STATISTICAL MEASUREMENT OF INTEROBSERVER AGREEMENT

STATISTISCHE METING VAN

OVEREENSTEMMING TUSSEN BEOORDELAARS

Proefschrift

ter verkrijging van de graad van doctor in de
geneeskunde
aan de Erasmus Universiteit Rotterdam
op gezag van de Rector Magnificus
Prof. Dr. M.W. van Hof
en volgens besluit van het College van Dekanen.
De openbare verdediging zal plaatsvinden op
woensdag 20 november 1985 te 15.45 uur

door

HUBERTUS JOSEPH ANTONIUS SCHOUTEN

geboren te Utrecht

drukkerij Elinkwijk B.V. - Utrecht

Promotiecommissie

Promotoren:	Prof. R. van Strik
	Prof. Dr. W. Molenaar
Overige leden:	Prof. Dr. A.P.J. Abrahamse
	Prof. Dr. G.J. Mellenbergh

To those observers
who are not convinced
that they are right

To Jeanne, Marijn and Hedi
To my mother
To the memory of my father

Contents

PREFACE

In 1976 the neurosurgeon R. Braakman stimulated me to think about inter-observer agreement. He asked for statistical advice concerning several studies that were designed to investigate and possibly improve the reproducibility of certain medical judgements. R. van Strik mentioned some relevant statistical papers from which I concluded that the then existing methodology was far from complete and I started to think of possible extensions. Building on the papers by Scott(1955), Cohen(1960, 1968), Fleiss(1971) and many others, I am now able to present a rather complete statistical methodology concerning the analysis of agreements and disagreements between observers who classify subjects. This methodology is based on the estimated probabilities that the one observer uses category i and the other observer uses category j, and on statistics derived from these estimated probabilities.

This thesis could not have been written without the help given by other people. R. van Strik, W. Molenaar and R. Popping constructively criticized all versions of the text and their suggestions greatly improved the substance of this book. Of course, all remaining errors, omissions and obscurities are my responsibility, and I will eagerly receive useful criticisms from any readers.

Hubert J.A. Schouten

INLEIDING EN SAMENVATTING
(INTRODUCTION AND SUMMARY in Dutch)

Wanneer een patiënte door een arts wordt onderzocht, is het wenselijk dat de bevindingen (diagnose, symptomen) niet anders uitvallen dan wanneer zij door een andere arts wordt onderzocht. Het boezemt geen vertrouwen in als artsen onderling ernstig van mening verschillen. De patiënte zal concluderen dat deze artsen het niet weten. Ze kunnen immers niet allemaal gelijk hebben. Het voorgaande geldt natuurlijk niet alleen voor artsen maar voor beoordelaars in het algemeen, vooral in situaties waarin het niet goed mogelijk is de waarheid objectief vast te stellen.

Bij zowel medisch als niet-medisch wetenschappelijk onderzoek kunnen verbanden tussen verschillende kenmerken (karaktereigenschappen, woonsituatie, diagnose, medische voorgeschiedenis, mate van genezing) ernstig worden versluierd door fouten bij het beoordelen van deze kenmerken. Daardoor kunnen belangrijke verbanden over het hoofd worden gezien. Wel ontdekte verbanden kunnen soms nauwelijks worden geïnterpreteerd doordat verschillende studies niet tot dezelfde kwantitatieve conclusies komen.

Het in kaart brengen van meningsverschillen kan de eerste stap zijn naar een betere onderlinge overeenstemming tussen beoordelaars: Welke meningsverschillen treden vaak op en welke beoordelaars zijn het vaak niet met elkaar eens? Dit kan leiden tot een verdere standaardisatie van de wijze van beoordelen en tot regelmatige training aan de hand van moeilijke gevallen; op deze tweede stap wordt echter in dit boek niet ingegaan.

De waarde van beoordelingen, zowel voor de dagelijkse praktijk als voor wetenschappelijk onderzoek, kan wezenlijk toenemen als de onderlinge overeenstemming tussen beoordelaars merkbaar wordt verbeterd. Maar volledigheidshalve moet worden opgemerkt dat zelfs volledige overeenstemming tussen beoordelaars niet garandeert dat de betreffende beoordelingen enige waarde hebben. Daarom moeten we bij het vereenvoudigen van een beoordelingssysteem vermijden dat belangrijke

informatie verloren gaat.

In dit boek worden experimenten beschouwd waarbij subjecten (bijvoorbeeld patiënten) worden ingedeeld in categorieën door minstens twee beoordelaars (bijvoorbeeld artsen). De categorieën staan bij voorbaat vast en zijn voor alle beoordelaars hetzelfde. Een subject wordt door een beoordelaar aan precies één categorie toegewezen. Het is niet noodzakelijk dat alle subjecten door dezelfde beoordelaars worden beoordeeld en ook het aantal beoordelaars per subject mag variëren. Voor een bepaald subject is er volledige overeenstemming tussen beoordelaars als zij dit subject aan dezelfde categorie toewijzen. De mogelijke meningsverschillen tussen twee beoordelaars zijn niet alle even ernstig als de categorieën geordend zijn.

In het eerste hoofdstuk wordt bekeken hoe de overeenstemming tussen twee beoordelaars kan worden onderzocht. Het is niet voldoende om na te gaan of de verdeling van de beoordelingen over de categorieën voor beide beoordelaars ongeveer hetzelfde is. Het vergelijken van gemiddelden is nog minder bevredigend. De correlatiecoëfficiënt en de fractie overeenstemming zijn evenmin goede maten van overeenstemming tussen twee beoordelaars. Bij het onderzoeken van overeenstemming tussen twee beoordelaars moet rekening worden gehouden met de volgende punten:

i) Als beide beoordelingen niet afhangen van het te beoordelen subject, zijn deze beoordelingen statistisch onafhankelijk. Zelfs dan kan vaak overeenstemming optreden, louter en alleen door het toeval.

ii) Het is te verwachten dat een bepaald meningsverschil vaker zal optreden naarmate de erbij betrokken categorieën vaker worden gebruikt.

De overeenstemmingsmaten kappa, upsilon en gewogen kappa worden daarom zo gedefinieerd dat ze de waarde nul hebben bij toevallige overeenstemming en de waarde één bij perfecte overeenstemming.

Kappa wordt gedefinieerd als de fractie meer dan toevallige overeenstemming. Een belangrijke eigenschap van kappa komt tot uiting

in de volgende interpretatie van een nieuwe wiskundige stelling: Als het samenvoegen van categorieën leidt tot een hogere kappa-waarde, dan zijn de betreffende categorieën relatief moeilijk van elkaar te onderscheiden.

Upsilon is een goede maat van overeenstemming als de categorieën geordend zijn, mits we bereid zijn de toewijzing aan een categorie te beschouwen als een kwantitatieve beoordeling. Upsilon is gelijk aan de correlatiecoëfficiënt als de beoordelingen van twee beoordelaars hetzelfde gemiddelde en dezelfde standaardafwijking hebben. Anders is upsilon kleiner dan de correlatiecoëfficiënt Van een kenmerk met een lage upsilon-waarde valt niet te verwachten dat het sterk gecorreleerd zal zijn met een ander kenmerk. Kappa en upsilon zijn gelijk aan elkaar als er slechts twee categorieën zijn.

In het tweede en derde hoofdstuk wordt de onderlinge overeenstemming binnen een bepaalde groep beoordelaars onderzocht. Dit wordt gedaan aan de hand van de overeenstemming tussen twee willekeurige beoordelaars (d.w.z. aselect uit die groep gekozen, zonder "terugleggíng"). Dat heeft de volgende voordelen:

i) De getalswaarden van de gebruikte overeenstemmingsmaten zijn niet afhankelijk van het totale aantal beoordelaars.

ii) Door op directe wijze te generaliseren van twee naar meer dan twee beoordelaars ontstaat een uniforme statistische methode: hetgeen voor twee beoordelaars waardevol is, heeft ook waarde voor meer dan twee beoordelaars.

Samen impliceren deze voordelen een algemene toepasbaarheid. Op aseaselecte wijze ontbrekende beoordelingen vormen geen probleem en het aantal beoordelaars per subject hoeft niet constant te zijn.

De beschreven statistische methoden worden toegelicht aan de hand van voorbeelden uit de medische diagnostiek. Bij een groep van zeven pathologen, die ieder afzonderlijk 118 coupes (weefselsneden) van biopten uit de baarmoederhals op kanker beoordelen, wordt geprobeerd één of meer homogene deelgroepen van pathologen te ontdekken. Uit een ander voorbeeld, waarin zes artsen bij 28 patiënten in coma de pupilreactie op licht beoordelen, blijkt dat op aselecte wijze ontbrekende

beoordelingen geen onoverkomelijk probleem vormen. Voor een wisselende groep psychiaters, die 30 patiënten indelen in vijf diagnostische categorieën, blijken bepaalde categorieën relatief moeilijk van elkaar te onderscheiden.

Belangrijke informatie kan ook worden verkregen door de overeenstemming binnen beoordelaars te onderzoeken. Met dat doel worden subjecten een tweede keer door dezelfde beoordelaars in klassen ingedeeld. De verkregen resultaten kunnen worden geanalyseerd met de in dit boek beschreven statistische methoden.

In alle voorbeelden wordt de standaard-knipmes-methode (Engels: standard jackknife technique) toegepast om de betrouwbaarheid van de getrokken conclusies aan te geven. In het vierde hoofdstuk worden ook de delta-methode en de methode-von-Münchhausen (Engels: bootstrap) beschouwd.

INTRODUCTION AND SUMMARY

When a patient is examined by a physician, it is desirable that the findings (diagnosis, symptoms) do not change when she is examined by a different physician. The patient is not inspired with confidence if physicians seriously differ in opinion, just because they cannot all be right. Of course, the foregoing does not only apply to physicians, but applies to judging observers in general, especially in situations where it is impossible to establish the truth in an objective way.

In medical and non-medical scientific research associations between different characteristics (personality aspects, housing conditions, diagnosis, medical history, degree of recovery) may be blurred by errors in the assessment of these characteristics. Thereby important associations may be overlooked. Associations that are discovered may be hard to interpret because different studies may result in different quantitative conclusions.

The specification of disagreement may be the first step to a better agreement among observers: Which disagreements do frequently occur and which observers often have different opinions? This may lead to a further standardization of the way of judging and to regular training on the basis of difficult cases; this second step, however, is not considered in this book.

The value of judgements, in daily practice and in scientific research, may greatly increase by substantially improved interobserver agreement. For the sake of completeness, however, it must be mentioned that even perfect interobserver agreement does not guarantee that judgements have any value. Therefore, when simplifying a judgement system, one must avoid to lose important information.

In the present book experiments are considered where subjects (e.g. patients) are placed in categories by at least two observers (e.g. physicians). The categories are determined in advance and are the same for all observers. A subject is assigned by an observer to one category. It is not necessary that all subjects are judged by the same observers,

and also the number of observers per subject may vary. With respect to a particular subject, there is perfect agreement among the observers if they all use the same category. The possible disagreements between two observers are not equally serious if the categories are ordered.

In the first chapter it is considered in which way agreement between two observers may be investigated. It is not sufficient to check if the distribution of the judgements over the categories is about the same for both observers. Comparing averages is even less satisfactory. Neither the correlation coefficient nor the proportion of agreement are good measures of agreement between two observers. When investigating the agreement between two observers, the following points must be taken into account:

i) If the two judgements do not depend on the subject that is to be judged, these judgements are statistically independent. Even then agreement may often occur, solely and alone by chance.

ii) It is to be expected that a certain disagreement will more frequently occur according as the corresponding categories are more frequently used.

Therefore the measures of agreement kappa, upsilon and weighted kappa are defined in such a way that they equal zero in case of pure chance agreement and they equal one in case of perfect agreement.

Kappa is defined as the proportion of agreement in excess of what is to be expected by chance. An important property of kappa is expressed through the following interpretation of a new mathematical theorem: When the combining of categories results in a higher kappa value, then these categories are relatively hard to distinguish.

Upsilon is a good measure of agreement if the categories are ordered, on the understanding that we are willing to consider the assignment to a category as a quantitative judgement. Upsilon equals the correlation coefficient if the judgements by two observers have the same mean and the same standard deviation. Otherwise upsilon is smaller than the correlation coefficient. From a characteristic with a low upsilon value it is not to be expected that it will be strongly correlated with any other characteristic. Kappa equals upsilon if there are only two

categories.

In the second and third chapter the mutual agreement within a certain group of observers is investigated on the basis of the agreement between two random observers. This approach has the following advantages:

i) The numerical value of a measure of agreement is independent of the total number of observers.

ii) By generalizing in a direct way from two to more than two observers, a uniform statistical methodology arises: what is useful in the case of two observers is also useful in the case of many observers.

These two advantages imply a general applicability. Randomly missing judgements present no problems, and the number of observers per subject may vary.

The statistical procedures are illustrated within the context of some clinical diagnosis studies. In a group of seven pathologists, who separately judge 118 biopsy slides from the uterine cervix with respect to cancer, it is tried to discover one or more homogeneous subgroups of pathologists. From another example, where six physicians assess the pupil reaction to light in 28 coma patients, it is apparent that randomly missing judgements do not form an insurmountable problem. For a varying group of psychiatrists, who place 30 patients in five diagnostic categories, it appears that certain categories are relatively hard to distinguish.

Important information may also be obtained by investigating intraobserver agreement. For that purpose subjects are placed in categories a second time by the same observers. The resulting data can be analyzed using the statistical methods described in this book.

In all examples the standard jackknife technique is applied to indicate the stability of the conclusions that are drawn. In the fourth chapter the delta method and the bootstrap are also considered.

Chapter 1

AGREEMENT BETWEEN TWO FIXED OBSERVERS

1.1 The Problem

When two observers assign subjects to categories, they may have different opinions. In many situations, however, it is desirable that judgements are reproducible from one observer to another: When two observers judge the same subject, they should use the same category. This is the case, for example, when physicians assess a diagnosis or a symptom, and when coders classify answers to open interview questions. In such situations judgements have little value if there is little agreement between the observers, and it is important to measure the degree of agreement. After sources of disagreement have been detected, action may be taken to increase the degree of agreement.

In order to measure the agreement between two fixed observers, the observers separately assign each of a sample of subjects to one of L categories. Although agreement may be regarded as a special case of association, a measure of association is not necessarily a measure of agreement. Table 1.1-1 shows full association: If we know which category is used by the one observer, we also know which category is used by the other observer. But table 1.1-1 also shows complete disagreement: The main diagonal is empty.

TABLE 1.1-1
Hypothetical Frequencies Showing
Full Association but Complete Disagreement

Category by Observer 1	Category by Observer 2				Total
	1	2	3	4	
1	0	25	0	0	25
2	0	0	0	25	25
3	25	0	0	0	25
4	0	0	25	0	25
Total	25	25	25	25	100

Suppose the L-point scale is a quantitative scale, and observer 1 always scores one point higher on the scale than observer 2. Then there is perfect correlation but no agreement. So the correlation coefficient is not a good measure of agreement.

