Chapter 5

Reliability and Measurement Error

Introduction

A test should give the same or similar results when administered repeatedly to the same individual within a time too short for real biological variation to take place. Results should be consistent whether the test is repeated by the same observer or instrument or by different observers or instruments. This desirable characteristic of a test is called "reliability" or "reproducibility."

Measures of reliability quantify differences between distinct measurements of the same thing. These differences can be random, if there is no particular pattern to the disagreements, or systematic if the disagreements tend to occur in one direction. How reliability is quantified depends on whether the result of the measurement is expressed as a number or a category.

Of course, just because two measurements agree with each other does not mean they are both giving the right answer. In Chapters 2–4 we assumed that there was a "gold standard" that allowed us to determine accuracy – how often a test gave the right answer in different groups of patients. However, in some situations, there is no gold standard and we need to settle for reliability. Although reliability is no guarantee of accuracy, an unreliable test cannot be very accurate.

Types of Variables

How we assess reliability of a measurement depends on whether the scale of measurement is numeric, the number of possible values, and whether they are ordered. Dichotomous variables, like alive or dead, have only two possible values. Nominal variables like blood type, race, or cardiac rhythm can take on a limited number of separate values and have no inherent order. Ordinal variables, such as pain that is rated "none," "mild," "moderate," or "severe," have an inherent order. Many scores or scales used in medicine, such as the Glasgow Coma Score, are ordinal variables.

Numeric variables are continuous if they can take on an infinite number of values, such as weight, serum glucose, or peak expiratory flow. In contrast, numeric variables are discrete if they can take on only a finite number of values, like the number of previous pregnancies or heart attacks, or the number of decayed, missing or filled teeth. If discrete numeric variables take on many possible values, they behave like continuous variables; if there are just a few possible values we can treat them as ordinal variables. Either continuous or discrete numeric variables can be grouped to create ordinal or dichotomous variables.

In this chapter, we will learn about the kappa statistic for measuring intra- and inter-rater reliability of nominal measurements and about the weighted kappa statistic for ordinal measurements. Assessment of intra-rater or intra-method reliability of a continuous test requires measurement of either the within-subject standard deviation or the within-subject coefficient of variation (depending on whether the random error is proportional to the level of the measurement). A Bland–Altman plot [1] can help estimate both systematic bias and random error. While correlation coefficients are often used to assess intra- and inter-rater reliability of a continuous measurement, we will see that they are generally inappropriate for assessing random error and useless for assessing systematic error (bias). We conclude with a brief discussion of calibration in which a continuous measurement is compared to a reference standard that is also continuous.

Measuring Interobserver Agreement for Categorical Variables

Agreement

When there are two observers or when the same observer repeats a categorical measurement on two occasions, the agreement can be summarized in a k × k table, where k is the number of categories. The simplest measure of interobserver agreement is the concordance or observed agreement rate, that is, the proportion of observations on which the two observers agree. This can be obtained by summing the numbers along the diagonal of the k × k table from the upper left to the lower right and dividing by the total number of observations.

We start by looking at some simple 2 × 2 (yes or no) examples. Later in this chapter, we will look at examples with more categories.

Example 5.1 Suppose you wish to measure inter-radiologist agreement at classifying 200 X-rays as either "normal" or "abnormal." Because there are two possible values, you can put the results in a 2 × 2 table.

Classification of 200 X-rays by two radiologists

		Radiologist #2		
		Abnormal	**Normal**	**Total**
Radiologist #1	**Abnormal**	40	30	70
	Normal	20	110	130
	Total	**60**	**140**	**200**

In this example, out of 200 X-rays, there were 40 that both radiologists classified as abnormal (upper left) and 110 that both radiologists classified as normal (lower right), for an observed agreement rate of (40 + 110)/200 = 75%.

When the observations are not evenly distributed among the categories (e.g., when the proportion "abnormal" on a dichotomous test is substantially different from 50%), the observed agreement rate can be misleading.

Example 5.2 If two radiologists each rate only 5 of 200 X-rays as abnormal (2.5%), but do not agree at all on which ones are abnormal, their observed agreement will still be (0 + 190)/200 = 95%.

		Radiologist #2		
		Abnormal	**Normal**	**Total**
Radiologist #1	**Abnormal**	0	5	5
	Normal	5	190	195
	Total	**5**	**195**	**200**

In fact, if two observers both know an abnormality is uncommon, they can have nearly perfect agreement just by never or rarely saying that it is present.

Kappa for Dichotomous Variables

To address this problem, another measure of interobserver agreement, called kappa (the Greek letter κ), is sometimes used. Kappa measures the extent of agreement inside a table, such as the ones in Examples 5.1 and 5.2, beyond what would be expected from the observers' overall estimates of the frequency of the different categories. The observers' estimated frequency of observations in each category is found from the totals for each row and column on the outside of the table. These outside totals are called the marginals in the table. Thus, kappa measures agreement beyond what would be expected from the marginals. Kappa ranges from −1 (perfect disagreement) to +1 (perfect agreement). A kappa of 0 indicates that the amount of agreement was exactly what would be expected from the marginals. Kappa is calculated as:

$$\text{Kappa} = \frac{\text{Observed \% agreement} - \text{Expected \% agreement}}{100\% - \text{Expected \% agreement}} \tag{5.1}$$

Observed % agreement is the same as the concordance rate.

Calculating Expected Agreement

We obtain expected agreement by adding the expected agreement in each cell along the diagonal. For each cell, the number of agreements expected from the marginals is the proportion of total observations found in that cell's row (the row total divided by the sample size) times the total number of observations found in that cell's column (the column total). We will illustrate why this is so in the next section.

In Example 5.1, the expected number in the "Abnormal/Abnormal" cell is $60/200 \times 70 = 21$. The expected number in the "Normal/Normal" cell is $140/200 \times 130 = 0.7 \times 130 = 91$. So, the total expected number of agreements is 21 + 91 = 112, and the expected % agreement is 112/200 = 56%. In contrast, in Example 5.2, in which both observers agree that abnormality was uncommon, the expected % agreement is much higher:

$$[(5/200 \times 5) + (195/200 \times 195)]/200 \sim 95\ \%.$$

Understanding Expected Agreement

The expected agreement used in calculating kappa is sometimes referred to as the agreement expected by chance alone. We prefer to call it agreement expected from the marginals,

Table 5.1 Formula for kappa

		Rater #2		
		+	−	Total
Rater #1	+	a	b	R_1
	−	c	d	R_2
	Total	C_1	C_2	N

Observed % agreement (sum along diagonal and divide by N):	(a + d)/N
Expected number for +/+ cell:	$R_1/N \times C_1$
Expected number for −/− cell:	$R_2/N \times C_2$
Expected % agreement (sum expected numbers along diagonal and divide by N):	$(R_1/N \times C_1 + R_2/N \times C_2)/N = (R_1 \times C_1 + R_2 \times C_2)/N^2$

$$\text{Kappa} = \frac{\text{Observed \% agreement} - \text{Expected \% agreement}}{100\% - \text{Expected \% agreement}}$$

because it is the agreement expected by chance only under the assumption that the marginals are fixed and known to the observers, which is generally not the case.

