studs. Anim Reprod Sci 2005;85:201–8.
[16] Colazo MG, Ambrose DJ, Kastelic JP, Small JA. Comparison of two enzyme immunoassays and a radioimmunoassay for measurement of progesterone concentrations in bovine plasma, skim milk, and whole milk. Can J Vet Res 2008;72:32–6.
[17] Lin LI-K. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989;45:255–68.
[18] Quist MA, LeBlanc SJ, Hand KJ, Lazenby D, Miglior F, Kelton DF. Agreement of predicted 305-day milk yields relative to actual 305-day milk weight yields. J Dairy Sci 2007;90:4684–92.
[19] Barlund CS, Carruthers TD, Waldner CL, Palmer CW. A comparison of diagnostic techniques for postpartum endometritis in dairy cattle. Theriogenology 2008;69:714–23.
[20] Streiner DR, Norman GL. Health Measurement Scales: A Practical Guide to Their Development and Use, 3rd edition, Oxford University Press; 2003.
[21] Dunn G. Design and Analysis of Reliability Studies: The Statistical Evaluation of Measurement Errors. Oxford University Press; 1989.
[22] Blackman N. Reproducibility of clinical data II: categorical outcomes. Pharm Stat 2004;3:109–22.
[23] Shroukri MM. Measurement of agreement. In: Armitage P, Colton T, editors. Encyclopedia of Biostatistics. 2nd edition, Wiley; 2005. p. 123–37.
[24] Banerjee M, Capozzoli M, Mcsweeney L, Sinha D. Beyond kappa: a review of interrater agreement measures. Can J Stat 2007;27:3–23.
[25] de Vet H. Observer reliability and agreement. In: Armitage P, Colton T, editors. Encyclopedia of Biostatistics. 2nd edition, Wiley; 2005. p. 3801–5.
[26] Fleiss JL, Levin B, Paik MC. Statistical Methods for Rates & Proportions, 3rd edition, Wiley Interscience; 2003.
[27] David I, Bodin L, Lagriffoul G, Leymarie C, Manfredi E, Robert-Granié C. Genetic analysis of male and female fertility after artificial insemination in sheep: comparison of single-trait and joint models. J Dairy Sci 2007;90:3917–23.
[28] Wagner EC, Prevolsek JS, Wynne-Edwards KE, Williams TD. Hematological changes associated with egg production: estrogen dependence and repeatability. J Exp Biol 2008;211:400–8.
[29] Bonnett DG. Sample size requirements for estimating intraclass correlations with desired precision. Stat Med 2002;21:1331–5.
[30] Shoukri MM, Asyali MH, Donner A. Sample size requirements for the design of reliability study: review and new results. Stat Methods Med Res 2004;13:251–71.
[31] Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Physical Ther 2005;85:257–68.
[32] Machin D, Campbell M, Fayers P, Pinol A. Sample Size Tables for Clinical Studies, 2nd edition, Blackwell Science; 1997.
[33] Machin D, Campbell MJ. Design of Studies for Medical Research. Wiley; 2005.