Now, suppose that observers 1 and 2 are not judging but gambling, just because they like gambling and they do not take their task seriously. They gamble according to the following chance mechanism. When they have to judge a subject, on a 3-point scale, observers 1 and 2 each throw a die. If the die shows 6 points, it is thrown again until it shows less than 6 points; this leaves the 6 out of it. If the die shows less

TABLE 1.1-2
Hypothetical Frequencies Showing Sixty-Six
Percent Agreement Between Observers 1 and 2

Category by Observer 1	Category by Observer 2			Total
	1	2	3	
1	1	8	1	10
2	8	64	8	80
3	1	8	1	10
Total	10	80	10	100

than 5 points, the subject is assigned to category 2; in eighty percent of the cases category 2 is used. If the die shows 5 points, heads or tails of a coin decides which of the categories 1 and 3 is used; each of the categories 1 and 3 is used in ten percent of the cases. The two judgements are statistically independent and table 1.1-2 contains the frequencies that are to be expected, on the basis of the marginal totals, when observers 1 and 2 have to judge a hundred subjects. There is sixty-six percent agreement: in sixty-six percent of the cases the two observers use the same category, merely by chance. There is, however, also sixty-six percent agreement between observers 3 and 4 in table 1.1-3 who take their task seriously. Observers 3 and 4 agree in a systematic way: would they have been gambling, then by table 1.1-4 only thirty-three percent agreement was to be expected. The last percentage is based on the marginal totals in table 1.1-3, assuming statistical independence. It is clear that the probability of agreement by chance, that is under statistical independence, is not the same for all pairs of observers and may be rather high. There is no doubt that chance agreement has to be taken into account. The percentage of agreement, or the proportion of agreement, is not a good measure of agreement.

TABLE 1.1-3
Hypothetical Frequencies Showing Sixty-Six Percent Agreement Between Observers 3 and 4

Category by Observer 3	Category by Observer 4			Total
	1	2	3	
1	24	13	3	40
2	5	20	5	30
3	1	7	22	30
Total	30	40	30	100

From tables 1.1-2 and 1.1-3 still another conclusion can be

drawn if it is assumed that the 3-point scale is a quantitative scale. Since observers 1 and 2 use the same chance mechanism, they have the same mean score, as can be deduced easily from the marginal totals in table 1.1-2, but their scores have no meaning. From the marginal totals in table 1.1-3, however, it is clear (without computation) that observers 3 and 4 have different means, although their judgements may be meaningful. The conclusion is unavoidable: it is not sufficient to look at means, and it is also not sufficient to consider only the univariate distributions of the individual observers.

TABLE 1.1-4
Expected Frequencies Showing Thirty-Three Percent Agreement Between Observers 3 and 4 When These Observers Would Have Been Gambling

Category by Observer 3	Category by Observer 4			Total
	1	2	3	
1	12	16	12	40
2	9	12	9	30
3	9	12	9	30
Total	30	40	30	100

Bibliographic Notes, 1.1

Cohen(1960) explained that the agreement between two observers often is investigated in a misleading way. He offered an alternative that is considered in the next section; see also the lucid review by Bartko and Carpenter(1967). Rogot and Goldberg(1966) stressed that chance agreement must be taken into account.

Spodick(1975), Koran(1975, 1976), Helzer et al.(1977), the Department of Clinical Epidemiology and Biostatistics in the McMaster University(1980) and Wulff(1981) argue that it is important to measure interobserver agreement. They give many examples of a disappointingly low degree of agreement among observers.

1.2 A Measure of Agreement

Suppose we wish to measure the agreement between two observers who separately classify each of a random sample of N subjects into one of L categories. The two observers are fixed, that is all subjects are judged by the same two observers. The case of varying observers, where the observers judging one subject are not necessarily the same as those judging another subject, is treated in section 3.3.

As an example we consider a study that was designed to investigate the observer variability in the histological classification of carcinoma in situ and related lesions of the uterine cervix. Seven pathologists separately classified N=118 biopsy slides into one of the following L=5 categories based on the most involved lesion:

Category 1: Negative
Category 2: Atypical Squamous Hyperplasia
Category 3: Carcinoma in Situ
Category 4: Squamous Carcinoma with Early Stromal Invasion
Category 5: Invasive Carcinoma

In the next chapter the agreement among all seven pathologists will be investigated, but in the present chapter only pathologists 1 and 2 are considered. Holmquist, McMahan and Williams(1967) described the study and presented the data.

Table 1.2-1 shows the observed frequencies f(i,j) of biopsy slides assigned to category i by pathologist 1 and to category j by pathologist 2; f(i,+) denotes the total number of slides assigned to category i by pathologist 1, and f(+,j) denotes the total number of slides assigned to category j by pathologist 2. The observed proportion of agreement between both pathologists is

$$o = \frac{1}{N} \sum_{i=1}^{L} f(i,i) \qquad (1.2\text{-}1)$$

$$= (22 + 7 + 36 + 7 + 3)/118 = .64 .$$

TABLE 1.2-1
Observed Frequencies f(i,j) and Expected Frequencies g(i,j) of Biopsy Slides Classified by Two Pathologists According to Most Involved Histological Lesion of the Uterine Cervix

Category by Pathologist 1		Category by Pathologist 2 j=1	j=2	j=3	j=4	j=5	Total f(i,+)
i=1	Observed	22	2	2	0	0	26
	Expected	5.9	2.6	15.2	1.5	0.7	
i=2	Observed	5	7	14	0	0	26
	Expected	5.9	2.6	15.2	1.5	0.7	
i=3	Observed	0	2	36	0	0	38
	Expected	8.7	3.9	22.2	2.3	1.0	
i=4	Observed	0	1	14	7	0	22
	Expected	5.0	2.2	12.9	1.3	0.6	
i=5	Observed	0	0	3	0	3	6
	Expected	1.4	0.6	3.5	0.4	0.2	
Total	f(+,j)	27	12	69	7	3	N=118

Category 1: Negative
Category 2: Atypical Squamous Hyperplasia
Category 3: Carcinoma in Situ
Category 4: Squamous Carcinoma with Early Stromal Invasion
Category 5: Invasive Carcinoma

If the assignments by the two pathologists would be independently distributed, these assignments would have no value: with respect to the probability of agreement, under statistical independence it does not matter whether the two pathologists judge the same slides or different slides. Therefore it is necessary to investigate if the observed frequencies f(i,j) substantially differ from the frequencies that are to be expected under the null hypothesis of independence, especially regarding the main diagonal where there is perfect agreement. Under the null hypothesis of independence

$$g(i,j) = f(i,+)f(+,j)/N \qquad (1.2\text{-}2)$$

is an estimate of the frequency of biopsy slides expected to be assigned to category i by pathologist 1 and to category j by pathologist 2, based on the marginal frequencies f(i,+) and f(+,j). An estimate of the proportion of agreement that is to be expected under independence is

$$e = \frac{1}{N} \sum_{i=1}^{L} g(i,i) \qquad (1.2\text{-}3)$$

$$= (5.9 + 2.6 + 22.2 + 1.3 + 0.2)/118 = .27 \ ,$$

which may be interpreted as the proportion of agreement that is to be expected solely on the basis of chance when judgements do not depend on the judged subject. It may also be interpreted as the proportion of agreement that is to be expected when pathologists 1 and 2 do not classify the same but different biopsy slides.

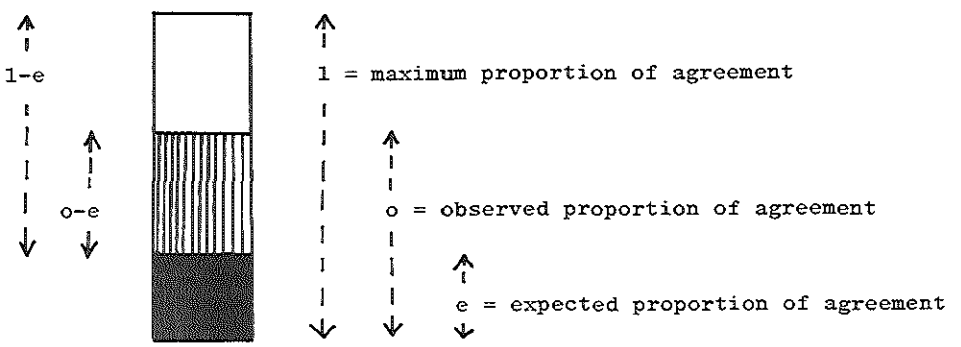

FIGURE 1.2-1 Interpretation of kappa = $\frac{o-e}{1-e}$

The difference o - e represents the proportion of agreement in excess of what is to be expected under independence, and 1 - e represents the maximum possible excess. Thus the coefficient kappa, defined by

$$k = \text{kappa} = \frac{o - e}{1 - e} \ , \qquad (1.2\text{-}4)$$

represents the proportional excess beyond what is to be expected under independence. In our example kappa = .50; see figure 1.2-1. Due to rounding errors in o = .64 and e = .27 the kappa value is not exactly equal to (.64 - .27)/(1.00 - .27). Kappa equals zero if the observed proportion of agreement is just what is to be expected under independence, and kappa equals one if the observers always perfectly agree. Cohen(1960) introduced the coefficient kappa as a modification of the coefficient alpha proposed by Scott(1955).

By way of exception it may occur that only one category is used by the two observers, e.g. f(1,1) = N. In such a case 1 - e = 0 and kappa cannot be computed. The argumentation below shows that we must accept the fact that kappa sometimes cannot be determined. Suppose that f(1,1) = N = 100 and that one more subject will be judged. Then it may happen that f(2,2) becomes equal to 1, in which case kappa = 1 represents perfect agreement. But it may also happen that f(1,2) becomes equal to 1, in which case kappa = 0 represents merely chance agreement. When only one category is used by the two observers, it is virtually impossible to make any reasonable prediction about the value of kappa in other samples, and so the degree of agreement is completely unknown. Of course, judgements do not yield any information about subjects if *always* all subjects will be assigned to the same category and a different category will *never* be used.

In view of generalizations to the case of more than two observers in the next chapters, it is useful to derive from the frequencies in table 1.2-1 the corresponding proportions shown in table 1.2-2. The notation used in table 1.2-2 is presented in table 1.2-3 and explained below. The observed proportion p(i,j) = f(i,j)/N is an unbiased estimate of the probability that a randomly selected subject is assigned to category i by observer 1 and to category j by observer 2. The marginal proportion $m_1(i) = f(i,+)/N = \sum_j p(i,j)$ is an unbiased estimate of the probability that a randomly selected subject is assigned to category i by observer 1, and $m_2(j) = f(+,j)/N = \sum_i p(i,j)$ is an unbiased estimate of the probability that category j is used by observer 2. The expected proportion

TABLE 1.2-2

Observed and Expected (Under Independence) Proportions of 118 Biopsy Slides Classified by Two Pathologists According to Most Involved Histological Lesion of the Uterine Cervix

Category by Pathologist 1		Category by Pathologist 2 1	2	3	4	5	Total
1	Observed	.19	.02	.02	.00	.00	.22
	Expected	.05	.02	.13	.01	.01	
2	Observed	.04	.06	.12	.00	.00	.22
	Expected	.05	.02	.13	.01	.01	
3	Observed	.00	.02	.31	.00	.00	.32
	Expected	.07	.03	.19	.02	.01	
4	Observed	.00	.01	.12	.06	.00	.19
	Expected	.04	.02	.11	.01	.00	
5	Observed	.00	.00	.03	.00	.03	.05
	Expected	.01	.01	.03	.00	.00	
Total		.23	.10	.58	.06	.03	1.00
	$p_{1\|2}$	.81	.58	.52	1.00	1.00	
	$p_{2\|1}$	.85	.27	.95	.32	.50	

Category 1: Negative
Category 2: Atypical Squamous Hyperplasia
Category 3: Carcinoma in Situ
Category 4: Squamous Carcinoma with Early Stromal Invasion
Category 5: Invasive Carcinoma

$$q(i,j) = m_1(i)m_2(j) , \qquad (1.2\text{-}5)$$

which equals $g(i,j)/N$, is an estimate of the probability under independence that a randomly selected subject is assigned to category i by observer 1 and to category j by observer 2. The observed proportion of agreement

TABLE 1.2-3
Format Used to Show Observed Proportions p(i,j), Expected (Under Independence) Proportions q(i,j), Marginal Proportions $m_1(i)$ and $m_2(j)$, and Conditional Proportions $p_{1|2}(j|j)$ and $p_{2|1}(i|i)$ When the Two Observers Used Three Categories: L=3

Category by Observer 1		Category by Observer 2 j=1	j=2	j=3	Total
i=1	Observed	p(1,1)	p(1,2)	p(1,3)	$m_1(1)$
	Expected	q(1,1)	q(1,2)	q(1,3)	
i=2	Observed	p(2,1)	p(2,2)	p(2,3)	$m_1(2)$
	Expected	q(2,1)	q(2,2)	q(2,3)	
i=3	Observed	p(3,1)	p(3,2)	p(3,3)	$m_1(3)$
	Expected	q(3,1)	q(3,2)	q(3,3)	
Total		$m_2(1)$	$m_2(2)$	$m_2(3)$	1.00
$p_{1\|2}$		$p_{1\|2}(1\|1)$	$p_{1\|2}(2\|2)$	$p_{1\|2}(3\|3)=p(3,3)/m_2(3)$	
$p_{2\|1}$		$p_{2\|1}(1\|1)$	$p_{2\|1}(2\|2)$	$p_{2\|1}(3\|3)=p(3,3)/m_1(3)$	

$$o = \sum_{i=1}^{L} p(i,i) \qquad (1.2\text{-}6)$$

is an unbiased estimate of the probability that a randomly selected subject is assigned to the same category by both observers, and the expected proportion of agreement

$$e = \sum_{i=1}^{L} q(i,i) \qquad (1.2\text{-}7)$$

is an estimate of the probability that the two observers agree by chance. For the sake of clarity: the expected proportions q(i,j) and e are conditional expectations given the marginal proportions $m_1(i)$ and $m_2(j)$. Considered unconditionally, the expected proportions q(i,j) and e are only estimates of expectations under complete independence, because

the margins are not fixed in advance.

On the diagonal in table 1.2-2 the observed proportions are larger than is to be expected under independence, but the two pathologists often differed one point on the five-point scale. Pathologist 2 assigned more than half of the biopsy slides to category 3.

From the slides assigned to category i by pathologist 1 a proportion

$$p_{2|1}(i|i) = p(i,i)/m_1(i) \qquad (1.2\text{-}8)$$

was also assigned to that category by pathologist 2. These conditional proportions are shown in the last row of table 1.2-2. Given that a slide was assigned to category i by pathologist 1, $p_{2|1}(i|i)$ is an estimate of the conditional probability that pathologist 2 assigns that slide to the same category. The conditional proportion

$$p_{1|2}(i|i) = p(i,i)/m_2(i) \qquad (1.2\text{-}9)$$

has an analogous interpretation. From the slides assigned to category 3 by pathologist 2 only a proportion .52 was assigned to that same category by pathologist 1; see the last row but one in table 1.2-2. Five of the ten conditional proportions are disappointingly low.

There is still another way to derive kappa. Since o is the observed proportion of agreement, d = 1 - o is the observed proportion of disagreement. The estimate c = 1 - e of the expected proportion of disagreement under independence may be regarded as the possible reduction in disagreement that could be obtained, and c - d may be regarded as the actual reduction in disagreement that is obtained. Thus the coefficient kappa, rewritten as

$$k = \text{kappa} = \frac{c - d}{c} \quad , \qquad (1.2\text{-}10)$$

may be regarded as the proportional reduction in disagreement, compared to the proportion of disagreement expected under independence.