To understand where the expected agreement comes from, consider the following thought experiment. After the initial reading that resulted in Table 5.1, suppose our two radiologists are each given back their stack of 200 films with a jellybean jar containing numbers of red and green jelly beans corresponding to their initial readings. For example, since Radiologist #1 rated 70 of the films abnormal and 130 normal, she would get a jar with exactly 70 red and 130 green jellybeans. Her instruction is then to close her eyes and draw out a jellybean for each X-ray in the stack. If the jellybean is red, she calls the film abnormal. If the jellybean is green, she calls the film normal. After she has "read" the film, she eats the jellybean. (This is known in statistics as "sampling without replacement.") When she is finished, she takes the stack of 200 films to Radiologist #2 (and retreats to the privacy of the reading room to reflect on the wisdom of eating 200 jellybeans). Radiologist #2 is given the same instructions; only his bottle has the numbers of colored jellybeans in proportion to his initial reading, that is, 60 red jellybeans and 140 green ones. The average agreement between the two radiologists over many repetitions of the jellybean method is the expected agreement, given their marginals.

If both radiologists have mostly green or mostly red jellybeans, their expected agreement will be more than 50%. In fact, in the extreme example, where both observers call all the films normal or abnormal, they will be given all green or all red jellybeans, and their "expected" agreement will be 100%. Kappa addresses the question: How well did the observers do compared with how well they would have done if they had jars of colored jelly beans in proportion to their totals (marginals), and they had used the jellybean color to read the film?

Now, why does multiplying the proportion in each cell's row by the number in that cell's column give you the expected number in that cell? Because if Radiologist #1 thinks 35% of the films are abnormal and agrees with Radiologist #2 no more than at a level expected from

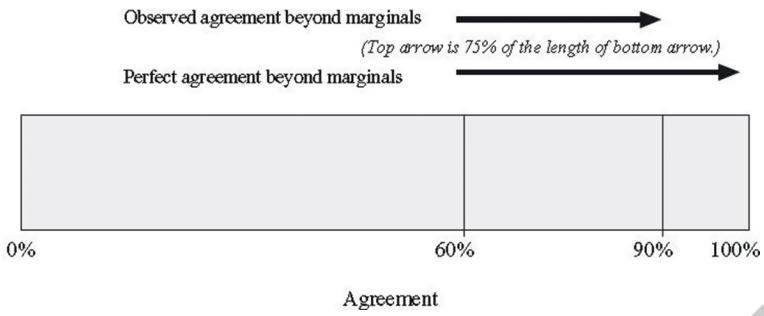

Figure 5.1 Visualizing kappa as the proportion of the way from expected to perfect agreement the observed agreement was for Example 5.3.

that, then she should think 35% of the films rated by Radiologist #2 are abnormal, regardless of how they are rated by Radiologist #2.[1]

"Wait a minute!" we hear you cry, "In real-life studies, the marginals are seldom fixed." In general, no one tells the participants what proportion of the subjects are normal. You might think that if they manage to agree on the fact that most are normal they should get some credit. This is, in fact, what can be counterintuitive about kappa. But that is how kappa is defined, so if you want to give credit for agreement on the marginals, you will need to use another statistic.

Understanding the Kappa Formula

Before we calculate some values of kappa, let us make sure you understand Eq. (5.1). The numerator is how much better the agreement was than what would be expected from the marginals. The denominator is how much better it could have been, if it were perfect. So, kappa can be understood as the percent of the way from the expected agreement to perfect agreement the observed agreement was.

Example 5.3 Fill in the blanks: If expected agreement is 60% and observed agreement is 90%, then kappa would be __ because 90% is __% of the way from 60% to 100% (Figure 5.1).

Answers: 0.75, 75.

Example 5.4 Fill in the blanks: If expected agreement is 70% and observed agreement is 80%, then kappa would be __ because 80% is __% of the way from 70% to 100%.

Answers: 0.33, 33

Returning to Example 5.1, because the observed agreement is 75% and expected agreement 56%, kappa is (75% − 56%)/(100% − 56%) = 0.43. That is, the agreement beyond expected, 75% − 56% = 19%, is 43% of the maximum agreement beyond expected, 100% − 56% = 44%. This is respectable, if somewhat less impressive than 75% agreement.

[1] In probability terms, if the two observers are *independent* (that is, not looking at the films, just guessing using jellybeans), the probability that a film will receive a particular rating by Radiologist # 1 and another particular rating by Radiologist #2 is just the product of the two marginal probabilities.

Similarly, in Example 5.2, kappa is (95% − 95%)/(100% − 95%) = 0, indicating that the degree of agreement was only what would be expected based on the marginals.

Impact of the Marginals

If the percent agreement stays roughly the same, kappa will decrease as the proportion of positive ratings becomes more extreme (farther from 50%). This is because, as the expected agreement increases, the room for agreement beyond expected is reduced. Although this has been called a paradox [2], it only feels that way because of our ambivalence about whether two observers should get credit for recognizing how rare or common a finding is.

Example 5.5 Yen et al. [3] compared abdominal exam findings suggestive of appendicitis, such as tenderness to palpation and absence of bowel sounds, between pediatric emergency physicians and pediatric surgical residents. Abdominal tenderness was present in roughly 60% of the patients and bowel sounds were absent in only about 6% of patients. The physicians agreed on the presence or absence of tenderness only 65% of the time, and the kappa was 0.34. In contrast, they agreed on the presence or absence of bowel sounds an impressive 89% of the time, but because absence of bowel sounds was rare, kappa was essentially zero (−0.04). They got no credit for agreeing that absence of bowel sounds was a rare finding.

Balanced versus Unbalanced Disagreement

When, as is often the case, kappa is disappointing, we should try to understand why. In many cases, additional investigation, and then targeted training, can help improve reliability.

One reason for poor reliability that suggests a possible solution is when disagreement is *unbalanced*. Unbalanced disagreement can occur if one observer has a lower threshold than the other for stating that a finding is present. (This is reminiscent of different cutoffs for defining a positive test for which we used ROC curves in Chapter 3.) For example, if one of the observers is hard of hearing, he won't hear as many heart murmurs as the other unless he gets an amplified stethoscope. Alternatively, if one observer attaches greater importance to avoiding false negatives (i.e., to not missing a finding), he will call more questionable findings present than an observer who wants to avoid false-positives. While two such observers will have different overall prevalence of positive findings (as reflected in their marginals), a better way to look for this kind of unbalanced disagreement is to focus on the subjects about whom they disagree. That is, compare the numbers above and below the agreement diagonal.

Example 5.6A The age at which girls in the United States first experience breast development has been dropping over the last generation, [4] likely due to some combination of increasing obesity and exposure to environmental pollutants [5, 6]. Terry et al. [7] compared ratings of mothers and trained clinicians as to whether breast development had begun among 282 girls aged 6–15 years old (mean age 9.5 years). Results are shown in Table 5.2. You can see that clinicians detected breast development in 44% of the girls, slightly more than the 37% detected by the mothers, but this difference does not seem particularly impressive.

However, if you look at the girls on whom they disagreed, a clear pattern is evident: there were 28 girls in whom the clinician but not the mother believed breast development had begun, but only 9 in whom the opposite was the case. This imbalance can be tested statistically using McNemar's test; the P-value is 0.003.[2]

Table 5.2 Comparison of mothers' and clinicians' determinations of whether breast development had begun in 282 girls

	Clinicians			
Mothers	**No**	**Yes**	**Total N**	**Total (%)**
No	150	28	178	**63**
Yes	9	95	104	**37**
Total N	159	123	282	**100**
Total %	**56%**	**44%**	**100%**	

Data from Terry et al. [7].

Although less common, unbalanced disagreement of ratings can occur in studies of intra-rater reliability (comparing the same observer at different time points) as well as inter-rater reliability (comparing two or more different observers). For example, imagine a radiologist reviewing the same set of X-rays before and after being sued for missing an abnormality. We might expect the readings to differ systematically, with unbalanced disagreements favoring abnormality on the second reading that were normal on the first reading.

