VIII. ANALYSIS OF OBSERVER VARIABILITY

Before using a measuring or diagnostic technique routinely, a researcher may wish to quantify the extent to which two determinations of the measurement, made by two different observers or measuring devices, disagree (inter-observer variability). He may also wish to quantify the repeatability of one observer in making the measurement at different times (intra-observer variability). To make these assessment, he has each observer make the measurement for each of a number of experimental units (e.g., patients).

The measurements being analyzed may be continuous, ordinal, or binary (yes/no). Ordinal measurements must be coded such that distances between values reflects the relative importance of disagreement. For example, if a measurement has the values 1, 2, 3 for poor, fair, good, it is assumed that "good" is as different from "fair" as "fair" is from "poor". If this is not the case, a different coding should be used, such as coding 0 for "poor" if poor should be twice as far from "fair" as "fair" is from "good". Measurements that are yes/no or positive/negative should be coded as 1 or 0. The reason for this will be seen below.

There are dozens of statistical methods for quantifying inter- and intra-observer variability. Correlation coefficients are frequently reported, but a perfect correlation can result even when the measurements disagree by a factor of 10. Variance components analysis is often used, but this analysis makes many assumptions, does not handle missing data very well, and yields quantities that are difficult to interpret. Some analysts, in assessing inter-observer agreement when each observer makes several determinations, compute differences between the average determinations for each observer. This method clearly yields a biased measurement of inter-observer agreement because it cancels the intra-observer variability.

A general and descriptive method for assessing observer variability will now be presented. For definiteness, the analysis for 3 observers and 2 readings per observer will be used. When designing such a study, the researcher should remember that the number of experimental units is usually the critical factor in determining the precision of estimates. There is not much to gain from having each observer make more than two readings or from having many observers in the study (although if few observers are used, these are assumed to be "typical" observers).

The intra-observer disagreement for a single patient or unit is defined as the average of the intra-observer absolute measurement differences. In other words, intra-observer disagreement is the average absolute difference between any two measurement from the same observer. The inter-observer disagreement for one unit is defined as the average absolute difference between any two readings from different observers. These measurements are computed separately for each unit and combined over units (by taking the mean or median, for example) to get an overall summary measure. When a reading is missing, that reading does not enter into any calculation and the denominator used in finding the mean disagreement is reduced by one.

Suppose that for one patient, observers A, B, and C make the following determinations in a blood chemistry test on two separate occasions, all on the same patient:

A	B	C
5,7	8,5	6,7

For that patient, the mean intra-observer difference is ($|5-7| + |8-5| + |6-7|$)/3 = (2+3+1)/3 = 2. The mean inter-observer difference is ($|5-8| + |5-5| + |5-6| + |5-7| + |7-8| + |7-5| + |7-6| + |7-7| + |8-6| + |8-7| + |5-6| + |5-7|$)/12 = (3+0+1+2+1+2+1+0+2+1+1+2)/12 = 16/12 = 1.33. If the first reading for observer A were unobtainable, the mean intra-observer difference for that patient would be ($|8-5| + |6-7|$)/2 = (3+1)/2 = 2 and the mean inter-observer difference would be ($|7-8| + |7-5| + |7-6| + |7-7| + |8-6| + |8-7| + |5-6| + |5-7|$)/8 = (1+2+1+0+2+1+1+2)/8 = 10/8 = 1.25.

The computations are carried out in like manner for each patient and summarized as follows:

Patient	Intra-observer Difference	Inter-observer Difference
1	2.00	1.33
2	1.00	3.50
3	1.50	2.66
.	.	.
.	.	.
n	.	.
Overall Average (or median)	1.77	2.23
$Q_{.25}$	.30	.38
$Q_{.75}$	2.15	2.84

When the measurement of interest is a yes/no determination such as presence or absence of a disease, these difference statistics are generalizations of the fraction of units in which there is exact agreement in the yes/no determination, when the absolute differences are summarized by averaging. To see this, consider the following data with only one observer:

Patient	Determinations:		D_1, D_2		Agreement?	$\lvert D_1 - D_2 \rvert$
1	Y	Y	1	1	Y	0
2	Y	N	1	0	N	1
3	N	Y	0	1	N	1
4	N	N	0	0	Y	0
5	N	N	0	0	Y	0
6	Y	N	1	0	N	1

The average $|D_1-D_2|$ is 3/6=.5 which is equal to the proportion of cases in which the two readings disagree.

An advantage of this method of summarizing observer differences is that the investigator can judge what is an acceptable difference and he can relate this directly to the summary disagreement statistic.

Comparison of Measurements with Standard

When the true measurement is known for each unit (or the true diagnosis is known for each patient), similar calculations can be used to quantify the extent of errors in the measurements. For each unit, the average (over

observers) difference from the true value is computed and these differences are summarized over the units. For example, if for unit #1 observer A measures 5 and 7, observer B measured 8 and 5, and the true value is 6, the average absolute error is $(|5-6| + |7-6| + |8-6| + |5-6|)/4 = (1+1+2+1)4 = 5/4 = 1.25$.

Assessing Agreement Between Two Binary Variables

Measuring Agreement Between Two Observers

Suppose that each of n patients undergoes two diagnostic tests that can yield only the values positive and negative. The data can be summarized in the following frequency table.