Bibliographic Notes, 1.2

Holmquist, McMahan and Williams(1967) presented the data in table 1.2-1 and gave a detailed description of the observer variability study.

Cohen(1960) introduced the coefficient kappa and discussed its properties. He remarked that, given the marginal distributions, the maximum possible value of kappa can be substantially less than one if the two marginal distributions are substantially different. This is a desirable property of kappa because different marginal distributions are to be considered a source of disagreement; see also section 1.5.

Popping(1983b, 1985) convincingly showed that, with respect to the measurement of interobserver agreement, kappa statistics have more desirable properties than other coefficients. One of the properties of kappa he mentioned is that perfect agreement is a transitive relationship: if observer 1 perfectly agrees with both observers 2 and 3, then observers 2 and 3 perfectly agree with one another.

The estimation procedure proposed by Scott(1955) is appropriate for the case of two varying observers, where subjects are judged by different pairs of observers; see also section 3.3.

It is also possible to measure the agreement between two observers who are not using the same nominal or ordinal scale, and who possibly use different numbers of categories. For that situation the interested reader is referred to Popping(1983a, 1984).

Landis and Koch(1977a) and Fleiss(1981, section 13.1) characterized different ranges of kappa values, with respect to the degree of agreement they suggest, by assigning labels to these ranges. At the end of his review Fleiss(1981, section 13.3) suggested many useful applications of kappa beyond the measurement of interobserver agreement.

1.3 *Combining Categories that are Easily Confused*

An important source of disagreement may be detected by investigating which categories are easily confused. When two particular categories are frequently used, however, it is to be expected that disagreements concerning these two categories frequently occur. This is taken into account when kappa is used, as the theorem below shows. The theorem is discussed after its proof. An interesting application is presented in section 3.3.

Theorem

Combining categories i and j increases kappa if and only if

$$\frac{p(i,j) + p(j,i)}{q(i,j) + q(j,i)} > \frac{1 - o}{1 - e} = 1 - \text{kappa} \qquad (1.3\text{-}1)$$

Notice that p(i,j) + p(j,i) is the observed proportion of disagreement among the combined categories, whereas 1 - o is the observed proportion of disagreement among all categories. The denominators contain the corresponding expected proportions of disagreement.

Proof

Let o and e denote the observed and expected proportion of agreement before categories are combined. After the two categories i and j have been combined, the observed proportion of agreement has become $o_c = o + p(i,j) + p(j,i)$, and the expected proportion of agreement has become $e_c = e + q(i,j) + q(j,i)$; here p(i,j), p(j,i), q(i,j) and q(j,i) denote proportions before categories are combined. It is convenient to define $\Delta_o = o_c - o = p(i,j) + p(j,i)$ and $\Delta_e = e_c - e = q(i,j) + q(j,i)$. When kappa_c denotes the kappa value after categories i and j have been combined, it is easy to see that

$$1 - \text{kappa}_c = \frac{1 - o_c}{1 - e_c} = \frac{1-o - \Delta_o}{1-e - \Delta_e}$$

The theorem now follows from the equivalent inequalities below.

$$\begin{aligned} \text{kappa}_c &> \text{kappa} \\ (1-e)(1-e-\Delta_e)(1-\text{kappa}_c) &< (1-\text{kappa})(1-e)(1-e-\Delta_e) \\ -\Delta_o(1-e) &< -\Delta_e(1-o) \end{aligned}$$

$$\frac{p(i,j)+p(j,i)}{q(i,j)+q(j,i)} = \frac{\Delta_o}{\Delta_e} > \frac{1-o}{1-e} \qquad \text{Q.E.D.}$$

From the proof it is easy to see how the theorem can be generalized to the case of combining more than two categories. Kappa is increased if and only if the ratio of observed to expected disagreement among the combined categories is larger than the ratio of observed to expected disagreement among all categories. When the ratio of observed to expected disagreement among certain categories is relatively high, the interpretation is that these categories are relatively hard to distinguish. So kappa is increased by combining certain categories if and only if these categories are relatively hard to distinguish. This property of kappa is used in section 3.3 to identify such categories.

When the L categories are ordered, as is the case in our example, most disagreements will be near the diagonal. Then, because of the above theorem, it is to be expected that kappa will be increased if the number of categories is decreased. This implies that kappa values from different samples are not quite comparable if the number of categories is not constant.

In our example categories 1 and 2, representing absence of carcinoma, are considered less serious clinically than categories 3, 4 and 5, representing presence of carcinoma. If the final diagnosis for a patient is characterized by categories 3, 4 or 5, this cancer patient should be operated. Therefore it is meaningful to consider categories 1 and 2 combined versus categories 3, 4 and 5 combined; see table 1.3-1. Given that pathologist 1 thinks there is no carcinoma, there is an estimated probability .69 that pathologist 2 also thinks there is no carcinoma. The

kappa value .66 for the two-point scale is higher than the kappa value .50 for the five-point scale. Of course, this certainly does not imply that the five-point scale should not be used. The example just illustrates that the kappa value depends on the number of categories used.

TABLE 1.3-1
Observed and Expected (Under Independence) Proportions of 118 Biopsy Slides Classified by Two Pathologists According to Presence(+) or Absence(-) of Carcinoma

		Pathologist 2		
Pathologist 1		-	+	Total
-	Observed	.31	.14	.44
	Expected	.15	.30	
+	Observed	.03	.53	.56
	Expected	.18	.37	
Total		.33	.67	1.00
$p_{1\|2}$		.92	.80	
$p_{2\|1}$		.69	.95	

Bibliographic Notes, 1.3

The theorem is new, but one may find an indication in the paper by Cohen(1968, p.216) where he discussed in what situations weighted kappa is larger than unweighted kappa.

1.4 Agreement on a Particular Category

The intraclass kappa coefficient, which is a special case of kappa, may be used to measure the agreement on a particular category. A high(low) intraclass kappa value k(i) means that the observers clearly(hardly) distinguish between category i on the one hand and the remaining L-1 categories on the other hand.

Consider a subject who is assigned to category r by the first observer and to category s by the second observer. There is agreement on category i if r=s=i. But there is also agreement on category i if r≠i and s≠i, even if r≠s. The observed proportion of agreement

$$o(i) = p(i,i) + \sum_{\substack{r=1 \\ r\neq i}}^{L} \sum_{\substack{s=1 \\ s\neq i}}^{L} p(r,s) = 1 - \sum_{\substack{j=1 \\ j\neq i}}^{L} (p(i,j) + p(j,i)) \qquad (1.4\text{-}1)$$

is an unbiased estimate of the probability that a randomly selected subject is assigned to category i by both observers or is assigned to the remaining L-1 categories by both observers. The estimate of the corresponding probability under independence is the expected proportion of agreement

$$e(i) = 1 - \sum_{\substack{j=1 \\ j\neq i}}^{L} (q(i,j) + q(j,i)) \qquad (1.4\text{-}2)$$

The observed and expected proportions of disagreement can be written as

$$d(i) = 1 - o(i) = m_1(i) + m_2(i) - 2p(i,i) \qquad (1.4\text{-}3)$$

$$c(i) = 1 - e(i) = m_1(i) + m_2(i) - 2q(i,i) \qquad (1.4\text{-}4)$$

The intraclass kappa coefficient

$$k(i) = \frac{o(i) - e(i)}{1 - e(i)} = \frac{c(i) - d(i)}{c(i)} \qquad (1.4\text{-}5)$$

is a chance corrected measure of agreement on category i. Of course,

when L=2 both intraclass kappas are equal to kappa.

The coefficient kappa can be written as a weighted average of the intraclass kappa coefficients:

$$k = \text{kappa} = \frac{\sum_{i=1}^{L} c(i)k(i)}{\sum_{i=1}^{L} c(i)} \qquad (1.4\text{-}6)$$

This implies that kappa lies between the smallest and the largest intraclass kappa. In our example kappa = .50, while k(1) = .78, k(2) = .27, k(3) = .44, k(4) = .43 and k(5) = .65. Since the five categories are ordered, the result that the first and last intraclass kappas are higher than the other ones is plausible.

Bibliographic Notes, 1.4

In essence, the formulation in this section has been given by Fleiss(1981, section 13.1). However, many other proposals have been published in order to measure the agreement on a particular category; see e.g. Fleiss(1971), Bishop, Fienberg and Holland(1975, section 11.4), Landis and Koch(1977c), Cicchetti, Lee, Fontana and Dowds(1978) and Schouten (1980).

For i≠j Schouten(1982a) defined the interclass kappa coefficient k(i,j) = (q(i,j) - p(i,j))/q(i,j) as a proportional reduction in disagreement. A high(low) interclass kappa value k(i,j) means that the observers clearly (hardly) distinguish between the two categories i and j. The coefficient k(i,j) has been described earlier by Hildebrand, Laing and Rosenthal (1977), but in a different context and under a different name. Landis and Koch(1977c), Schouten(1980) and Holman et al.(1982) proposed other interclass measures of agreement. My present opinion, however, is that the p(i,j) and q(i,j) should be inspected, rather than the k(i,j), because the proportions p(i,j) and q(i,j) are easier to understand; see also the theorem in section 1.3. This holds equally well when there are more than two observers.

1.5 Upsilon

In this section the categories are assumed to be ordered and equidistant; the L-point scale is treated as a quantitative scale. When a subject is assigned to category i by the first and to category j by the second observer, the degree of disagreement between both observers may be represented by the squared difference $(i - j)^2$. In the sample of N subjects the mean squared difference is

$$d(u) = \sum_{i=1}^{L} \sum_{j=1}^{L} p(i,j)(i - j)^2 \qquad (1.5\text{-}1)$$

When the judgements by both observers are independently distributed, the chance squared difference

$$c(u) = \sum_{i=1}^{L} \sum_{j=1}^{L} q(i,j)(i - j)^2 \qquad (1.5\text{-}2)$$

is an estimate of the expected squared difference. Regarding c(u) as the possible reduction in squared difference that could be obtained, and c(u) - d(u) as the actual reduction in squared difference that is obtained, the coefficient

$$u = \text{upsilon} = \frac{c(u) - d(u)}{c(u)} \qquad (1.5\text{-}3)$$

is the proportional reduction in squared difference, compared to the chance squared difference. The coefficient upsilon is a measure of agreement between two observers: u = 1 if the observers never have different opinions, and u = 0 (apart from sampling fluctuations) if both judgements are independently distributed. In our example we have d(u) = .52, c(u) = 2.33 and upsilon = .78. The two pathologists disagree much less than is to be expected by chance. When there are only two categories, upsilon equals kappa.

The mean and variance of the judgements y_a by observer a (a = 1,2) may be estimated by

$$\bar{y}_a = \sum_{i=1}^{L} m_a(i)i \qquad (1.5\text{-}4)$$

$$\mathrm{var}(y_a) = \sum_{i=1}^{L} m_a(i)(i - \bar{y}_a)^2 \qquad (1.5\text{-}5)$$

The covariance of both judgements may be estimated by

$$\mathrm{cov}(y_1,y_2) = \sum_{i=1}^{L}\sum_{j=1}^{L} p(i,j)(i - \bar{y}_1)(j - \bar{y}_2) \qquad (1.5\text{-}6)$$

Note in passing that the estimates in (1.5-5) and (1.5-6) are biased. Now it can be shown that the mean squared difference, the chance squared difference and upsilon can be rewritten as

$$d(u) = \mathrm{var}(y_1) + \mathrm{var}(y_2) - 2\mathrm{cov}(y_1,y_2) + (\bar{y}_1 - \bar{y}_2)^2 \qquad (1.5\text{-}7)$$

$$c(u) = \mathrm{var}(y_1) + \mathrm{var}(y_2) + (\bar{y}_1 - \bar{y}_2)^2 \qquad (1.5\text{-}8)$$

$$\text{upsilon} = \frac{2\mathrm{cov}(y_1,y_2)}{\mathrm{var}(y_1) + \mathrm{var}(y_2) + (\bar{y}_1 - \bar{y}_2)^2} \qquad (1.5\text{-}9)$$

Since $\mathrm{var}(y_1) + \mathrm{var}(y_2) \geq 2\sqrt{(\mathrm{var}(y_1)\mathrm{var}(y_2))}$, which is easily proved by squaring, upsilon is less than or equal to the Pearson coefficient of correlation between the quantitative judgements by the two observers, provided that the covariance is positive. In the case of equal means and equal variances upsilon equals the Pearson coefficient of correlation.

The formulas above also hold if each estimate is replaced by the population parameter that is estimated. If the biased estimates $\mathrm{var}(y_a)$ and $\mathrm{cov}(y_1,y_2)$ in formulas (1.5-5) and (1.5-6) are replaced by the corresponding unbiased estimates, which are $N/(N-1)$ times larger, then formulas (1.5-7) to (1.5-9) inclusive only approximately hold. Regarding the mathematical notation, above var(.) and cov(.,.) have been used to denote the sample variance and covariance. In section 1.9 and in chapter 4, however, a separate notation for the sample variance and the population variance is needed: $s^2(.)$ and $\sigma^2(.)$.

In section 3.4 it is shown that the Pearson coefficient of correlation between two different random variables y and y' is bounded by the upsilon values of y and y'; see also the end of section 1.8.

The coefficient upsilon may also prove its usefulness in the case of continuous measurements. When y_{ha} denotes the measurement on subject h (h = 1,2,...,N) by observer a (a = 1,2), d(u) and c(u) can be rewritten as

$$d(u) = \frac{1}{N} \sum_{h=1}^{N} (y_{h1} - y_{h2})^2 \qquad (1.5\text{-}10)$$

$$c(u) = \frac{1}{N^2} \sum_{g=1}^{N} \sum_{h=1}^{N} (y_{g1} - y_{h2})^2 \qquad (1.5\text{-}11)$$

Now the mean and variance of the judgements by observer a (a = 1,2) may be estimated by

$$\bar{y}_a = \frac{1}{N} \sum_{h=1}^{N} y_{ha} \qquad (1.5\text{-}12)$$

$$var(y_a) = \frac{1}{N} \sum_{h=1}^{N} (y_{ha} - \bar{y}_a)^2 \qquad (1.5\text{-}13)$$

and the covariance between the judgements by the two observers may be estimated by

$$cov(y_1,y_2) = \frac{1}{N} \sum_{h=1}^{N} (y_{h1} - \bar{y}_1)(y_{h2} - \bar{y}_2) \qquad (1.5\text{-}14)$$

Formulas (1.5-12) to (1.5-14) inclusive are equivalent to formulas (1.5-4) to (1.5-6) inclusive, and formulas (1.5-7) to (1.5-9) inclusive are valid.

Bibliographic Notes, 1.5

Until now upsilon was called weighted kappa using quadratic weights; see the next section. In my opinion a new name is necessary because many non-statisticians tend to confuse kappa and weighted kappa, like they tend to confuse the Pearson correlation coefficient and intraclass correlation coefficients.