Kappa versus Sensitivity and Specificity

In the study described in Example 5.6A, the girls themselves were also asked to determine their stage of pubertal development (on questionnaires that used line drawings to depict the five Tanner stages of puberty), so their answers could also be compared with those of their mothers and the clinicians. In addition to providing kappa statistics, the authors reported sensitivity and specificity of the girls (for those at least 10 years old) and their mothers at determining whether breast development had begun, using clinicians as the gold standard. They found similar sensitivities (both 83%), but lower specificity of the girls compared with their mothers (61% vs. 79%).

As discussed in Chapter 4, if the clinicians were an imperfect gold standard, the sensitivity and specificity reported above could be misleading – either too high, if errors were correlated, or too low if not. To help justify their use of clinicians as a gold standard, the authors also reported inter-rater reliability between clinicians. This was excellent, with kappa ranging from 0.94 to 1.00 for pairwise comparisons between the three raters, suggesting that using them as a gold standard was reasonable.

[2] It is easy to find web-based calculators to do the McNemar test: just Google "McNemar test calculator." (Don't be alarmed if the test mentions matched case–control studies; the same test is used for them.) If you play around a little you can find that the result depends only on the disagreement cells – just the two numbers. This makes sense: the number of times the observers agree provides no information about whether disagreement is unbalanced.

In contrast, Siew et al. [8] studied reliability of telemedicine for the assessment of seriously ill children. They compared ratings of seven items on a respiratory observation checklist between observers performing the examination at the bedside and telemedicine observers watching via FaceTime® on an iPad®. They found good interobserver agreement between the bedside and telemedicine observers (kappa 0.6–0.8 for different items).

Importantly, they did *not* use the observations of the bedside observer as the gold standard to estimate sensitivity and specificity, which would have suggested that disagreements were due to errors by the telemedicine observer. Instead, they measured interobserver agreement between two bedside observers and found kappa values in the same 0.6–0.8 range, supporting their conclusion that telemedicine observations were similar in reliability to bedside observations.

Sometimes kappa is used, rather than sensitivity and specificity, even when there is a gold standard. This is most common when there are multiple possible diagnoses being considered simultaneously, so that both the test result and the gold standard are nominal variables. For example, Perry et al. [9] compared results of pre-operative CT scans with operative results in adults with nontraumatic abdominal pain. They classified CT scans and operative findings by anatomic location (e.g., upper gastrointestinal, lower gastrointestinal, etc.) and pathology (perforation, obstruction, bleeding, etc.). They found that kappa for agreement between CT and operative findings was similar for scans performed during regular working hours and scans performed on nights and weekends.

Kappa for Three or More Categories

Unweighted Kappa

So far, our examples for calculating kappa have been dichotomous ratings, like abnormal versus normal radiographs or presence versus absence of breast development. When there are three or more nominal (not ordered) categories, the calculation of kappa is the same: observed agreement is still calculated by looking at the proportion of observations along the diagonal, and expected agreement is calculated as before: for each cell along the diagonal the expected proportion agreement is the proportion in that row times the proportion in that column.

Of course, the more categories there are, all else being equal, the less likely it is that observers will agree on the category by chance alone. The key feature of unweighted kappa is that there is no credit for being close: only the numbers along the perfect agreement diagonal are counted towards the observers' agreement.

Weighted Kappa

Linear Weights

When there are more than two categories, it is important to distinguish between ordinal variables and nominal variables. For ordinal variables, kappa fails to capture all the information in the data, because it does not give partial credit for ratings that are similar, but not the same. Weighted kappa allows for such partial credit. The formula for weighted kappa is the same as that for regular kappa, except that observed and expected agreement are calculated by summing cells, not just along the diagonal, but for the whole table, with each cell first multiplied by a weight for that cell.

The weights for partial agreement can be anything you want, as long as they are used to calculate both the observed and expected levels of agreement. The most straightforward way to do the weights (and the default for most statistical packages) is to assign a weight of

Table 5.3 Linear weights for three categories

		Rater #2		
		Category 1	**Category 2**	**Category 3**
	Category 1	1	1/2	0
Rater #1	Category 2	1/2	1	1/2
	Category 3	0	1/2	1

0 when the two raters are maximally far apart (i.e., the upper right and lower left corners of the k × k table), a weight of 1 when there is exact agreement (along the diagonal from upper left to lower right), and weights proportionally spaced in between for intermediate levels of agreement. Because a plot of these weights against the number of categories between the ratings of the two observers yields a straight line, these are sometimes called "linear weights." We will give you the formula below, but it is easier to just look at some examples and see what we mean.

If there are three categories, the ratings can be at most two categories apart. The cells in the upper right and lower left corners, with maximal disagreement, get a weight of zero. The cells along the diagonal get a weight of 1 for perfect agreement. There is only one other group of cells, and they are half way between the other cells, so it makes sense that they get a weight of 1/2 (Table 5.3).

Similar logic holds for larger numbers of categories; if there are four categories, the weights would be 0, 1/3, 2/3, and 1.

Now for the formula: If there are k ordered categories, for each cell take the number of categories between the two raters, divide by $k - 1$ (the farthest they could be apart) and subtract this from 1. That is,

$$\text{Linear weight for the Cell in "Row i, Column j"} = 1 - \frac{|\, i - j \,|}{k - 1} \tag{5.2}$$

Along the diagonal, i = j and the weight is 1, for perfect agreement. At the upper right and lower left corners, $|i - j| = (k - 1)$, and the weights are 0. You can think of the second part of the weight, which gets subtracted from 1, as the penalty for not being exactly right, which is maximized when the raters disagree completely.

Example 5.6B The mothers, daughters, and clinicians in the study by Perry et al. actually rated the breast development of the girls according to five Tanner stages, with stage 1 indicating no breast development and stage 5 indicating complete breast development. Results comparing mothers to clinicians are shown in Table 5.4. Because Tanner stage is an ordinal variable, it makes sense to use a weighted kappa. The authors reported an unweighted kappa of 0.54 and weighted kappa of 0.72 for these results. That the weighted kappa was higher than unweighted kappa is typical; you can see that the observations cluster along the main diagonal and along the diagonals just above and below it. This means that most of the time when there was not perfect agreement, the mothers and clinicians disagreed by only one Tanner stage. For five categories, linear kappa weights are 1.0, 0.75, 0.5, 0.25 and 0, so parents and clinicians got 75% credit for those (27 + 7 + 4 + 8 + 8 + 7 + 4 + 4 =) 69 girls on whom they disagreed by only one stage.

Table 5.4 Comparison of mothers' and clinicians' Tanner stage ratings for breast development of 282 girls

	Clinicians					
Mothers	**1**	**2**	**3**	**4**	**5**	**Total**
1	150	27	1	0	0	**178**
2	8	25	7	3	0	**43**
3	1	7	19	4	1	**32**
4	0	1	4	12	8	**25**
5	0	0	0	4	0	**4**
Total	**159**	**60**	**31**	**23**	**9**	**282**

Data from Terry et al. [7].

Quadratic Weights

A commonly used alternative to linear weights is quadratic weights. With quadratic weights, the penalty for disagreement at each level, $|i - j|/(k - 1)$, is squared:

$$\text{Quadratic weight for cell at Row i and Column j} = 1 - \left(\frac{i - j}{k - 1}\right)^2 \tag{5.3}$$

Because this penalty $|i - j|/(k - 1)$ is less than 1, squaring it makes it smaller. Smaller penalties mean that quadratic weights give *more credit* for partial agreement. For example, if there are three categories, the weight for partial agreement is $1 - (1/2)^2 = 0.75$, rather than 0.5; if there are four categories, the weight for being off by one category is $1 - (1/3)^2 = 8/9$, rather than 2/3.