		Test 2 +	Test 2 -	
Test 1	+	a	b	g
	-	c	d	h
		e	f	n

An estimate of the probability that the two tests agree is $p_A=(a+d)/n$. A 95% confidence interval for the true probability is derived from

$$p_A \pm 1.96 \sqrt{p_A (1-p_A)/n}$$

If the disease being tested is very rare or very common, the two tests will agree with high probability by chance alone. The K statistic is one way to measure agreement that is corrected by chance (1):

$$K = \frac{p_A - p_C}{1 - p_C}$$

where p_C is the expected agreement proportion if the two observers are completely independent. The statistic can be simplified to

$$K = \frac{2 (ad-bc)}{gf + eh}$$

If the two tests are in perfect agreement, K=1. If the two agree at the level expected by chance, K=0. If the level of agreement is less than one would obtain by chance alone, K<0.

A formal test of significance of the difference in the probabilities of + for the two tests is obtained using McNemar's test. The null hypothesis is that the probability of + for test 1 is equal to the probability of + for test 2, or equivalently that the probability of observing a +- is the same as that of observing -+. The normal deviate test statistic is given by

$$z = \frac{b-c}{\sqrt{b+c}}$$

Measuring Agreement Between One Observer and a Standard

Suppose that each of n patients is studied with a diagnostic test and that the true diagnosis is determined, resulting in the following frequency table:

		Diagnosis +	Diagnosis -	
Test	+	a	b	g
	-	c	d	h
		e	f	n

The following measures are frequently used to describe the agreement between the test and the true diagnosis. Here T^+ denotes a positive test, D^- denotes no disease, etc.

Quantity	Probability Being Estimated	Formula
Correct diagnosis rate	Prob (T = D)	(a+d)/n
Sensitivity	Prob $(T^+ \mid D^+)$	a/e
Specificity	Prob $(T^- \mid D^-)$	d/f

Note that when the disease is very rare or very common, the correct diagnosis rate will be high by chance alone. Since the sensitivity and specificity are calculated conditional on the diagnosis, the prevalence of disease does not affect these measures. For this reason sensitivity and

specificity are often preferred over the proportion of correct classifications as measures of accuracy.

When estimating any of these quantities, confidence intervals are useful adjunct statistics. A 95% confidence interval is obtained from

$$p \pm 1.96 \sqrt{p(1-p)/m},$$

where p is the proportion and m is its denominator.

Problems

1. Three technicians, using different machines, make 3 readings each. For the data which follow, calculate estimates of inter- and intra-technician discrepancy.

Technician 1			Technician 2			Technician 3		
Reading			Reading			Reading		
1	2	3	1	2	3	1	2	3
18	17	14	16	15	16	12	15	12
20	21	20	14	12		13		
26	20	23	18	20		22	24	
19	17		16			21	23	
28	24		32	29		29	25	

2. Forty-one patients each receive two tests, yielding the frequency table shown below. Calculate a measure of agreement (or disagreement) along with an associated 95% confidence interval. Also calculate a chance-corrected measure of agreement. Test the null hypothesis that the two tests have the same probability of being positive and the same probability of being negative. In other words, test the hypothesis that the chance of observing +- is the same as observing -+.

		Test 2	
		+	-
Test 1	+	29	8
	-	0	4

References

1. Landis JR, Koch GG: A review of statistical methods in the analysis of data arising from observer reliability studies (Part II). Statistica Neerlandica 29:151-61, 1975.
2. Landis JR, Koch GG: An application of hierarchical Kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics 33:363-74, 1977.

NOTE: Zhouwen Liu and Frank Harrell have developed much more general purpose software for this in R.

VIII. ANALYSIS OF OBSERVER VARIABILITY

The following example SAS program will calculate the general measures of intra- and inter-observer disagreement described above for the case where there are up to 4 observers (A-D) and up to 3 readings per observer on each experimental unit. Missing data are handled by the program. The first line of input data is used to define the pattern of observers. For example, the first 3 readings are for observer A.

```
DATA disagree;
*Define which measurements, corresponding to x1-x12, come from
 the same observer;
RETAIN o1-o12 ' '; IF _n_=1 THEN INPUT o1-o12;
*Input a line of data;
INPUT x1-x12;
*Compute measures of intra- and inter-observer disagreement for
 each experimental unit;
ARRAY x[*] x1-x12;   ARRAY obs[*] o1-o12;
nintra=0; sintra=0; ninter=0; sinter=0;
     DO j=1 TO 11;
          DO k=j+1 TO 12;
          d=abs(x[j]-x[k]);
          IF d>. THEN DO;
               IF obs[j]=obs[k] THEN DO; *From same observer;
                    nintra=nintra+1; sintra=sintra+d;
                    END;
               ELSE DO;
                    ninter=ninter+1; sinter=sinter+d;
                    END;
               END;
          END;
```

```
     END;
intra=sintra/nintra; inter=sinter/ninter;
KEEP x1-x12 intra inter ninter nintra;
CARDS;
A A A B B B C C C D D D
1 2 3 . . . . . .
1 3 . 2 7 . . . . . .
1 2 3 4 5 6 7 8 9 7 8 9
PROC PRINT;
*Now summarize measures over experimental units;
PROC UNIVARIATE;VAR intra inter;
```