The coefficient upsilon is defined without making analysis of variance assumptions. Unlike many intraclass correlation coefficients, the coefficient upsilon really is a measure of agreement. For large N, however, Fleiss and Cohen(1973) proved the approximate equivalence of upsilon and one of the many intraclass correlation coefficients, analogous to formula (1.5-9); see also Krippendorff(1970) and Cohen(1960, 1968). In unbalanced designs, however, as considered in chapter 3, there may be a great difference between the sample value of upsilon and the sample value of that intraclass correlation coefficient; so the two statistics should not be confused.

1.6 Weighted Kappa

When a subject is assigned to category i by the first and to category j by the second observer, the degree of disagreement between both observers may be represented by a disagreement weight v(i,j), where

$$v(i,i) = 0 \leq v(i,j) = v(j,i) \ , \qquad (1.6\text{-}1)$$

even if the L-point scale is a nominal scale. The special case $v(i,j) = (i - j)^2$ was discussed in section 1.5, where the L-point scale was treated as an interval scale. When the disagreement weights are averaged over the N subjects, the observed degree of disagreement

$$d(v) = \sum_{i=1}^{L} \sum_{j=1}^{L} p(i,j)v(i,j) \qquad (1.6\text{-}2)$$

is obtained. The chance degree of disagreement

$$c(v) = \sum_{i=1}^{L} \sum_{j=1}^{L} q(i,j)v(i,j) \qquad (1.6\text{-}3)$$

is an estimate of the degree of disagreement that is to be expected by chance when both judgements are independently distributed.

The chance disagreement c(v) may be regarded as the possible reduction in disagreement that could be obtained, and c(v) - d(v) may be regarded as the actual reduction in disagreement that is obtained. Thus the coefficient

$$\text{weighted kappa} = \frac{c(v) - d(v)}{c(v)} \qquad (1.6\text{-}4)$$

may be regarded as the proportional reduction in disagreement, compared to the chance disagreement. Weighted kappa varies from negative values for more than chance disagreement, through zero for chance disagreement, to one when there is no disagreement at all. Kappa and upsilon are special cases of weighted kappa. Weighted kappa equals kappa if $v(i,j) = 1$ for $i \neq j$, and weighted kappa equals upsilon if $v(i,j) = (i - j)^2$.

Cohen(1968) introduced weighted kappa as a generalization of kappa.

Let v(max) denote the largest disagreement weight. Since v(i,j) represents the degree of disagreement between the categories i and j, v(i,j)/v(max) may be regarded as the proportion of disagreement between these categories, and the agreement weight

$$w(i,j) = 1 - \frac{v(i,j)}{v(max)} \qquad (1.6\text{-}5)$$

may be regarded as the proportion of agreement between these categories;

$$0 \leq w(i,j) = w(j,i) \leq 1 = w(i,i) \ . \qquad (1.6\text{-}6)$$

Of course, agreement weights can be defined without defining disagreement weights.

With reference to the agreement weights w(i,j) the observed degree of agreement is

$$o(w) = \sum_{i=1}^{L} \sum_{j=1}^{L} p(i,j)w(i,j) \qquad (1.6\text{-}7)$$

The chance degree of agreement

$$e(w) = \sum_{i=1}^{L} \sum_{j=1}^{L} q(i,j)w(i,j) \qquad (1.6\text{-}8)$$

is an estimate of the degree of agreement that is to be expected by chance when the judgements by the two observers are independently distributed. The difference o(w) - e(w) represents the excess agreement beyond chance, and 1 - e(w) represents the maximum possible excess agreement beyond chance. Thus the coefficient

$$k(w) = \text{weighted kappa} = \frac{o(w) - e(w)}{1 - e(w)} \qquad (1.6\text{-}9)$$

represents the proportional excess agreement beyond the degree of agreement that is to be expected under independence. Weighted kappa varies from negative values for less than chance agreement, through

zero for chance agreement, to one for perfect agreement. It is easy to see that $k(w) = (c(v) - d(v))/c(v)$ if (1.6-5) holds.

The coefficient weighted kappa is a very general measure of agreement, and the coefficients considered in sections 1.2 to 1.5 inclusive are special cases of weighted kappa. It is easy to see that the weighted kappa coefficient $k(w)$ equals the intraclass kappa coefficient $k(2)$ if the first set of agreement weights in table 1.6-1 is used. If the second set of weights is used, the weighted kappa value equals the kappa value corresponding to the condensed table 1.3-1.

TABLE 1.6-1
Two Sets of Agreement Weights

Category 2 Versus Categories 1, 3, 4 and 5 Combined

Category	1	2	3	4	5
1	1	0	1	1	1
2	0	1	0	0	0
3	1	0	1	1	1
4	1	0	1	1	1
5	1	0	1	1	1

Categories 1 and 2 Combined Versus Categories 3, 4 and 5 Combined

Category	1	2	3	4	5
1	1	1	0	0	0
2	1	1	0	0	0
3	0	0	1	1	1
4	0	0	1	1	1
5	0	0	1	1	1

Bibliographic Notes, 1.6

Cohen(1968) introduced the coefficient weighted kappa and discussed its main properties.

In case of an ordinal scale Hall(1974) preferred to use the linear disagreement weights $v(i,j) = |i - j|$, thus considering the disagreement between categories 2 and 4 twice as serious as the disagreement between categories 3 and 4. When using upsilon, the disagreement between categories 2 and 4 is considered four times as serious as the disagreement between categories 3 and 4. When using kappa, all disagreements are considered equally serious. Because of the relations between upsilon and the correlation coefficient, see sections 1.5 and 3.4, I slightly prefer upsilon, but I have to admit that it is a matter of taste. However, if there are good reasons to choose certain weights, or if the investigator strongly prefers certain weights, I am the last man to protest.

1.7 Missing Judgements

For the sake of simplicity, the discussion in this section is focussed on the statistic kappa, but it also applies to other agreement statistics. The discussion is not confined to the case of only two observers: subjects may be judged by two or more observers.

The statistic kappa especially proves its usefulness when two or more kappa values are compared with one another; see sections 2.1 to 2.4 inclusive and 3.1 to 3.3 inclusive. In many situations missing judgements will not cause a substantial bias in kappa values. When missing judgements do cause a substantial bias, however, there is a problem, especially if the kappa values are not biased to the same extent. Three different situations are to be distinguished:

i) Judgements are missing by chance
With regard to a certain observer, assume that the unclassified subjects do not systematically differ from the classified subjects. When this assumption holds for all observers, computations can be based on the subjects that are judged by at least two observers. The example in sections 3.1 and 3.2 concerns six fixed observers, but no subject is judged by all six observers.

ii) Subjects may be assigned to the "Other" category
When some subjects are not classified by an observer because this observer thinks these subjects do not belong to one of the L categories, just create a new category L+1: the "Other" category. A subject is assigned to category L+1 if the observer thinks the subject does not belong to one of the categories 1 to L inclusive. Now o, e and kappa can be computed from an L+1 by L+1 table of observed and expected proportions, treating category L+1 in exactly the same way as the categories 1 to L inclusive. It is also possible to treat the assignments to category L+1 as if they were missing by chance, see under *(i)*, but that means a loss of information. Section 3.3 contains an example.

iii) Some subjects are hard to classify

The situation becomes more complicated when a subject is not classified because this subject is hard to classify, although the observer thinks the subject belongs to one of the L categories. When there are more than two observers, and a table of kappa values is to be inspected (as is done in sections 2.1 and 3.1), there may be a problem if kappa is biased upwards for some pairs of observers. When the effect of training on degree of agreement is to be assessed, the observed effect may have a bias if the number of missing assignments is much larger before than after the training. When two groups of observers are to be compared, the two kappas may be biased to a different extent if many more assignments are missing in the one group than in the other. If possible, this situation must be avoided. In my opinion this problem has no satisfactory solution. This situation may be treated as indicated under *(i)* or *(ii)*. Another approach is as follows.

Suppose M of the N subjects have been classified by both of two observers (at the end of this section the case of more than two observers is considered). For these M subjects o denotes the observed and e denotes the chance proportion of agreement. Suppose that, when the remaining N-M subjects also had been classified by both observers, for these N-M subjects the observed proportion of agreement would have been equal to the chance proportion of agreement e. This leads to an adjusted observed proportion of agreement

$$o_{adj.} = \frac{Mo + (N-M)e}{N} \qquad (1.7\text{-}1)$$

and an adjusted kappa value

$$\text{kappa}_{adj.} = \frac{o_{adj.} - e}{1 - e} = \frac{M}{N}\,\text{kappa} \qquad (1.7\text{-}2)$$

Since the assumption above may be somewhat pessimistic, the last formula must be considered a lower bound.

When there are more than two fixed observers, the most right expression in formula (2.2-7) in the next chapter suggests how to

proceed. In the case of varying observers kappa may be multiplied by the ratio of actual number of assignments and maximum possible number of assignments, analogous to formula (1.7-2).

Bibliographic Notes, 1.7

Fleiss(1971) gave an example where the "Other" category has been used. Kraemer(1980) concluded that, regarding that example, the approaches under *(i)* and *(ii)* do not lead to substantially different results. The example is reanalyzed in section 3.3.

Formulas (1.7-1) and (1.7-2) have not been published before. In 1982 I made a different proposal, see Schouten(1982b, section 7), but I am no longer behind that proposal. It has been based on the irrational assumption that there would have been complete disagreement on the N-M subjects not classified by both observers.

1.8 Influence of Population Characteristics

In this section it is demonstrated that kappa not only depends on the judging observers, but also depends on the characteristics of the population of subjects.

Consider a population of individuals who are diagnosed by two physicians A and B regarding presence or absence of a certain disease. Table 1.8-1 shows the agreement between both physicians, separately for individuals with and without the disease. For each physician the specificity is .90: if an individual actually does not have the disease, there is a probability .90 that the physician also declares that the individual does not have the disease. For each physician the sensitivity is also .90: if an individual actually has the disease, there is a probability .90 that the physician also declares that the individual has the disease. In ten percent of the cases the wrong diagnosis is assessed.

TABLE 1.8-1
Probabilities of Diagnoses by Physicians A and B
According to Presence(+) or Absence(-) of the disease

Individuals Without the Disease				Individuals With the Disease			
	B				B		
A	-	+	Total	A	-	+	Total
-	.84	.06	.90	-	.04	.06	.10
+	.06	.04	.10	+	.06	.84	.90
Total	.90	.10	1.00	Total	.10	.90	1.00
Specificity = .90				Sensitivity = .90			

The coefficient kappa depends on the prevalence of the disease in the population of individuals; the prevalence is defined as the proportion of individuals who actually have the disease. Table 1.8-2 shows the agreement between both physicians when the prevalence is

TABLE 1.8-2
Probabilities of Diagnoses by Physicians A and B
According to Presence(+) or Absence(-) of the Disease

All Individuals
Prevalence = .10

	B		
A	-	+	Total
-	.76	.06	.82
+	.06	.12	.18
Total	.82	.18	1.00

Kappa = .59
PV(-) = .99
PV(+) = .50

All Individuals
Prevalence = .50

	B		
A	-	+	Total
-	.44	.06	.50
+	.06	.44	.50
Total	.50	.50	1.00

Kappa = .76
PV(-) = .90
PV(+) = .90

.10 and when the prevalence is .50. The two kappa values are substantially different.

For the sake of completeness the predictive values are also considered. The predictive value of a positive diagnosis is denoted by PV(+) and the predictive value of a negative diagnosis is denoted by PV(-). Given that the physician declares that the individual has the disease, then PV(+) is the probability that this positive diagnosis is correct. Given that the physician declares that the individual does not have the disease, then PV(-) is the probability that this negative diagnosis is correct. It is well known that the predictive values strongly depend on the prevalence of the disease. Bayes theorem may be used to compute PV(+) and PV(-), assuming that the sensitivity and the specificity do not depend on the prevalence of the disease and are constant over populations of individuals. Notice that, in table 1.8-2, the lower kappa value corresponds with a rather low PV(+). Since both physicians have the same marginal distribution, kappa equals the correlation coefficient, which is often denoted as phi for this dichotomous case; see section 1.5.

Now, suppose there are two symptoms S1 and S2, where S1 is

always present if S2 is present, and vice versa. Because of errors made by observers the perfect agreement between S1 and S2 is unnoted. Suppose further that table 1.8-1 applies to both S1 and S2, physician A judges the presence or absence of S1 and physician B judges the presence or absence of S2. Then table 1.8-2 shows the association between S1 and S2 as observed by physicians A and B. If the association between S1 and S2 is to be investigated in a certain number of individuals, then the usual chi square test has greater power when the prevalence is .50 than when the prevalence is .10. This illustrates that the influence of population characteristics may be considered a desirable property of kappa; see also section 3.4. In addition, in many situations there is little reason to assume that sensitivity and specificity are constant over different populations:

i) Knowledge of the prevalence probably will have some influence on the diagnosis; in view of Bayes theorem, applied to compute the predictive values, this influence may be desirable.

ii) In the subpopulation of patients with the disease, the percentage of seriously ill patients probably will not be constant all over the world. As a consequence, the sensitivity probably will not be constant.

Bibliographic Notes, 1.8

Kraemer(1979, 1982) showed that the influence of population characteristics may be considered a desirable property of kappa; see also section 3.4. In a scientific study of certain associations, judgements will have little value if nearly all individuals are assigned to the same category. An association cannot even be investigated if all individuals fall into the same category.

1.9 Sampling Theory: The Standard Jackknife

When randomly selected subjects have been judged by two fixed observers, the resulting kappa value is an estimate of the corresponding parameter κ for the population of subjects and the two observers under consideration. In section 1.2 the parameter κ is implicitly defined as a function of the probabilities that are estimated by the proportions p(i,j) and q(i,j). Inferences about the parameter κ can be drawn by computing an approximate confidence interval on the basis of the statistic kappa and its standard error. For the sake of clarity: in this section independence between observers is not assumed and marginal proportions are not considered fixed.

This section is not confined to the case of only two observers. The results below also apply to the situations with more than two observers that are discussed in chapters 2 and 3. When randomly selected subjects have been judged by more than two fixed observers, the resulting kappa value is an estimate of the corresponding parameter κ for the population of subjects and the fixed group of observers under consideration. In the case of varying observers, treated in section 3.3, the statistic kappa corresponds with the parameter κ for the population of subjects and the population of observers under consideration.

In chapter 4 several statistical methods are described to estimate the standard error of kappa and to perform statistical tests. When the most simple method in section 4.2 is used, sampling fluctuations in the chance proportion of agreement e are not taken into account and the true standard error may be seriously overestimated; see Cicchetti and Fleiss(1977) and the bibliographic notes to section 4.2. The delta method, the jackknife and the bootstrap do account for sampling fluctuations in e. From Parr and Tolley(1982), Efron and Gong(1983, especially the last sentence in section 4) and Efron(1982, especially pages 13, 14 and 21) it can be concluded that the standard jackknife probably is a (very) good method. Moreover, with respect to the problems considered in chapters 2 and 3, among the good methods the standard jackknife is most easy to apply, although it is necessary to use a computer.

Therefore throughout chapters 1, 2 and 3 the standard jackknife technique is applied. According to Parr and Tolley(1982): If y is a real function (such as kappa) of multinomial proportions, with continuous first and second partial derivatives, in large samples of subjects y approximately follows a normal distribution and the standard jackknife may be applied. Below this technique is described in general terms.