Table 5.5 shows the quadratic weights for a five-category variable like Tanner stage.

Recall that the linear weight for being off by one in a 5 × 5 table was 0.75, so the penalty was 0.25, or 1/4. If we square that penalty, we get 1/16, or 0.0625, so the quadratic weight is 0.9375. Not surprisingly, calculating kappa for Table 5.4 using quadratic weights gives an even higher value: 0.86. (Recall unweighted kappa was 0.54 and weighted kappa was 0.72.)

Table 5.5 Quadratic weights for five Tanner stages

	Clinicians				
Mothers	**1**	**2**	**3**	**4**	**5**
1	1	0.9375	0.75	0.4375	0
2	0.9375	1	0.9375	0.75	0.4375
3	0.75	0.9375	1	0.9375	0.75
4	0.4375	0.75	0.9375	1	0.9375
5	0	0.4375	0.75	0.9375	1

Quadratic weighted kappa will generally be higher (and hence look better) than linear weighted kappa, because the penalty for anything other than complete disagreement is smaller. Thus, a simple manipulation available to authors studying reproducibility who want to report higher kappas without actually improving inter-rater agreement is to use quadratic weights.

Custom Weights

Of course, these linear and quadratic weights are just two ways to do the weighting. If you want more generous weights than linear weights, quadratic weights will do the trick. But if you want weights less generous than linear weights, or you believe some disagreements are much worse than others that differ by the same number of categories, you can create your own weights.

Example 5.7 The Glasgow Coma Scale (GCS) is commonly used in emergency department patients to quantify the level of consciousness. Gill et al. [10] examined the reliability of the components of the GCS by comparing scores of two emergency physicians independently assessing the same patient. They chose to use custom weights, giving half-credit for disagreements differing by only one category and no credit for disagreements differing by two or three categories. Their custom weights for the eye-opening component of the GCS are shown below.

Custom weights for the four eye-opening ratings

	None	To pain	To command	Spontaneous
None	1	0.5	0	0
To pain	0.5	1	0.5	0
To command	0	0.5	1	0.5
Spontaneous	0	0	0.5	1

In this case, the kappa using custom weights was greater than the unweighted kappa and slightly less than the linear weighted kappa. This makes sense because linear weighted kappa gave more credit for disagreements differing by one category (2/3 vs. 1/2 weight) or two categories (1/3 vs. 0 weight).

Although weighted kappa is generally used for ordinal variables, it can be used for nominal variables as well, if some types of disagreement are more significant than others. For example, in the previously described study of CT scans for nontraumatic abdominal pain [9], the authors used a weighting scheme that gave more credit if the CT scan's anatomic location was close to that found at operation (e.g., small bowel vs. colon) than if it was farther away (e.g., stomach vs. colon). This might make sense if, for example, the CT scan determined the location of the incision made by the surgeon.

Whatever weighting scheme we choose, it should be symmetric along the diagonal so that the answer does not depend on which observer is #1 and which is #2.

Kappa also generalizes to more than two observers, although then it is much easier to use a statistics software package. When there are more than two observers, calculation of kappa does not require that each of the observers rate each subject in the sample. The number of raters can vary across subjects and the number of subjects can vary across raters; the number of raters can vary from subject to subject. This same approach works with pairs of observers when the observers in each pair vary. Systematic (unbalanced) disagreement can be suspected if observers have very different marginals and confirmed by comparing ratings of observers that have rated the same subjects. Additional information on multi-rater Kappa is provided in Appendix 5.A.

What is a good kappa?

Students frequently ask us what constitutes a good kappa. Two proposed classifications are shown in Table 5.6. For reasons discussed below, the classifications in Table 5.6 are probably most appropriate for dichotomous variables. However, what constitutes a good kappa also depends on the clinical context. The classifications in Table 5.6 seem appropriate for physical examination findings on which agreement is often moderate or worse, and which generally determine which tests to do or at most whether to start treatments that are not particularly onerous. On the other hand, if the kappa is describing agreement between pathologists whose diagnosis could mean the difference between a sigh of relief and a long course of chemotherapy, we would hesitate to call a kappa of 0.81 "almost perfect" or even "very good"!

Table 5.6 Kappa classifications

Kappa	Sackett et al. [11]	Altman [12]
0—0.2	Slight	Poor
0.2—0.4	Fair	Fair
0.4—0.6	Moderate	Moderate
0.6—0.8	Substantial	Good
0.8—1.0	Almost perfect	Very good

We also saw that with two categories and a given level of observed agreement, kappa depends on both the overall prevalence of the abnormality and whether disagreements are balanced or unbalanced. In our discussion of kappa for three or more categories, another problem became apparent. It should be clear from Example 5.6A that kappa depends on the weights used: the same dataset generated kappa values from 0.54 for unweighted kappa to 0.86 using quadratic weights. The kappa classifications in Table 5.6 are probably generous for linear weighted kappa and not appropriate for quadratic weighted kappa.

Reliability of Continuous Measurements

With continuous variables, just as with categorical and ordinal variables, we are interested in the variability in repeated measurements by the same observer or instrument and in the differences between measurements made by two different observers or instruments.

Test–Retest Reliability

The random variability of some continuous measurements is well known. Blood pressures, peak expiratory flows, and grip strengths will vary between measurements done in rapid succession. Because they are easily repeatable, most clinical research studies use the average of several repetitions rather than a single value. We tend to assume that the variability between measurements is random, not systematic, but this may not be the case. For example, the first grip strength measurement might fatigue the subject so that the second measurement is systematically lower. Similarly, a patient might get better at peak flow measurement with practice. In these cases of systematic variability, we cannot assess test–retest reliability, because the quantities (grip strength and peak flow) are changed by the measurement process itself.

When the variability can be assumed to be purely random, it can be approximately constant across all magnitudes of the measurement or it can vary (usually increase) with the magnitude of the measurement.

Within-Subject Standard Deviation and Repeatability

The simplest description of a continuous measurement's variability is the within-subject standard deviation, S_w [13]. This requires a dataset of several subjects on whom the measurement was repeated multiple times. You calculate each subject's sample variance according to the following formula:

$$\text{Single subject sample variance} = \frac{(M_1 - M_{avg})^2 + (M_2 - M_{avg})^2 + (M_3 - M_{avg})^2 + \cdots + (M_N - M_{avg})^2}{(N - 1)} \tag{5.4}$$

where
N is the number of repeated measurements on a single subject,
M_1, M_2, M_3, . . ., M_N are the repeated measurements on a single subject, and
M_{avg} is the average of all N measurements on a single subject.

Then, you average these sample variances across all the subjects in the sample and take the square root to get S_w. When there are only two measurements per subject, the formula for within-subject sample variance simplifies to

$$\text{Sample variance for two measurements} = \frac{(M_1 - M_2)^2}{2} \tag{5.5}$$[3]

Example 5.8 Suppose you want to assess the test–retest reliability of a new pocket blood glucose meter. You measure the finger-stick glucose twice on each of 10 different subjects:

Calculation of within-subject standard deviation on duplicate glucose measurements

	Glucose measurement (mg/dL)			
Specimen	**#1**	**#2**	**Difference**	**Variance = $(M_1 - M_2)^2/2$**
1	80	92	−12	72
2	89	92	−3	4.5
3	93	109	−16	128
4	97	106	−9	40.5
5	103	87	16	128
6	107	104	3	4.5
7	100	105	−5	12.5
8	112	104	8	32
9	123	110	13	84.5
10	127	120	7	24.5
		Average variance = 53.1		
				S_w = 7.3

[3] The 2 in the denominator may look odd, but you will see it is correct if you substitute $(M_1 + M_2)/2$ for M_{avg} in Eq. (5.4) and do the algebra.