Let y be the statistic (such as kappa) computed from the complete sample of N subjects and let $y_{(-h)}$ be the value of the statistic when the h-th subject is deleted from the sample. The h-th pseudovalue is computed as

$$y^{(h)} = Ny - (N-1)y_{(-h)} \cdot \qquad (1.9\text{-}1)$$

The jackknife estimate

$$y^{(.)} = \frac{1}{N} \sum_{h=1}^{N} y^{(h)} \qquad (1.9\text{-}2)$$

has a smaller bias than y. In most applications, however, the difference between $y^{(.)}$ and y is negligible compared to their standard errors; see e.g. Efron(1982, p.8). For the sake of simple interpretation, throughout chapters 1, 2 and 3 the value of y is presented and not the value of $y^{(.)}$. The true standard errors of $y^{(.)}$ and y are denoted by $\sigma(y^{(.)})$ and $\sigma(y)$, and the true variances by $\sigma^2(y^{(.)})$ and $\sigma^2(y)$. In my opinion, it is not necessary to use a different mathematical notation to denote an observed value and the corresponding random variable, at least in this book. Although $\sigma^2(y^{(.)})$ and $\sigma^2(y)$ are not exactly equal, the same estimate, namely

$$s^2(y^{(.)}) = s^2(y) = \frac{1}{N(N-1)} \sum_{h=1}^{N} (y^{(h)} - y^{(.)})^2 , \qquad (1.9\text{-}3)$$

is used to estimate both $\sigma^2(y^{(.)})$ and $\sigma^2(y)$; see e.g. Parr and Tolley (1982, the end of section 1) and Efron(1982, note on page 13). The estimated standard error, which is the square root of $s^2(y^{(.)}) = s^2(y)$, is denoted by $s(y^{(.)})$, s(y), s.e. or shortly s. In small samples the

estimated standard error tends to be slightly greater, in expectation, than the true standard error; see Efron(1982, pages 21 and 42). This implies that statistical tests may be slightly conservative, that is the tail-probability may be slightly too large, especially in small samples.

The standard jackknife, described in the preceding paragraph where a single subject is deleted from the sample, is a special case of the grouped jackknife where a group of subjects is deleted from the sample; see section 4.4. The grouped jackknife requires considerably less computational effort, but results in less stable estimates.

When z_α denotes the 100α percentile point of the standard normal distribution, the interval $y^{(.)} \pm z_\alpha s$, or $y \pm z_\alpha s$, may be taken as an approximate $100(1 - 2\alpha)\%$ confidence interval for the population parameter corresponding with y. In many situations the skewness of the distribution of y is probably more important than the difference between $y^{(.)}$ and y.

Suppose the statistics y_1 and y_2 are computed from the same sample of N subjects. Examples may be found in sections 2.2 and 3.3: y_1 and y_2 may be kappa values before and after combining categories, or y_1 and y_2 may be kappa values corresponding to different groups of observers. Obviously, the dependent statistics y_1 and y_2 may be compared by defining the statistic $y = y_1 - y_2$ and applying the standard jackknife technique described above, with now $y^{(h)} = y_1^{(h)} - y_2^{(h)}$. The ratio $z = y^{(.)}/s(y^{(.)})$ may be referred to tables of the standard normal distribution to approximately test if y_1 and y_2 are significantly different. When not all subjects are involved in the computation of y_1 and y_2, e.g. due to missing judgements in the case that two groups of observers are to be compared, I suggest to restrict the statistical test to the data from those subjects that are involved in the computation of both y_1 and y_2.

Suppose the statistics y_1 and y_2 are computed from independent samples of subjects, and the standard jackknife technique results in the estimated variances $s^2(y_1^{(.)}) = s^2(y_1)$ and $s^2(y_2^{(.)}) = s^2(y_2)$. In large samples of subjects the independent statistics y_1 and y_2 may be compared by referring the value of the ratio

$$z = \frac{y_1^{(.)} - y_2^{(.)}}{\sqrt{(s^2(y_1^{(.)}) + s^2(y_2^{(.)}))}} \qquad (1.9\text{-}4)$$

to tables of the standard normal distribution.

If all observers assign all subjects to the same category, e = 1 and the statistic kappa cannot be computed; see also section 1.2, the paragraph between formulas (1.2-4) and (1.2-5). So the above described standard jackknife technique cannot be applied if all observers assign all subjects but one to the same category. In such a case asymptotic methods should not be applied and, in my opinion, it is also not useful to compute any measure of agreement because of too large sampling fluctuations.

When kappa = 1, its estimated standard error is s.e. = 0, however small the sample of subjects may be. In this case application of other methods also results in the estimate s.e. = 0. Only in the case that also $\kappa = 1$ we have $\sigma(k) = 0$.

In the case of two fixed observers and L categories there are at most L^2 different pseudovalues and a more efficient computation of the standard error is possible; see also Fleiss and Davies(1982). As in section 2.1, the observed frequency f(i,j) denotes the number of subjects assigned to category i by observer 1 and to category j by observer 2. Let y be the statistic computed from the complete data set regarding N subjects. For f(i,j) > 0, let $y_{(-i,-j)}$ be the value of the statistic if f(i,j) is replaced by f(i,j)-1. For f(i,j) = 0, let $y_{(-i,-j)}$ be some constant, e.g. $y_{(-i,-j)} = 0$ if f(i,j) = 0; below it will become clear that it does not matter which constant is chosen for $y_{(-i,-j)}$ if f(i,j) = 0. The pseudovalue

$$y^{(i,j)} = Ny - (N-1)y_{(-i,-j)} \qquad (1.9\text{-}5)$$

refers to the i-th row and j-th column of the frequency table. The jackknife estimate is computed as

$$y^{(.)} = \frac{1}{N} \sum_{i=1}^{L} \sum_{j=1}^{L} f(i,j)y^{(i,j)} \qquad (1.9\text{-}6)$$

and its estimated standard error is computed from

$$s^2(y^{(.)}) = s^2(y) = \frac{1}{N(N-1)} \sum_{i=1}^{L} \sum_{j=1}^{L} f(i,j)(y^{(i,j)} - y^{(.)})^2 \ . \quad (1.9\text{-}7)$$

Bibliographic Notes, 1.9

Quenouille(1956) introduced the jackknife technique to reduce the bias in an estimate of a population parameter. Tukey(1958) suggested the estimation procedure for the standard error of the improved estimator. Efron(1982), Efron and Gong(1983) and Parr(1983) compared the jackknife technique to other methods. Kraemer(1980) applied the jackknife technique in an interobserver agreement study.

Parr and Tolley(1982) compared the jackknife technique to the delta method in the case of a multinomial distribution. Fleiss and Davies (1982) applied the jackknife technique to estimate the standard error of kappa in the case of two varying observers and two categories; they also suggested a modification.

Fleiss(1981, section 13.2) described a simple chi square test to compare more than two independent kappa statistics. Weighted least squares regression may be used to perform a test for linear trend or to describe the dependence of kappa on several variables; see e.g. Steel and Torrie(1980, section 10.13) for the theory and Musch, Landis, Higgins, Gilson and Jones(1984) for an interesting example.

Chapter 2

AGREEMENT AMONG MANY OBSERVERS

When All Subjects are Judged by the Same Observers

2.1 A Clinical Diagnosis Example

The present chapter is concerned with experiments where each of a random sample of N subjects is classified into one of L categories by each of a fixed group of n observers. Conclusions drawn have reference to the n fixed observers and the population of subjects from which a random sample is taken.

The statistical procedures are illustrated within the context of a clinical diagnosis study described by Holmquist, McMahan and Williams (1967). The first two observers in this study were also considered in sections 1.2 to 1.5 inclusive. The study was designed to investigate the variability in the histological classification of carcinoma in situ and related lesions of the uterine cervix. Each of n=7 pathologists separately classified N=118 biopsy slides into one of the following L=5 ordered categories based on the most involved lesion:

TABLE 2.1-1
Separate Classification by Seven Pathologists of Most Involved Histological Lesion of the Uterine Cervix

	Pathologist								Pathologist								Pathologist							
Slide	2	5	1	7	3	4	6		Slide	2	5	1	7	3	4	6	Slide	2	5	1	7	3	4	6
1	3	3	4	3	4	2	3		44	2	2	3	2	2	2	1	86	3	3	3	3	2	3	3
2	1	1	1	1	1	1	1		45	1	2	1	1	1	1	1	87	3	3	4	3	3	3	3
3	3	3	3	3	3	3	3		46	3	3	2	3	1	2	1	88	2	3	4	3	3	2	2
4	3	3	4	3	3	4	3		47	4	3	4	3	4	3	3	89	3	4	2	3	2	2	1
5	3	3	3	3	3	3	3		48	3	3	3	3	3	2	2	90	3	4	3	3	3	2	2
6	1	1	2	1	2	1	1		49	2	2	3	1	2	2	1	91	3	3	3	2	2	1	2
7	1	2	1	1	1	1	1		51	3	2	2	2	2	2	2	92	4	4	4	3	3	2	1
8	3	2	3	3	2	3	2		52	3	3	3	3	3	4	2	93	3	3	3	2	2	2	2
9	2	3	2	2	2	2	1		53	3	3	4	3	3	3	5	94	1	2	1	1	2	1	1
10	1	2	1	1	1	1	1		54	3	4	3	3	2	2	2	95	3	4	3	3	3	2	3
11	5	5	5	5	5	4	5		55	3	3	3	3	3	3	2	96	3	2	4	2	1	1	1
12	1	2	1	1	1	1	1		56	2	2	2	2	2	1	2	98	3	4	4	3	3	4	3
13	3	3	3	3	3	2	3		57	3	3	2	3	2	2	1	99	2	2	1	2	2	1	1
15	2	1	2	2	2	1	1		58	1	1	1	1	1	1	1	100	3	4	3	3	3	2	2
16	3	3	4	3	3	2	2		59	3	3	3	3	3	3	3	101	4	4	4	4	3	4	3
17	3	3	3	3	2	3	3		60	1	1	1	1	2	1	1	102	3	3	3	3	2	2	3
18	3	3	2	3	2	2	2		61	3	2	1	1	2	1	1	103	1	1	1	1	1	1	1
19	1	2	2	1	2	1	1		62	3	3	4	3	3	3	2	104	3	4	2	2	2	2	1
22	3	2	2	3	2	2	1		63	3	2	1	2	2	2	1	105	3	3	3	3	3	3	2
23	1	1	1	1	2	1	1		64	3	3	2	3	2	2	2	106	3	3	2	1	1	1	1
24	3	3	4	3	3	4	3		65	3	3	4	3	3	3	3	107	3	3	3	3	2	2	2
25	1	2	1	1	2	1	1		66	3	3	3	4	3	4	2	108	3	3	3	3	2	2	1
26	1	1	1	1	1	1	1		67	1	1	1	1	1	1	1	110	2	2	2	1	1	1	1
27	1	2	2	2	2	2	1		68	3	3	2	2	2	2	2	111	1	2	1	1	1	1	1
28	4	4	4	3	4	2	3		69	3	3	3	3	2	3	1	112	3	2	3	3	2	2	2
29	3	3	3	3	3	2	2		70	1	1	1	1	1	1	1	113	3	2	3	2	2	2	1
30	3	3	3	3	3	3	2		71	3	3	4	3	3	3	3	114	3	2	2	1	1	1	1
31	1	1	1	1	1	1	1		72	3	3	3	3	3	2	1	115	3	3	3	3	2	2	2
32	3	3	4	3	3	3	2		73	3	3	3	3	3	3	2	116	1	2	1	1	1	1	1
33	3	3	3	3	3	3	3		74	3	3	4	3	1	3	2	117	3	3	3	3	3	2	2
34	1	1	1	1	1	1	1		76	2	1	1	1	1	1	1	118	3	3	3	3	2	2	1
35	3	3	3	3	3	2	1		77	2	2	2	2	1	2	1	119	1	2	1	1	1	1	1
36	2	3	2	2	2	2	1		78	3	3	2	2	2	1	2	120	1	1	1	1	1	1	1
37	3	3	3	3	2	2	1		79	1	1	2	1	1	2	1	121	2	2	2	2	1	1	1
38	3	4	5	3	3	3	1		80	4	4	4	3	3	2	1	122	3	3	5	3	4	2	4
39	1	2	2	1	1	1	1		81	1	1	1	1	1	1	1	123	3	4	4	3	4	2	1
40	3	3	3	3	2	2	1		82	4	4	4	3	3	3	3	124	1	2	1	1	1	1	1
41	3	3	3	3	3	3	2		83	5	5	5	4	1	4	5	126	3	2	2	2	1	1	1
42	5	5	5	5	5	5	5		84	3	2	2	2	2	2	1								
43	3	3	5	3	3	2	2		85	4	5	4	3	4	2	1								

Category 1: Negative
Category 2: Atypical Squamous Hyperplasia
Category 3: Carcinoma in Situ
Category 4: Squamous Carcinoma with Early Stromal Invasion
Category 5: Invasive Carcinoma

If the final diagnosis for a patient is characterized by categories 3, 4 or 5, this patient has to be operated upon for cancer. The full data resulting from this classification are presented in table 2.1-1; the data resulting from the classification by pathologists 1 and 2 were also presented in table 1.2-1. Slide number 85 was assigned to all five categories. Each of the slides 38, 74, 80, 89, 92, 96, 104, 122 and 123 were assigned to four different categories. These slides may be used in training sessions in order to improve future interobserver agreement. The statistical analysis below is based on the complete data set.

Table 2.1-2 shows how the histological slides were distributed

TABLE 2.1-2
Marginal Proportions of Biopsy Slides Classified by Seven Pathologists According to Most Involved Histological Lesion of the Uterine Cervix

	Category							
Pathologist	1	2	3	4	5	3,4 or 5	$\bar{x}$	s.d.
2	.23	.10	.58	.06	.03	.67	2.6	1.2
5	.14	.26	.45	.12	.03	.60	2.7	1.0
1	.22	.22	.32	.19	.05	.56	2.6	1.2
7	.27	.17	.52	.03	.02	.56	2.3	1.0
3	.26	.36	.31	.05	.02	.38	2.2	0.9
4	.32	.41	.19	.07	.01	.27	2.0	0.9
6	.53	.26	.17	.01	.03	.21	1.8	1.0

Category 1: Negative
Category 2: Atypical Squamous Hyperplasia
Category 3: Carcinoma in Situ
Category 4: Squamous Carcinoma with Early Stromal Invasion
Category 5: Invasive Carcinoma

over the categories by each pathologist. The seventh column is most important since the treatments may be different for patients whose final diagnosis is characterized by categories 1 or 2, representing absence of carcinoma, as compared to categories 3, 4 or 5, representing presence of carcinoma. The distributions are remarkably different. In this table, and in all further tables, the pathologists are ordered according to the seventh column in table 2.1-2.

Treating the N=118 classifications as interval scores, for each pathologist the mean score $\bar{x}$ and standard deviation s.d. has been computed; see the last two columns in table 2.1-2. The ordering of the pathologists would not have been changed dramatically if the mean scores $\bar{x}$ had been used to order the pathologists.