For Subject 1, the difference between the two measurements was −12. You square this to get 144 and divide by 2 to get a within-subject variance of 72. Averaging together all 10 variances yields 53.1, so the within-subject standard deviation $\mathbf{S_w}$ is $\sqrt{53.1}$ or 7.3.[4]

If we assume that the measurement error is distributed in a normal (Gaussian or bell-shaped) distribution, then about 95% of our measurements (on a single specimen) will be within 1.96 $\mathbf{S_w}$ of the theoretical true value for that specimen. In this case, 1.96 × 7.3 = 14.3 mg/dL. So about 95% of the meter readings will be within about 14.3 mg/dL of the true value. The difference between two measurements on the same subject is expected to be within $\left(1.96 \times \sqrt{2} =\right) 2.77 \times \mathbf{S_w}$ 95% of the time.[5] In this example, $2.77 \times S_w = 2.77 \times 7.3 = 20.2$. This is called the repeatability. We can expect the difference between two measurements on the same specimen to be less than 20.2 mg/dL 95% of the time.

Why Not Use Average Standard Deviation?

Rather than take the square root of the variance for each subject (that subject's standard deviation) and then average those to get $\mathbf{S_w}$, we first averaged the variances and then took the square root. We did this to preserve desirable mathematical properties – the same general reason that we use the standard deviation (the square root of the mean square deviation) rather than average deviation. However, because the quantities we are going to average are squared, the effect of outliers (subjects from whom the measurement error was much larger than average) is magnified.

Why Not Use the Correlation Coefficient?

A scatterplot of the data in Example 5.8 is shown in Figure 5.2.

You may recall from your basic statistics course that the correlation coefficient measures linear correlation between two measurements, ranging from −1 (for perfect inverse correlation) to 1 (for perfect correlation) with a value of 0 if the two variables have no linear correlation or are independent.[6] For these data, the correlation coefficient is 0.67. Is this a good measure of test–retest reliability? Before you answer, see Example 5.9.

[4] If there are more than two measurements per subject, and especially if there are different numbers of measurements per subject, it is easiest to get the average within-subject variance by using a statistical package to perform a one-way analysis of variance (ANOVA). In the standard one-way ANOVA table, the residual mean square is the within-subject variance [13].

[5] The variance of the difference between two independent random variables is the sum of their individual variances. Since both measurements have variance equal to the within-specimen variance, the difference between them has variance equal to twice that of the within-specimen variance and the standard deviation of the difference is $\sqrt{2} \times S_{within}$. If the difference between the measurements is normally distributed, 95% of these differences will be within 1.96 standard deviations of the mean difference, which is 0.

[6] Our point here is going to be that two measurements can have poor agreement but a correlation coefficient close to 1 because a strong linear relationship does not necessarily imply good agreement. You should also know that two measurements can be closely related and have a correlation coefficient of 0, so long as the relationship isn't linear. For example, if values of x are centered around 0 (e.g., −3, −2, −1, 0, 1, 2, 3) and $y = x^2$, the correlation coefficient between y and x would be 0.

Example 5.9 Let us replace the last pair of measurements (127, 120) in Example 5.8 with a pair of measurements (300, 600) on a hyperglycemic specimen. This pair of measurements does not show very good reliability. The glucose level of 300 mg/dL might or might not prompt a patient using the pocket glucose meter to adjust his insulin dose. The glucose level of 600 mg/dL should prompt him to call his doctor. Here are the new data:

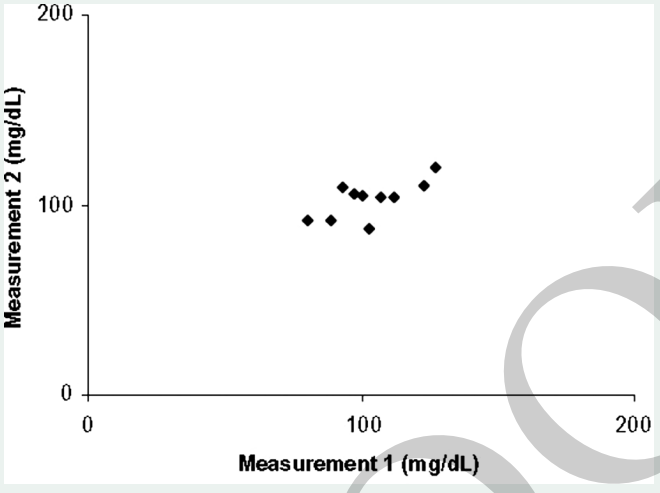

Figure 5.2 Scatterplot of the blood glucose meter readings in Example 5.8. Correlation coefficient = 0.67.

Duplicate glucose measurements from Example 5.8 (except for the last observation)

	Glucose measurement (mg/dL)			
Specimen	**#1**	**#2**	**Difference**	**Variance**
1	80	92	−12	72.0
2	89	92	−3	4.5
3	93	109	−16	128.0
4	97	106	−9	40.5
5	103	87	16	128.0
6	107	104	3	4.5
7	100	105	−5	12.5
8	112	104	8	32.0
9	123	110	13	84.5
10	300	600	−300	45,000.0
		Average Variance = 4,550.7		
				S_w = 67.5

And the new scatterplot is shown in Figure 5.3.

The correlation coefficient for these data is 0.99. We have added a single pair of measurements that do not even agree with each other very well, and yet the correlation coefficient has gone from 0.67 to 0.99 (almost perfect). Meanwhile, the within-subject standard deviation $\mathbf{S_w}$ has increased from 7.3 to 67.5 mg/dL (it has gotten much worse), and the repeatability has increased from 20.2 to 186.9 mg/dL.

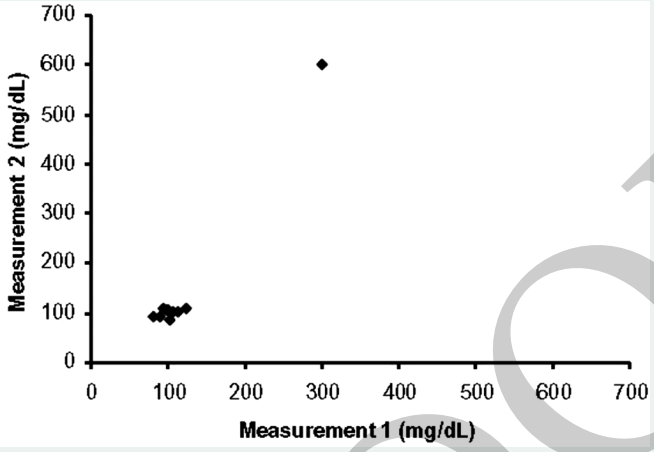

Figure 5.3 Scatterplot of the glucose meter readings in Example 5.9. Correlation coefficient = 0.99.

Although it is tempting to use the correlation coefficient between the first and second measurements on each subject as a measure of reliability, here is why that's usually a bad idea [14]:

1. As we just saw, the correlation coefficient is very sensitive to outliers.
2. (Related to 1): The correlation coefficient will automatically be higher if the range of measurements is higher, even though the precision of the measurement stays the same.
3. The correlation coefficient is high for any linear relationship, not just when the first measurement equals the second measurement. If the second measurement is always 300 mg/dL higher or 40% lower than the first measurement, the correlation coefficient is 1, although the measurements do not agree at all.
4. The test of significance for the correlation coefficient uses the absence of relationship as the null hypothesis. This will almost invariably be rejected because of course there is likely to be a relationship between the first and second measurements, even if they do not agree with each other very well.

Measurement Error Proportional to Magnitude

Sometimes the random error of a measurement is proportional to the magnitude of the measurement. An example of this is where the measurement is accurate to ±5% rather than ± a fixed number of mg/dL. We can visually assess whether error increases with magnitude by graphing the absolute difference between the two measurements versus their average. The glucose meter readings from Example 5.8 show no clear trend of error increasing with the magnitude of the measurement, as shown in Figure 5.4.

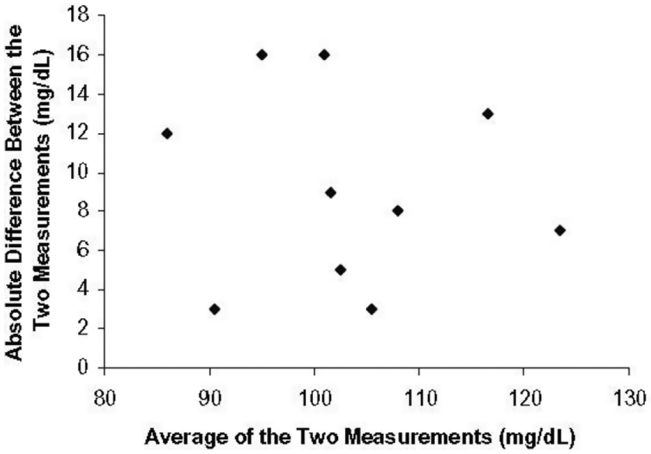

Figure 5.4 Plot of the absolute difference between the two measurements against their average for the data in Example 5.8.

If the random error of a measurement is proportional to the magnitude of the measurement, we cannot use a single value for the within-subject (or within-specimen) standard deviation because it increases with the level of the measurement. In cases like this, rather than estimating the within-subject standard deviation, the variability could be better summarized by the within-subject coefficient of variation (CV), equal to the standard deviation divided by the mean.

Example 5.10 Sometimes, the difference between measurements tends to increase as the magnitude of the measurements increases.

Duplicate glucose measurements illustrating increasing error proportional to the mean

	Glucose measurement (mg/dL)					
Specimen	**#1**	**#2**	**Difference (mg/dL)**	**Mean (mg/dL)**	**SD (mg/dL)**	**CV (%)**
11	93	107	−14	100	9.9	9.9
12	132	117	15	124.5	10.6	8.5
13	174	199	−25	186.5	17.7	9.5
14	233	277	−44	255	31.1	12.2
15	371	332	39	351.5	27.6	7.8
16	364	421	−57	392.5	40.3	10.3
17	465	397	68	431	48.1	11.2
18	518	446	72	482	50.9	10.6
19	606	540	66	573	46.7	8.1
20	682	806	−124	744	87.7	11.8

Again, this is better appreciated by plotting the absolute value of the difference between the two measurements against their average (Figure 5.5).

Note that, in the table for this example, the CV remains relatively constant at about 10%.

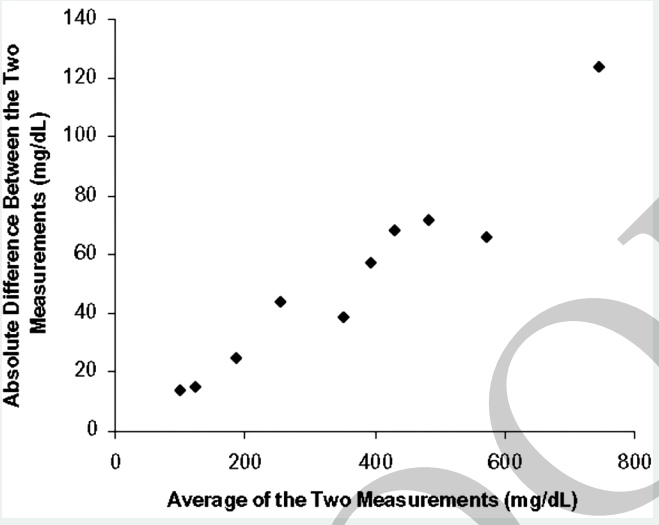

Figure 5.5 Check for error proportional to the mean by plotting the absolute value of the difference between the two measurements against their average. This is done for the data in Example 5.10.

Method Comparison

In our discussion of test–retest reliability, we assumed that no systematic difference existed between initial and repeat applications of a single test. When we compare two different testing methods (or two different instruments or two different testers), we can make no such assumption. Oral temperatures are usually lower than rectal temperatures, abdominal aortic aneurysm diameters are usually lower when assessed by ultrasound than by computed tomography (CT) [15], and mean arterial pressures are usually lower when measured by a finger cuff than by a line in the radial artery [16]. So, when we are comparing two methods, we need to quantify both systematic bias and random differences between the measurements.

Researchers comparing methods of measurement often present their data by plotting the first method's measurement versus the second method's measurement, and by calculating a regression line and a correlation coefficient. We have seen that the correlation coefficient is not good for assessing measurement agreement.

Example 5.11 We compare two methods of measuring bone mineral density (BMD) in children in Figure 5.6: quantitative CT (qCT) and dual-energy X-ray absorptiometry (DXA) [18]. Both qCT and DXA results are reported as Z scores, equal to the number of standard deviations the measurement is from the mean among normals.[7] We would like to get the same Z score whether we use qCT or DXA.

[7] If m and s are the mean BMD and standard deviation in the normal population, then Z = (BMD – m)/s.

In the dataset depicted in Figure 5.6, we can replace a pair of measurements showing moderate agreement (Z_{DXA} = 3, Z_{CT} = 2.25) with a pair of measurements showing *perfect* agreement (Z_{DXA} = 0, Z_{CT} = 0), and the correlation coefficient *decreases* from 0.61 to 0.51. The regression line is not particularly informative either, because we want the two methods of measurement to be interchangeable, not just linearly related. Rather than graph the regression line, most of us would prefer the line of identity on which the measurement by the second method *equals* the measurement by the first method (Figure 5.7).

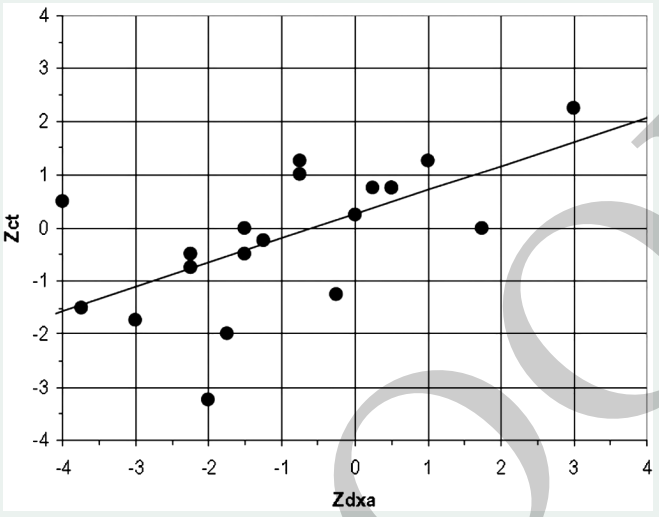

Figure 5.6 Comparison of BMD Z scores obtained by quantitative CT (Z_{CT}) and DXA (Z_{DXA}) (r = 0.61). Fictional data based on Wren et al. [18]

Looking at the points relative to the line of identity in Figure 5.7 reveals that in this dataset, where most of the measurements are negative, qCT gives a higher (less negative) measurement than DXA. This is easier to see by plotting the differences between the