Table 2.1-3 shows the raw data when categories 1 and 2 combined, representing absence of carcinoma(-), versus categories 3, 4 and 5 combined, representing presence of carcinoma(+), are considered. Less than the half of the slides is assigned to the same category, - or +, by

TABLE 2.1-3
Frequencies of Biopsy Slides Classified by Seven Pathologists According to Presence(+) or Absence(-) of Carcinoma in Situ of the Uterine Cervix

Pathologist 2517346	Frequency	Pathologist 2517346	Frequency
-------	34	+++----	7
		-++++--	1
+------	6	+-++-+-	1
-+-----	2		
--+----	2	+++++--	13
		++++-+-	2
++-----	4	++++--+	1
+-+----	2		
+--+---	1	++++++-	10
		++++-+	5
+++----	2	++++-++	3
++-+---	5		
+-++---	1	+++++++	16

all pathologists. If a patient has cancer(+), she should be operated.

For each pair of pathologists table 2.1-4 shows the kappa value reflecting the degree of agreement between both pathologists in classifying biopsy slides, using the 5-point scale. There is a relatively high degree of agreement among pathologists 1, 2, 5 and 7. Pathologists 4 and especially 6 are less in agreement with the others.

Table 2.1-5 shows the kappa values reflecting the degree of agreement in classifying biopsy slides according to presence(+: categories 3, 4 and 5 combined) or absence(-: categories 1 and 2 combined) of carcinoma in situ of the uterine cervix. From this table and the seventh column in table 2.1-2 it is obvious that, using the 2-point scale, kappa is high if both pathologists have about the same marginal distribution, while kappa is low if the corresponding marginal distributions are very

TABLE 2.1-4
Kappa Values Reflecting the Degree of Agreement in Classifying N=118 Biopsy Slides According to Most Involved Histological Lesion of the Uterine Cervix; L=5 Categories

Pathologist		2	5	1	7	3	4	6
2	Kappa		.50	.50	.63	.36	.29	.21
	s.e.		.06	.06	.06	.06	.05	.05
5	Kappa	.50		.38	.47	.32	.21	.13
	s.e.	.06		.06	.06	.06	.06	.05
1	Kappa	.50	.38		.47	.38	.33	.18
	s.e.	.06	.06		.06	.06	.06	.05
7	Kappa	.63	.47	.47		.51	.44	.31
	s.e.	.06	.06	.06		.06	.06	.05
3	Kappa	.36	.32	.38	.51		.42	.30
	s.e.	.06	.06	.06	.06		.06	.06
4	Kappa	.29	.21	.33	.44	.42		.34
	s.e.	.05	.06	.06	.06	.06		.06
6	Kappa	.21	.13	.18	.31	.30	.34	
	s.e.	.05	.05	.05	.05	.06	.06	

different. Again, there is a relatively high degree of agreement among pathologists 1, 2, 5 and 7, whereas pathologists 4 and especially 6 are less in agreement with the others.

Table 2.1-6 shows the upsilon values reflecting the degree of agreement when using a 5-point scale and quadratic disagreement weights; see sections 1.5 and 1.6. For the third time there is a relatively high degree of agreement among pathologists 1, 2, 5 and 7. Pathologists 3, 4 and especially 6 are less in agreement with the others. Pathologist 7 has the highest upsilon values.

From each of the last three tables the conclusion has been drawn that pathologists 1, 2, 5 and 7 form a homogeneous subgroup. In this respect it does not matter which of the three coefficients is used. Notice, however, that the kappa values in table 2.1-5 are much larger than the

TABLE 2.1-5
Kappa Values Reflecting the Degree of Agreement in Classifying N=118 Biopsy Slides According to Presence or Absence of Carcinoma in Situ of the Cervix; L=2 Categories

Pathologist		2	5	1	7	3	4	6
2	Kappa		.75	.66	.74	.44	.31	.23
	s.e.		.06	.07	.06	.07	.06	.05
5	Kappa	.75		.70	.81	.58	.36	.30
	s.e.	.06		.07	.05	.07	.06	.06
1	Kappa	.66	.70		.79	.65	.45	.35
	s.e.	.07	.07		.06	.06	.07	.06
7	Kappa	.74	.81	.79		.65	.45	.35
	s.e.	.06	.05	.06		.06	.07	.06
3	Kappa	.44	.58	.65	.65		.52	.45
	s.e.	.07	.07	.06	.06		.08	.08
4	Kappa	.31	.36	.45	.45	.52		.56
	s.e.	.06	.06	.07	.07	.08		.09
6	Kappa	.23	.30	.35	.35	.45	.56	
	s.e.	.05	.06	.06	.06	.08	.09	

kappa values in table 2.1-4 and much smaller than the upsilon values in table 2.1-6. It is obvious that the interpretation of the numerical value of a measure of agreement should not be absolute but relative. The value 0 is an exception: for all three coefficients 0 indicates just chance agreement. The value 1, representing perfect agreement, has not the same interpretation for a 5-point scale and a 2-point scale.

In the next section it is shown that the degree of agreement among pathologists 1, 2, 5 and 7 is significantly larger than the degree of agreement among all seven pathologists.

Bibliographic Notes, 2.1

Holmquist, McMahan and Williams(1967) described the study in more detail and presented the data contained in table 2.1-1.

TABLE 2.1-6

Upsilon Values Reflecting the Degree of Agreement in Classifying N=118 Biopsy Slides According to Most Involved Histological Lesion of the Uterine Cervix; L=5 Categories

Pathologist		2	5	1	7	3	4	6
2	Upsilon		.82	.78	.84	.63	.61	.46
	s.e.		.03	.04	.04	.08	.06	.07
5	Upsilon	.82		.74	.77	.62	.55	.40
	s.e.	.03		.04	.04	.07	.06	.08
1	Upsilon	.78	.74		.78	.68	.62	.50
	s.e.	.04	.04		.04	.07	.06	.07
7	Upsilon	.84	.77	.78		.75	.78	.57
	s.e.	.04	.04	.04		.06	.04	.07
3	Upsilon	.63	.62	.68	.75		.65	.56
	s.e.	.08	.07	.07	.06		.07	.09
4	Upsilon	.61	.55	.62	.78	.65		.68
	s.e.	.06	.06	.06	.04	.07		.05
6	Upsilon	.46	.40	.50	.57	.56	.68	
	s.e.	.07	.08	.07	.07	.09	.05	

2.2 *Agreement Within a Group of Observers*

The notation in this chapter is slightly different from the notation in chapter 1. This cannot be avoided in an acceptable manner, but the differences in notation are kept to a minimum and will hopefully cause no trouble.

The proportion of subjects assigned to category i by the a-th and to category j by the b-th observer is denoted by $p_{a,b}(i,j)$. With reference to the fixed group of n observers, the observed proportion

$$p(i,j) = \frac{1}{n(n-1)} \sum_{a=1}^{n} \sum_{\substack{b=1 \\ b \neq a}}^{n} p_{a,b}(i,j) \qquad (2.2\text{-}1)$$

is an estimate of the probability that a random subject is assigned to category i by the first and to category j by the second of two observers who are taken at random and without replacement from the whole group of n observers. Notice that the symmetry $p(i,j) = p(j,i)$ directly follows from the equality $p_{a,b}(i,j) = p_{b,a}(j,i)$.

The proportion of subjects assigned to category i by the a-th observer is denoted by $m_a(i)$. It is obvious that, for any $b \neq a$,

$$m_a(i) = \sum_{j=1}^{L} p_{a,b}(i,j) = \sum_{j=1}^{L} p_{b,a}(j,i) \; . \qquad (2.2\text{-}2)$$

These marginal proportions were already shown in table 2.1-2 and discussed in the preceding section.

Analogous to formula (1.2-5) the expected proportion

$$q_{a,b}(i,j) = m_a(i)m_b(j) \qquad (2.2\text{-}3)$$

is an estimate of the probability under independence that a random subject is assigned to category i by the a-th and to category j by the b-th observer. If all n assignments are independently distributed, the expected proportion

$$q(i,j) = \frac{1}{n(n-1)} \sum_{a=1}^{n} \sum_{\substack{b=1 \\ b \neq a}}^{n} q_{a,b}(i,j) \qquad (2.2\text{-}4)$$

is an estimate of the probability that a random subject is assigned to category i by the first and to category j by the second of two observers who are taken at random and without replacement from the whole group of n observers. Since $q_{a,b}(i,j) = q_{b,a}(j,i)$, for any $b \neq a$, the symmetry $q(i,j) = q(j,i)$ holds.

TABLE 2.2-1
Agreement Among All Seven Pathologists:
Observed and Expected (Under Independence) Proportions of
Biopsy Slides Classified by Two Random Pathologists

Category by the One Pathologist		Category by the Other Pathologist					
		1	2	3	4	5	Total
1	Observed	.19	.06	.02	.00	.00	.28
	Expected	.08	.07	.10	.02	.01	
2	Observed	.06	.09	.09	.01	.00	.25
	Expected	.07	.06	.09	.02	.01	
3	Observed	.02	.09	.22	.04	.00	.36
	Expected	.10	.09	.13	.03	.01	
4	Observed	.00	.01	.04	.02	.00	.07
	Expected	.02	.02	.03	.00	.00	
5	Observed	.00	.00	.00	.00	.02	.03
	Expected	.01	.01	.01	.00	.00	
Total		.28	.25	.36	.07	.03	1.00
Conditional		.68	.37	.60	.23	.64	

Category 1: Negative
Category 2: Atypical Squamous Hyperplasia
Category 3: Carcinoma in Situ
Category 4: Squamous Carcinoma with Early Stromal Invasion
Category 5: Invasive Carcinoma
Kappa = .36 with Standard Error .03
Upsilon = .65 with Standard Error .04

The observed proportions p(i,j) in table 2.2-1 show that the pathologists often differed one point on the 5-point scale, but seldom differed more than one point. On the diagonal the observed proportions are substantially larger than is to be expected under independence.

Notice that q(i,j) cannot be computed from the marginal proportions

$$\begin{aligned} p(i,+) &= \sum_{j=1}^{L} p(i,j) = \sum_{j=1}^{L} q(i,j) = \frac{1}{n} \sum_{a=1}^{n} m_a(i) \\ &= \sum_{j=1}^{L} p(j,i) = \sum_{j=1}^{L} q(j,i) = p(+,i) \end{aligned} \qquad (2.2\text{-}5)$$

because in general $q(i,j) \neq p(i,+)p(+,j)$, which is proved in the appendix directly following the present section. Even under the null hypothesis of n statistically independent assignments, the assignments by two random observers may be dependent because the random selection of observers is without replacement: the population of observers is changed by the selection of the first random observer. The second random observer is taken from the remaining n-1 observers, a subpopulation that completely depends on the first random observer. This may be important when the number of observers is very small, say $n<5$, in the case that the n marginal distributions are substantially different.

Analogous to formula (1.2-6) in section 1.2,

$$o_{a,b} = \sum_{i=1}^{L} p_{a,b}(i,i) \qquad (2.2\text{-}6)$$

is the observed proportion of agreement between the a-th and b-th observer. The observed proportion of agreement between two random observers, defined by

$$o = \sum_{i=1}^{L} p(i,i) = \frac{1}{n(n-1)} \sum_{a=1}^{n} \sum_{\substack{b=1 \\ b \neq a}}^{n} o_{a,b} \,, \qquad (2.2\text{-}7)$$

is an estimate of the probability that a random subject is assigned to the same category by both of two observers who are taken at random

and without replacement from the whole group of n observers. From the diagonal in table 2.2-1 we compute o = .54 .

Analogous to formula (1.2-7)

$$e_{a,b} = \sum_{i=1}^{L} q_{a,b}(i,i) \qquad (2.2\text{-}8)$$

is the expected proportion of agreement between the a-th and b-th observer under the null hypothesis of independence. The expected proportion of agreement between two random observers, defined by

$$e = \sum_{i=1}^{L} q(i,i) = \frac{1}{n(n-1)} \sum_{a=1}^{n} \sum_{\substack{b=1 \\ b \neq a}}^{n} e_{a,b} \ , \qquad (2.2\text{-}9)$$

is an estimate of the probability under independence that a random subject is assigned to the same category by both of two observers who are taken at random and without replacement from the whole group of n observers. In other words, e is an estimate of the probability that two random observers agree by chance. From the diagonal in table 2.2-1 we compute e = .27 . The observed proportion of agreement is twice what is to be expected under independence.

Analogous to formula (1.2-4) the kappa coefficient

$$k_{a,b} = \frac{o_{a,b} - e_{a,b}}{1 - e_{a,b}} \qquad (2.2\text{-}10)$$

is a measure of agreement between the a-th and b-th observer; see table 2.1-4. The kappa coefficient

$$k = \frac{o - e}{1 - e} \qquad (2.2\text{-}11)$$

is a measure of agreement between two random observers and may be considered a group measure of agreement among the observers within a group. The statistic k can be written as a weighted average of the statistics $k_{a,b}$:

$$k = \frac{\sum_{a=1}^{n} \sum_{\substack{b=1 \\ b \neq a}}^{n} (1 - e_{a,b}) k_{a,b}}{\sum_{a=1}^{n} \sum_{\substack{b=1 \\ b \neq a}}^{n} (1 - e_{a,b})} \qquad (2.2\text{-}12)$$

For the whole group of seven pathologists we have $k = .36$. In the preceding section we saw that the pathologists 1, 2, 5 and 7 seem to form a homogeneous subgroup. For this subgroup we have $k = .49$; see table 2.2-2. Application of the jackknife technique described in section 1.9 results in the standard normal value $z = 4.76$ that corresponds with the one-sided tail-probability $P_1 < .0001$. Even when we consider that there are 35 subgroups of four pathologists, and the test was guided by the data, we must conclude that the kappa values .36 and .49 are significantly different. According to the Bonferroni inequality, see e.g. Cox and Hinkley(1974, p.78), we have a small tail-probability $P < .0035$: under the null hypothesis that the degree of agreement is the same for all pairs of observers, there is a probability $P < .0035$ of obtaining a one-sided tail-probability $< .0001$ for at least one of the 35 subgroups of four pathologists. So among pathologists 1, 2, 5 and 7 the degree of agreement is significantly higher than among all seven pathologists. In other words, the subgroup is significantly more homogeneous than the whole group.

A remark is to be made that is especially important when the interobserver agreement study is part of a larger study, e.g. a clinical trial or epidemiological survey. We should not blindly go by significance tests and the decision whether an observer belongs to the homogeneous subgroup or not should also depend on the kappa values and on practical considerations.

Other coefficients, like intraclass kappa, upsilon and weighted kappa, can be computed from the proportions p(i,j) and q(i,j) in a straightforward manner. There is a significant difference between the upsilon value .65 for the whole group of seven pathologists and the upsilon value .79 for the subgroup consisting of pathologists 1, 2, 5

and 7 ($z = 5.50$, $P_1 < .0001$).