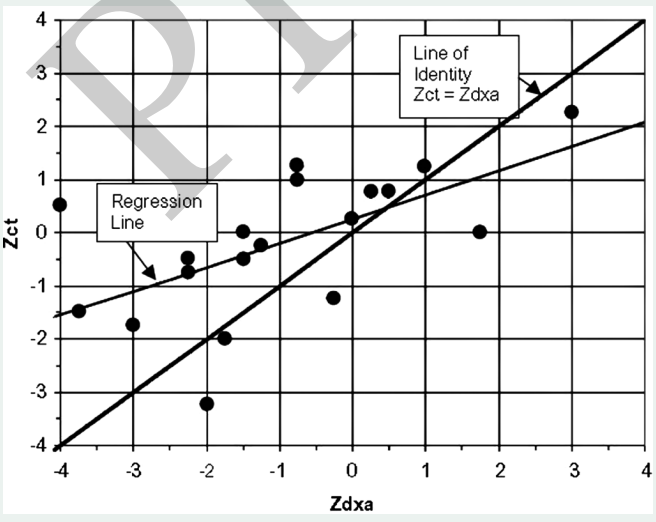

Figure 5.7 Line of identity where $Z_{CT} = Z_{DXA}$. Fictional data based on Wren et al. [18]

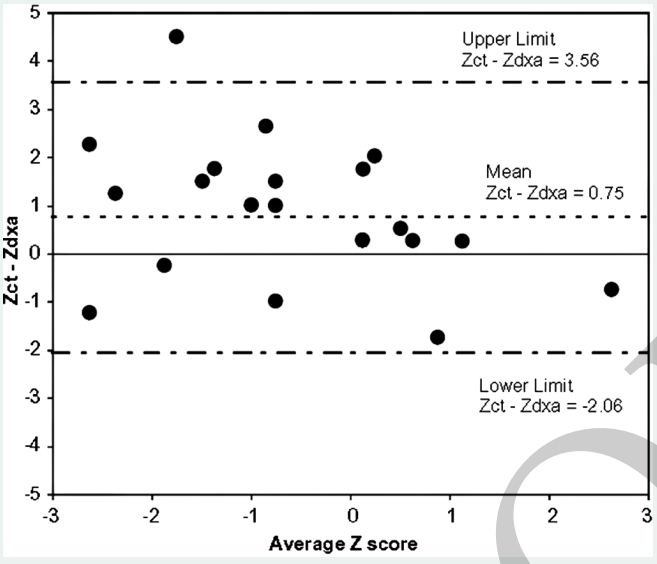

Figure 5.8 Bland–Altman plot showing the difference in BMD Z scores as measured by CT versus DXA with mean difference and 95% limits of agreement.

measurements versus their average, a so-called Bland–Altman plot (Figure 5.8) [1, 17].[8] The mean difference ($Z_{CT} - Z_{DXA}$) is 0.75. The mean difference between two measurements is also sometimes called the "bias." The standard deviation of the differences is 1.43. The 95% limits of agreement are $0.75 \pm (1.96 \times 1.43)$ or -2.06 to 3.56. This means that 95% of the time, the difference between the Z scores, as assessed by CT and DXA, will be within this range. A Bland–Altman plot shows the mean difference and the 95% limits of agreement.

That BMD Z scores by CT are, on average, higher than by DXA is not a severe problem. After all, we know that rectal temperatures are consistently higher than oral temperatures and can adjust accordingly. However, the large variability in the differences and the resulting wide limits of agreement make it hard to accept the use of these BMD measurements interchangeably.

Calibration

Method comparison becomes calibration when one of the two methods being compared using a plot like Figure 5.8 is considered the gold standard and gives the "true" value of a measurement. For a true calibration problem, the gold standard method should be much more precise than the method being compared against it. In fact, the test–retest variability of the gold standard method should be so low that it can be ignored. The x-axis of the plot should then correspond to the measurement by the gold standard method rather than the mean of the two measurements [19]. The plot then compares the difference between the methods versus the gold standard value. This is called a "modified Bland–Altman plot."[9]

[8] According to Krouwer [19], the 1986 article by Bland and Altman on method comparison is the most cited paper ever published in *The Lancet*, a medical journal that has been published weekly since 1823.

[9] This is despite the fact that Bland and Altman said they weren't talking about calibration: "Sometimes we compare an approximate method with a very precise one. This is a calibration problem and we will not discuss it further here." [20].

Example 5.12 Earlier, we assessed the variability of repeated glucose measurements by a finger-stick blood glucose meter. Now, we compare the meter's finger-stick measurement to a

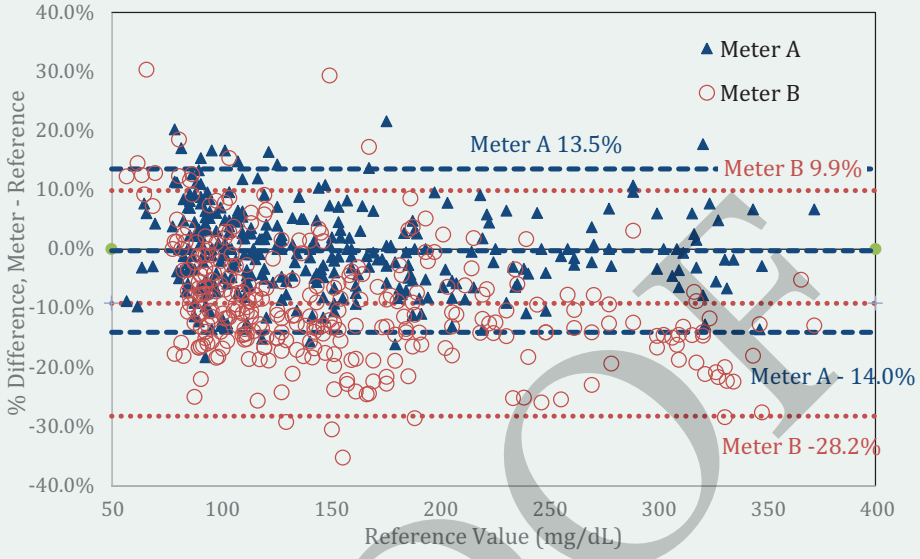

Figure 5.9 Modified Bland-Altman plots comparing finger-stick measurements on Meter A and Meter B to glucose level measured by the reference lab on a simultaneously obtained plasma specimen. Bias (Coefficient of Variation): Meter A −0.3% (7.0%); Meter B −9.2% (9.7%). Meter A = AgaMatrix CVS Advanced, Meter B = Advocate Redi-Code Plus; N = 318.
Data from [21]

laboratory measurement on a simultaneously drawn plasma specimen, which is the reference standard. In fact, we compare two meters, Meter A (AgaMatrix CVS Advanced) and Meter B (Advocate Redi-Code Plus) to the plasma specimen [21]. As in Example 5.10, the difference between the finger-stick and plasma measurements is proportional to the magnitude of the measurement, so we plot percent error rather than absolute error (Figure 5.9). As mentioned above, the mean error is also called the bias, which is reported along with the standard deviation of the error.

Figure 5.9 shows that Meter A provided an unbiased estimate (bias = −0.3%) of the plasma glucose level, while Meter B tended to understate the level (bias = −9.2%). Meter A also had a lower standard deviation, so the zone of 95% agreement is narrower.

We prefer calibration plots like those in Figure 5.9 that plot the differences between two measurements, so we can compare them to the horizontal zero line. It is a better way to visualize measurement agreement (or lack thereof) than plotting one measurement against the other and comparing results to a 45-degree diagonal as in Figure 5.7. Unfortunately, as we will see in Chapter 6, the standard calibration plot used to evaluate predictions uses the 45-degree diagonal.

Using Studies of Reliability from the Literature

In Chapter 4, we provided guidance on critical appraisal of studies of diagnostic test accuracy. In this section, we provide some tips on studies of reliability. First, consider the study subjects. In studies of reliability, there are really two sets of subjects: the patients, who

will have a particular distribution of results or findings, and the examiners. If we want to know whether results are applicable in our own clinical setting, the subjects in the study should be representative of those whom we or our colleagues are likely to test. Specifically, the way that we would like them to be representative is that their findings should be as easy or as difficult to discern as those of patients in our own clinical population – neither very subtle nor very obvious nor very extreme findings should be overrepresented. Watch for studies in which subjects with ambiguous or borderline results are under-sampled or not included at all.

Similarly, consider whether the examiners or instruments used in the study are representative. How were they selected? If examiners were selected because of their interest in interobserver variability or their location in a center that specializes in the problem, they may perform better than might be expected elsewhere. But the opposite is also possible: sometimes investigators are motivated to study interobserver variability or method comparison of instruments in a particular setting because they have the impression that it is poor.

In testing the reliability of two observers on a normal/abnormal rating, it does not make sense to include only subjects rated abnormal by at least one of the two raters. This excludes all the agreements where both raters thought the subject was normal. Nor does it make sense for the second rating to occur only if the first rating is abnormal. This excludes disagreements of type normal/abnormal and only allows type abnormal/normal.

Next, think about the measurements in the study. Were they performed similarly to how such measurements are performed in your clinical setting? Were there optimal conditions, such as a quiet room, good lighting, and/or regular maintenance of the instruments to do the measurements? Are the investigators studying the whole process for making the measurement, or have they selected only a single step? For example, a study of inter-rater reliability of interpretation of mammograms, in which two observers read the same films, will capture variability only in the interpretation of images, not in how the breast was imaged. This will probably overestimate reliability. On the other hand, cardiologists reading a videorecorded echocardiogram might show lower reliability than if they were performing the echocardiogram themselves.

Finally, did the authors investigate predictors of reliability? Associations between variables are often more generalizable across populations than are descriptive statistics on the variables themselves. For example, Terry et al. [7] found that reliability of breast staging was better in younger and slimmer girls. In fact, among girls over 10 years old with a body mass index above the 85th percentile, kappa was −0.06 for both the girls themselves and their mothers (compared with clinicians). Identifying predictors of poor reliability can suggest targeted interventions to improve it or help identify subsets of patients in whom a particular test may be too unreliable to use. In this case, the results suggest not relying on mothers' or daughters' self-assessments of breast development if the girl's body mass index is above the 85th percentile.

For further discussion of these issues, see Chapter 12 of *Designing Clinical Research, 4th Edition* [22].

Summary of Key Points

1. The methods used to quantify inter- and intra-rater reliability of measurements depend on variable type.

2. For categorical variables, the **observed % agreement** is simply the proportion of the measurements upon which both observers agree exactly.
3. Particularly when observers agree that the prevalence of any of the different categories is high or low, it may be desirable to calculate **kappa** (κ), an estimate of agreement beyond that expected based on the row and column totals ("marginals") in a table summarizing the results.
4. For ordered categories, weighted kappa provides partial credit for close but not exact agreement. Linear, quadratic, or custom weights can be used.
5. For continuous measurements, the within-subject standard deviation expresses the spread of repeated measurements around the subject's mean.
6. When measurement error increases with the value of the mean (e.g., a measurement is accurate to ±3%), the **coefficient of variation**, equal to the within-subject standard deviation divided by the mean, is a better way to express reproducibility.
7. Bland–Altman plots are helpful for comparing methods of measurement. They show the scatter of differences between the methods, whether the difference tends to increase with the magnitude of the measurement, and any systematic difference or bias.
8. Comparing an alternative method for making continuous measurements with a highly reliable reference standard is called calibration.

References

1. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;1(8476):307–10.
2. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol*. 1990;43 (6):543–49.
3. Yen K, Karpas A, Pinkerton HJ, Gorelick MH. Interexaminer reliability in physical examination of pediatric patients with abdominal pain. *Arch Pediatr Adolesc Med*. 2005;159(4):373–6.
4. Kaplowitz PB, Oberfield SE. Reexamination of the age limit for defining when puberty is precocious in girls in the United States: implications for evaluation and treatment. Drug and Therapeutics and Executive Committees of the Lawson Wilkins Pediatric Endocrine Society. *Pediatrics*. 1999;104(4 Pt 1):936–41.
5. Maron DF. Early puberty: causes and effects. *Sci Am*. 2015;312(5):28, 30.
6. Ozen S, Darcan S. Effects of environmental endocrine disruptors on pubertal development. *J Clin Res Pediatr Endocrinol*. 2011;3(1):1–6.
7. Terry MB, Goldberg M, Schechter S, et al. Comparison of clinical, maternal, and self pubertal assessments: implications for health studies. *Pediatrics*. 2016;138(1). https://pediatrics.aappublications.org/content/138/1/e20154571.
8. Siew L, Hsiao A, McCarthy P, et al. Reliability of telemedicine in the assessment of seriously ill children. *Pediatrics*. 2016;137(3):e20150712.
9. Perry H, Foley KG, Witherspoon J, et al. Relative accuracy of emergency CT in adults with non-traumatic abdominal pain. *Br J Radiol*. 2016;89(1059): 20150416.
10. Gill MR, Reiley DG, Green SM. Interrater reliability of Glasgow Coma Scale scores in the emergency department. *Ann Emerg Med*. 2004;43(2):215–23.
11. Sackett D, Haynes R, Guyatt G, Tugwell P. *Clinical epidemiology: a basic science for clinical medicine*. Boston: Little, Brown and Company; 1991. 52 p.
12. Altman DG. *Practical statistics for medical research*. 1st ed. London; New York: Chapman and Hall; 1991. xii, 611pp.
13. Bland JM, Altman DG. Measurement error. *BMJ*. 1996;313(7059):744.

14. Bland JM, Altman DG. Measurement error and correlation coefficients. *BMJ*. 1996;313 (7048):41–2.
15. Lederle FA, Wilson SE, Johnson GR, et al. Variability in measurement of abdominal aortic aneurysms. Abdominal Aortic Aneurysm Detection and Management Veterans Administration Cooperative Study Group. *J Vasc Surg*. 1995;21 (6):945–52.
16. Bos WJ, van Goudoever J, van Montfrans GA, van den Meiracker AH, Wesseling KH. Reconstruction of brachial artery pressure from noninvasive finger pressure measurements. *Circulation*. 1996;94 (8):1870–5.
17. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res*. 1999;8(2):135–60.
18. Wren TA, Liu X, Pitukcheewanont P, Gilsanz V. Bone densitometry in pediatric populations: discrepancies in the diagnosis of osteoporosis by DXA and CT. *J Pediatr*. 2005;146(6):776–9.
19. Krouwer JS. Why Bland–Altman plots should use X, not (Y+X)/2 when X is a reference method. *Stat Med*. 2008;27 (5):778–80.
20. Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. *The Statistician*. 1983;32:307–17.
21. Klonoff DC, Parkes JL, Kovatchev BP, et al. Investigation of the accuracy of 18 marketed blood glucose monitors. *Diabetes Care*. 2018;41 (8):1681–8.
22. Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB. *Designing clinical research*. 4th ed. Philadelphia: Wolters Kluwer/Lippincott Williams & Wilkins; 2013.