The last row in tables 2.2-1 and 2.2-2 shows the conditional proportions

$$p(i|i) = \frac{p(i,i)}{p(i,+)} . \qquad (2.2\text{-}13)$$

When a random observer uses category i, there is an estimated probability $p(i|i)$ that a second random observer also uses category i. For each of

TABLE 2.2-2
Agreement Among Pathologists 1, 2, 5 and 7:
Observed and Expected (Under Independence) Proportions of
Biopsy Slides Classified by Two Random Pathologists

Category by the One Pathologist		Category by the Other Pathologist					
		1	2	3	4	5	Total
1	Observed	.16	.05	.00	.00	.00	.21
	Expected	.04	.04	.10	.02	.01	
2	Observed	.05	.08	.05	.00	.00	.19
	Expected	.04	.03	.09	.02	.01	
3	Observed	.00	.05	.35	.06	.01	.47
	Expected	.10	.09	.22	.05	.02	
4	Observed	.00	.00	.06	.03	.00	.10
	Expected	.02	.02	.05	.01	.00	
5	Observed	.00	.00	.01	.00	.02	.03
	Expected	.01	.01	.02	.00	.00	
Total		.21	.19	.47	.10	.03	1.00
Conditional		.75	.44	.74	.32	.67	

Category 1: Negative
Category 2: Atypical Squamous Hyperplasia
Category 3: Carcinoma in Situ
Category 4: Squamous Carcinoma with Early Stromal Invasion
Category 5: Invasive Carcinoma
Kappa = .49 with Standard Error .04
Upsilon = .79 with Standard Error .03

TABLE 2.2-3

Agreement Among All Seven Pathologists: Proportions of Biopsy Slides Classified by Two Random Pathologists According to Presence(+) or Absence(-) of Carcinoma

Category by the One Pathologist		Category by the Other Pathologist		
		-	+	Total
-	Observed	.41	.12	.54
	Expected	.28	.25	
+	Observed	.12	.34	.46
	Expected	.25	.21	
Total		.54	.46	1.00
Conditional		.77	.74	

Kappa = .52 with Standard Error .04

TABLE 2.2-4

Agreement Among Pathologists 1,2,5 and 7: Proportions of Biopsy Slides Classified by Two Random Pathologists According to Presence(+) or Absence(-) of Carcinoma

Category by the One Pathologist		Category by the Other Pathologist		
		-	+	Total
-	Observed	.34	.06	.40
	Expected	.16	.24	
+	Observed	.06	.54	.60
	Expected	.24	.36	
Total		.40	.60	1.00
Conditional		.85	.90	

Kappa = .74 with Standard Error .04

the five categories in our example the conditional proportion is larger in the subgroup of pathologists 1, 2, 5 and 7 than in the whole group of seven pathologists.

In tables 2.2-3 and 2.2-4 categories 1 and 2 combined versus categories 3, 4 and 5 combined are considered, that is absence(-) versus presence(+) of carcinoma. The kappa value .74 for the homogeneous subgroup is significantly larger than the kappa value .52 for the whole group ($z = 6.00$, $P_1 < .0001$). Regarding the two conditional proportions, there is a substantial difference between both groups of pathologists. In section 1.3 it has been explained that it is to be expected that the kappa value for two categories is larger than the kappa value for five categories. Since the matrices p(i,j) and q(i,j) are symmetric, the theorem in that section now becomes:
Combining categories i and j increases kappa if and only if

$$\frac{p(i,j)}{q(i,j)} > 1 - \text{kappa} = 1 - k.$$

Bibliographic Notes, 2.2

Schouten(1982b) proposed formulas (2.2-1) and (2.2-4) to compute the proportions p(i,j) and q(i,j); formulas (2.2-7) and (2.2-9) are direct consequences of (2.2-1) and (2.2-4). Since the p(i,j) and q(i,j) clearly show which categories are confused, these proportions are important in themselves and not only useful to compute kappa statistics. Moreover, the problem of estimating any meaningful measure of agreement may be reduced to the problem of estimating the p(i,j) and q(i,j). Schouten (1982b) selected the homogeneous subgroup consisting of pathologists 1, 2, 5 and 7 in a stepwise manner; see also section 2.4.

Hubert(1977) and Conger(1980) indicated the efficient computation of kappa, but without explicitly mentioning the proportions p(i,j) and q(i,j). Light(1971) proposed to average the $k_{a,b}$; see also Conger(1980).

Appendix to section 2.2

In this appendix it is shown that the chance expected proportion $q(i,j)$ in general is not exactly equal to the product $p(i,+)p(+,j)$. First,

$$p(i,+) = \sum_{j=1}^{L} p(i,j) = \sum_{j=1}^{L} \frac{1}{n(n-1)} \sum_{a=1}^{n} \sum_{\substack{b=1 \\ b\neq a}}^{n} p_{a,b}(i,j)$$

$$= \frac{1}{n(n-1)} \sum_{a=1}^{n} \sum_{\substack{b=1 \\ b\neq a}}^{n} \sum_{j=1}^{L} p_{a,b}(i,j)$$

$$= \frac{1}{n(n-1)} \sum_{a=1}^{n} \sum_{\substack{b=1 \\ b\neq a}}^{n} m_a(i) = \frac{1}{n} \sum_{a=1}^{n} m_a(i) \ ,$$

and this equality is used below:

$$p(i,+)p(+,j) = \left(\frac{1}{n} \sum_{a=1}^{n} m_a(i) \right)\left(\frac{1}{n} \sum_{b=1}^{n} m_b(j) \right)$$

$$= \frac{1}{n^2} \sum_{a=1}^{n} \sum_{b=1}^{n} m_a(i)m_b(j)$$

$$= \frac{1}{n^2} \sum_{a=1}^{n} m_a(i)m_a(j) + \frac{1}{n^2} \sum_{a=1}^{n} \sum_{\substack{b=1 \\ b\neq a}}^{n} m_a(i)m_b(j)$$

$$= \frac{1}{n^2} \sum_{a=1}^{n} m_a(i)m_a(j) + \frac{1}{n^2} \sum_{a=1}^{n} \sum_{\substack{b=1 \\ b\neq a}}^{n} q_{a,b}(i,j)$$

$$= \frac{1}{n^2} \sum_{a=1}^{n} m_a(i)m_a(j) + \frac{n-1}{n} q(i,j) .$$

Thus

$$\frac{n-1}{n} p(i,+)p(+,j) + \frac{1}{n} p(i,+)p(+,j) = \frac{1}{n^2} \sum_{a=1}^{n} m_a(i)m_a(j) + \frac{n-1}{n} q(i,j)$$

and

$$p(i,+)p(+,j) - q(i,j) = \frac{1}{n-1} \left(\frac{1}{n} \sum_{a=1}^{n} m_a(i)m_a(j) - p(i,+)p(+,j) \right)$$

$$= \frac{1}{n(n-1)} \sum_{a=1}^{n} (m_a(i) - p(i,+))(m_a(j) - p(+,j)).$$

Notice that, in the last formula, the right hand side is of order 1/n. When there are more than five observers, the difference between $p(i,+)p(+,j)$ and $q(i,j)$ will often be negligible. The last formula shows that $p(i,+)p(+,j) = q(i,j)$ if the n marginal distributions are equal. For i=j we have

$$p(i,+)p(+,i) - q(i,i) = \frac{1}{n(n-1)} \sum_{a=1}^{n} (m_a(i) - p(i,+))^2 \geq 0$$

on the main diagonal. With respect to (unweighted) kappa = 1 - (1-o)/(1-e) this implies that

$$1 - \frac{1 - \Sigma p(i,i)}{1 - \Sigma p(i,+)p(+,i)} \leq 1 - \frac{1 - \Sigma p(i,i)}{1 - \Sigma q(i,i)} ,$$

a result earlier presented by Conger(1980).

2.3 *Agreement Between a Particular Observer and the Other Ones*

The observed proportion

$$p_a(i,j) = \frac{1}{n-1} \sum_{\substack{b=1 \\ b \neq a}}^{n} p_{a,b}(i,j) \qquad (2.3\text{-}1)$$

is an estimate of the probability that a random subject is assigned to category i by the a-th observer and to category j by an observer who is taken at random from the remaining n-1 observers. When the judgement by the a-th observer is not associated with any of the other n-1 judgements, the expected proportion

$$q_a(i,j) = \frac{1}{n-1} \sum_{\substack{b=1 \\ b \neq a}}^{n} q_{a,b}(i,j) \qquad (2.3\text{-}2)$$

is an estimate of the probability that the a-th observer uses category i and a randomly chosen other observer uses category j. Tables 2.3-1 and 2.3-2 show the observed and expected proportions $p_6(i,j)$ and $q_6(i,j)$. Pathologist 6 very often uses the next less serious category compared to the other pathologists. In these tables the expected proportions can be computed directly from the marginal proportions. Many of the conditional proportions, computed in analogy to formulas (1.2-8) and (1.2-9) in section 1.2, are extremely low.

The observed proportion of agreement

$$o_a = \sum_{i=1}^{L} p_a(i,i) = \frac{1}{n-1} \sum_{\substack{b=1 \\ b \neq a}}^{n} o_{a,b} \qquad (2.3\text{-}3)$$

and the expected proportion of agreement

$$e_a = \sum_{i=1}^{L} q_a(i,i) = \frac{1}{n-1} \sum_{\substack{b=1 \\ b \neq a}}^{n} e_{a,b} \qquad (2.3\text{-}4)$$

are used to compute the kappa statistic

$$k_a = \frac{o_a - e_a}{1 - e_a} \qquad (2.3\text{-}5)$$

as a measure of agreement between the a-th observer on the one hand and the remaining n-1 observers on the other hand.

The kappa statistic k for the whole group of n observers, as defined in the preceding section, is a weighted average of the kappa

TABLE 2.3-1
Agreement Between Pathologist 6 and the Other Ones: Observed and Expected (Under Independence) Proportions of Biopsy Slides Classified by Pathologists According to Most Involved Histological Lesion of the Uterine Cervix

Category by Pathologist 6		Category by the Other Pathologist 1	2	3	4	5	Total
1	Observed	.23	.17	.09	.02	.00	.53
	Expected	.13	.13	.21	.04	.01	
2	Observed	.01	.07	.17	.02	.00	.26
	Expected	.06	.07	.10	.02	.01	
3	Observed	.00	.01	.12	.04	.00	.17
	Expected	.04	.04	.07	.01	.00	
4	Observed	.00	.00	.00	.00	.00	.01
	Expected	.00	.00	.00	.00	.00	
5	Observed	.00	.00	.01	.01	.02	.03
	Expected	.01	.01	.01	.00	.00	
Total		.24	.25	.40	.08	.03	1.00
	$p_{6\|Other}$	.97	.27	.30	.02	.78	
	$p_{Other\|6}$	.44	.26	.70	.17	.58	

Category 1: Negative
Category 2: Atypical Squamous Hyperplasia
Category 3: Carcinoma in Situ
Category 4: Squamous Carcinoma with Early Stromal Invasion
Category 5: Invasive Carcinoma
Kappa = .24
Upsilon = .52

statistics k_a:

$$k = \frac{\sum_{a=1}^{n} (1 - e_a)k_a}{\sum_{a=1}^{n} (1 - e_a)} \quad . \qquad (2.3\text{-}6)$$

This implies that the group measure of agreement k increases by removing the observer with the smallest k_a value. A homogeneous subgroup of observers may be obtained by successively removing several observers from the group in a stepwise manner.

This section is concluded with the following remark: Since the correct diagnoses are unknown in the example, the results from the statistical analysis do not even suggest that pathologist 6 is making incorrect diagnoses. It is more likely that he is using different criteria than the other pathologists. Serious discussions about these criteria may help pathologists to speak the same language and to improve future interobserver agreement.

TABLE 2.3-2
Agreement Between Pathologist 6 and the Other Ones: Proportions of Biopsy Slides Classified According to Presence(+) or Absence(-) of Carcinoma

Category by Pathologist 6		Category by Another Pathologist		
		-	+	Total
-	Observed	.48	.31	.79
	Expected	.39	.40	
+	Observed	.01	.20	.21
	Expected	.10	.11	
Total		.49	.51	1.00
	$p_{6\|Other}$	.97	.61	
	$p_{Other\|6}$	.61	.07	

Kappa = .36

Bibliographic Notes, 2.3

The procedures in this section have been proposed by Schouten(1982b). Hubert(1977) indicated the efficient computation of k_a, but without mentioning the proportions $p_a(i,j)$ and $q_a(i,j)$.

2.4 *Agreement Between Two Groups of Observers*

Hierarchical Clustering

In order to detect important sources of disagreement we may try to divide the group of observers into several homogeneous subgroups, in such a way that the degree of interobserver agreement is higher within subgroups than between subgroups. Future interobserver agreement may be improved when we find out why and in which way subgroups differ in opinion. Kappa statistics may be used to identify such homogeneous subgroups which we call clusters. The cluster consisting of observers 1, 2, 5 and 7 is denoted by {1,2,5,7}, with analogous notation for other clusters.

For the cluster G consisting of $g \geq 2$ observers the observed proportion

$$p_G(i,j) = \frac{1}{g(g-1)} \sum_{a \in G} \sum_{\substack{b \in G \\ b \neq a}} p_{a,b}(i,j) \qquad (2.4\text{-}1)$$

is an estimate of the probability that a random subject is assigned to category i by the first and to category j by the second of two different observers taken at random from G. The expected proportion $q_G(i,j)$, the observed proportion of agreement o_G and the expected proportion of agreement e_G are defined in analogy to (2.4-1). The intracluster kappa coefficient

$$k_G = \frac{o_G - e_G}{1 - e_G} \qquad (2.4\text{-}2)$$

is a measure of interobserver agreement within G. In section 2.2 we compared the degree of interobserver agreement within clusters {1,2,3,4,5,6,7} and {1,2,5,7}.

For the two clusters G and H consisting of g and h observers, where no observer belongs to G and to H, the observed proportion

$$p_{G,H}(i,j) = \frac{1}{gh} \sum_{a \in G} \sum_{b \in H} p_{a,b}(i,j) \qquad (2.4\text{-}3)$$

is an estimate of the probability that a random subject is assigned to category i by a random observer from G and to category j by a random observer from H. The expected proportion $q_{G,H}(i,j)$, the observed proportion of agreement $o_{G,H}$ and the expected proportion of agreement $e_{G,H}$ are defined in analogy to (2.4-3). The intercluster kappa coefficient

$$k_{G,H} = \frac{o_{G,H} - e_{G,H}}{1 - e_{G,H}} \qquad (2.4\text{-}4)$$

TABLE 2.4-1
Intracluster (On the Diagonal) and Intercluster (Off the Diagonal) Kappa Coefficients Reflecting the Degree of Agreement in Classifying Presence or Absence of Carcinoma in Situ of the Uterine Cervix

Cluster	{1,2,5,7}	{3}	{4}	{6}
{1,2,5,7}	.74	.58	.39	.31
{3}	.58		.52	.45
{4}	.39	.52		.56
{6}	.31	.45	.56	

Cluster	{1,2,3,5,7}	{4}	{6}
{1,2,3,5,7}	.67	.42	.33
{4}	.42		.56
{6}	.33	.56	

Cluster	{1,2,3,5,7}	{4,6}
{1,2,3,5,7}	.67	.37
{4,6}		.56

is a measure of interobserver agreement between G and H. Notice that only the matrices $\{o_{a,b}\}$ and $\{e_{a,b}\}$ are needed to compute all possible intracluster and intercluster kappa statistics. The coefficient k_a in the preceding section is an intercluster (or intergroup) kappa coefficient. Similar definitions hold for the intercluster upsilon coefficient $u_{G,H}$ and the intercluster weighted kappa coefficient $k_{G,H}(w)$.

The hierarchical cluster analysis starts with n clusters, where each cluster holds one observer. Next, the observers within the two clusters with the highest intercluster kappa coefficient are grouped together and form a new cluster, and this may go on until finally all observers are considered to be in one cluster. When applying this method to the example, regarding presence or absence of carcinoma, the clusters {5,7}, {1,5,7}, {1,2,5,7} and {1,2,3,5,7} are successively formed with corresponding intracluster kappa coefficients .81, .77, .74 and .67. Then the cluster {4,6} is formed with intracluster kappa coefficient $k_{\{4,6\}} = k_{4,6} = .56$. Table 2.4-1 shows some steps in the cluster

TABLE 2.4-2
Agreement Among Pathologists 1,2,3,5,7: Proportions of Biopsy Slides Classified According to Presence(+) or Absence(-) of Carcinoma in Situ of the Uterine Cervix

Category by the One Pathologist		Category by the Other Pathologist		
		-	+	Total
-	Observed	.36	.08	.45
	Expected	.20	.25	
+	Observed	.08	.47	.55
	Expected	.25	.30	
Total		.45	.55	1.00
Conditional		.82	.85	

Kappa =.67

analysis. Tables 2.4-2, 2.4-3 and 2.4-4 show the agreement within and between clusters {1,2,3,5,7} and {4,6}. The two off diagonal cells in table 2.4-4 clearly demonstrate that pathologists 1, 2, 3, 5 and 7 often see a carcinoma while pathologists 4 and 6 do not see a carcinoma; the reverse seldom happens.

The above described stepwise clustering was stopped at a rather arbitrary moment. In section 2.1 we saw only one homogeneous subgroup consisting of pathologists 1, 2, 5 and 7. In view of tables 2.1-5 and 2.4-1 the stepwise clustering may stop after cluster {1,2,5,7} has been formed, thus presenting clusters {1,2,5,7}, {3}, {4} and {6}.

The estimation procedures described in the first three paragraphs of this section may also prove useful when the clusters are not obtained by cluster analysis but are defined in advance as special groups of observers, e.g. when neurosurgeons, neurologists and psychiatrists are to be compared.

TABLE 2.4-3
Agreement Between Pathologists 4 and 6: Proportions of Biopsy Slides Classified According to Presence(+) or Absence(-) of Carcinoma in Situ of the Uterine Cervix

Category by Pathologist 4		Category by Pathologist 6		
		-	+	Total
-	Observed	.68	.05	.73
	Expected	.57	.15	
+	Observed	.11	.16	.27
	Expected	.21	.06	
Total		.79	.21	1.00
	$p_{4\|6}$	.86	.76	
	$p_{6\|4}$	.93	.59	

Kappa =.56

TABLE 2.4-4
Agreement Between {1,2,3,5,7} and {4,6}: Proportions of Biopsy Slides Classified According to Presence(+) or Absence(-) of Carcinoma in Situ of the Uterine Cervix

Category by Pathologist From {1,2,3,5,7}		Category by Pathologist From {4,6} -	+	Total
-	Observed	.44	.01	.45
	Expected	.34	.11	
+	Observed	.32	.23	.55
	Expected	.42	.13	
Total		.76	.24	1.00
	$p_{G\|H}$	.58	.96	
	$p_{H\|G}$	.98	.42	

for Clusters G={1,2,3,5,7} and H={4,6}

Kappa = .37

Bibliographic Notes, 2.4

The procedures in this section have been proposed by Schouten(1982b).

For the two-point scale Fidler and Nagelkerke(1985) proposed a parametric model to describe the data in this example. They concluded that only five descriptive parameters are needed and indicated the clusters {2}, {1,5,7}, {3} and {4,6}.

2.5 *Agreement with the Majority Opinion*

When the investigator is willing to believe that the majority of the observers is (nearly) always right, there is an odd number of observers and there are only two categories, it is possible to estimate the sensitivity, specificity and predictive values for each of the observers. Since most observers are part of the majority, such estimates probably are too optimistic. Moreover, when there are two homogeneous subgroups of observers, the smaller group may be right but the larger group is put in the right. Despite these dangers many people like this approach because it is straightforward and it is pleasant to know who is right and who is wrong. If only for the sake of completeness this approach is now discussed.

The sensitivity is the probability that a positive subject receives a positive judgement, e.g. for a woman with a cervix carcinoma the sensitivity is the probability that her pathologist thinks she has a cervix carcinoma. The sensitivity is estimated by

$$Se = p(+ \text{ by Observer} \mid + \text{ by Majority}). \qquad (2.5\text{-}1)$$

From the subjects with a positive majority opinion, Se is the proportion with also a positive judgement by the observer under consideration.

The specificity is the probability that a negative subject receives a negative judgement. It is estimated by

$$Sp = p(- \text{ by Observer} \mid - \text{ by Majority}). \qquad (2.5\text{-}2)$$

From the subjects with a negative majority opinion, Sp is the proportion with also a negative judgement by the observer under consideration.

When the judgement by an observer is positive, the probability that the judged subject actually is positive is called the predictive value of a positive judgement. It is estimated by

$$PV+ = p(+ \text{ by Majority} \mid + \text{ by Observer}). \qquad (2.5\text{-}3)$$

From the subjects with a positive judgement by the the observer considered, PV+ is the proportion with also a positive majority opinion. The predictive value of a negative judgement is analogously defined and estimated by

$$PV- = p(- \text{ by Majority} \mid - \text{ by Observer}). \qquad (2.5\text{-}4)$$

Sensitivity, specificity and both predictive values all are conditional probabilities that the judgement by the considered observer is correct, estimated under the assumption that the majority is correct. Table 2.5-1 shows the estimates Se, Sp, PV+ and PV-, where the majority opinion of all seven pathologists is considered the truth. The prevalence is the proportion of positive subjects. When we also remember the seventh column in table 2.1-2, showing the proportion of subjects with a positive judgement, the following conclusion may be drawn: When the proportion of positive judgements by a certain pathologist lies near the prevalence, then that pathologist scores high in table 2.5-1.

TABLE 2.5-1
Estimated Sensitivity, Specificity and Predictive Values When the Majority Opinion of All Seven Pathologists is Considered to be the Truth.

Pathologist	Se	Sp	PV+	PV-
2	.98	.64	.73	.97
5	.98	.78	.82	.98
1	1.00	.88	.89	1.00
7	1.00	.88	.89	1.00
3	.76	1.00	1.00	.81
4	.54	1.00	1.00	.69
6	.42	1.00	1.00	.63

Prevalence = .50

Pathologists 4 and 6 are put in a still more unfavourable light if the majority opinion of the pathologists 1, 2, 3, 5 and 7 is considered the truth; see table 2.5-2. Is this an honest presentation of the data?

TABLE 2.5-2
Estimated Sensitivity, Specificity and Predictive Values When the Majority Opinion of the Pathologists 1,2,3,5 and 7 is Considered to be the Truth

Pathologist	Se	Sp	PV+	PV-
2	.99	.75	.84	.97
5	.97	.88	.92	.96
1	.93	.92	.94	.90
7	.97	.98	.98	.96
3	.67	1.00	1.00	.70
4	.48	1.00	1.00	.59
6	.37	1.00	1.00	.55

Prevalence = .57

Especially in the case of three observers the estimation procedure described above may result in misleadingly optimistic estimates of the sensitivity, specificity and predictive values. Suppose that 100 subjects have been judged by three observers, that the three judgements are statistically independent and the observers 1, 2 and 3 give 60, 50 and 40 positive judgements respectively. The characteristics in table 2.5-3 are estimated considering the majority opinion as the truth. The estimated sensitivities, specificities and predictive values are not very high, but

TABLE 2.5-3
Hypothetical Data Showing Statistical Independence

Observer 1 2 3	Frequency		
- - -	12	Observer 1: Se=.84	Sp=.64
+ - -	18	PV+=.70	PV-=.80
- + -	12		
- - +	8	Observer 2: Se=.76	Sp=.76
+ + -	18	PV+=.76	PV-=.76
+ - +	12		
- + +	8	Observer 3: Se=.64	Sp=.84
+ + +	12	PV+=.80	PV-=.70

they also do not indicate that the judgements may be useless, whereas there is no reason at all to think that the judgements have any value because there is merely chance agreement.

Bibliographic Notes, 2.5

The definitions of sensitivity, specificity and predictive values have been adopted from the article by the Department of Clinical Epidemiology and Biostatistics in the McMaster University(1983); see also Schechter and Sheps(1985) and Connell and Koepsell(1985).

Assuming conditional independence between judgements by physicians given the true state of an individual, healthy or diseased, Walter(1984) proposed a maximum likelihood estimation procedure for sensitivity and specificity of each observer. His assumption does not hold, however, if there is an underlying continuum corresponding with the healthy versus diseased dichotomy, which will often be the case. For that reason I am not convinced that his results depend more on the data than on the adopted mathematical model.

2.6 *Probability of Correct Judgement by a Majority of Observers*

Judgements by a majority of observers are expected to be more reliable than judgements by a single observer. A quantitative comparison is possible if there are only two categories. A majority judgement by an odd number of observers is not guaranteed if there are more than two categories. Therefore in this section it is assumed that there are only two categories.

In practice a two-out-of-three majority opinion may be obtained as follows. In the first instance a subject is judged by two observers. If both observers agree, their opinion is the two-out-of-three majority opinion. If both observers disagree, the opinion of a third observer is the two-out-of-three majority opinion.

Below a single subject is considered. We are interested in the probability that this particular subject is assigned to the right category by a majority of the n judging observers, where n is odd. Assume that there is a large population of observers from which a random sample of n observers may be taken. When p denotes the proportion of observers in the population that would assign the subject to the correct category, the number of correctly judging observers in the sample follows a binomial distribution with parameters n and p. Table 2.6-1 shows the probability of a correct majority judgement by a given number of observers and for a certain proportion of correctly judging observers in the population. Of course, if most observers would use the wrong

TABLE 2.6-1
Probability of Correct Judgement by a Majority of Observers

Number of Observers	Proportion of Correctly Judging Observers						
	.40	.50	.60	.70	.80	.90	.95
3	.35	.50	.65	.78	.90	.97	.99
5	.32	.50	.68	.84	.94	.99	1.00
7	.29	.50	.71	.87	.97	1.00	1.00
9	.27	.50	.73	.90	.98	1.00	1.00

category for the subject under consideration, the judgement by a single observer has a larger probability to be correct than a majority judgement, but such subjects will be exceptions. If eighty percent of the observers in the population would use the right category, there is a probability .80 that a single observer uses the right category, whereas there is a probability .90 that a two-out-of-three majority opinion is correct. My conclusion is that majority judgements should be used more frequently, in scientific research and in daily practice.

Bibliographic Notes, 2.6

Sandifer, Fleiss and Green(1968) considered sample selection by diagnosis using one, one-out-of-three, two, two-out-of-three and three concurring opinions.

2.7 *Multivariate Agreement Weights*

Let $p(i_1,i_2,\ldots,i_n)$ denote the proportion of subjects assigned to category i_1 by the first observer, to category i_2 by the second observer, ..., and to category i_n by the n-th observer. As before, let $m_a(j)$ denote the proportion of subjects assigned to category j by the a-th observer.

Until now we only considered bivariate agreement weights $w(i,j)$, where $0 \leq w(i,j) \leq 1$. Landis and Koch(1977b), however, analyzed the example in this chapter using multivariate agreement weights $W(i_1,i_2,\ldots,i_n)$, where $0 \leq W(i_1,i_2,\ldots,i_n) \leq 1$. They defined the observed proportion of agreement as

$$o(W) = \sum_{i_1=1}^{L} \sum_{i_2=1}^{L} \cdots \sum_{i_n=1}^{L} p(i_1,i_2,\ldots,i_n)W(i_1,i_2,\ldots,i_n) \qquad (2.7\text{-}1)$$

$$= \frac{1}{N} \sum_{h=1}^{L} W(i_{h1},i_{h2},\ldots,i_{hn})$$

where i_{ha} denotes the category to which the h-th subject is assigned by the a-th observer. The chance expected proportion of agreement is

$$e(W) = \sum_{i_1=1}^{L} \sum_{i_2=1}^{L} \cdots \sum_{i_n=1}^{L} q(i_1,i_2,\ldots,i_n)W(i_1,i_2,\ldots,i_n) \qquad (2.7\text{-}2)$$

where $q(i_1,i_2,\ldots,i_n) = m_1(i_1)m_2(i_2)\cdots m_n(i_n)$

and the coefficient weighted kappa is

$$k(W) = \frac{o(W) - e(W)}{1 - e(W)} \qquad (2.7\text{-}3)$$

Below some special cases of $k(W)$ are mentioned: pairwise agreement, simultaneous agreement and majority agreement.

Pairwise Agreement

When the multivariate agreement weight $W(i_1,i_2,\ldots,i_n)$ is the average, over the $n(n-1)$ pairs of observers, of the $n(n-1)$ bivariate agreement weights $w(i_a,i_b)$, that is

$$W(i_1,i_2,\ldots,i_n) = \frac{1}{n(n-1)} \sum_{a=1}^{n} \sum_{\substack{b=1 \\ b\neq a}}^{n} w(i_a,i_b) , \qquad (2.7\text{-}4)$$

it is easy to see that

$$\begin{aligned} o(W) &= \frac{1}{N} \sum_{h=1}^{N} \frac{1}{n(n-1)} \sum_{a=1}^{n} \sum_{\substack{b=1 \\ b\neq a}}^{n} w(i_{ha},i_{hb}) \\ &= \frac{1}{n(n-1)} \sum_{a=1}^{n} \sum_{\substack{b=1 \\ b\neq a}}^{n} \frac{1}{N} \sum_{h=1}^{N} w(i_{ha},i_{hb}) \\ &= \frac{1}{n(n-1)} \sum_{a=1}^{n} \sum_{\substack{b=1 \\ b\neq a}}^{n} o_{a,b}(w) = o(w) \end{aligned} \qquad (2.7\text{-}5)$$

where the observed degree of agreement $o(w)$ is defined analogously to (2.2-7). It also can be shown that $e(W) = e(w)$ and so $k(W) = k(w)$. Since $k(w)$ is a measure of agreement between two random observers, the interpretation of $k(w)$ does not depend on the total number of observers. Therefore $k(w)$ may be used to compare groups with the same or different numbers of observers regarding the degree of inter-observer agreement within those groups: see section 2.2. Moreover, missing values can be handled: see section 3.2.

Simultaneous Agreement

By definition, there is simultaneous agreement among n observers if they all use the same category:

$$\begin{aligned} W(i_1,i_2,\ldots,i_n) &= 1 \quad \text{if} \quad i_1 = i_2 = \ldots = i_n \\ &= 0 \quad \text{otherwise} \end{aligned} \qquad (2.7\text{-}6)$$

In this case, the interpretation of the coefficient k(W) depends on the total number of observers. Therefore this k(W) may be used to compare groups with the same number of observers regarding the degree of simultaneous agreement within those groups. It is not sensible to compare groups with different numbers of observers regarding the degree of simultaneous agreement.

Majority Agreement

For some integer f with $\frac{1}{2}n < f \leq n$, there is f-out-of-n majority agreement if at least f of the n observers use the same category:

$$W(i_1,i_2,\ldots,i_n) = 1 \quad \text{if at least f arguments are equal}$$
$$= 0 \quad \text{otherwise} \qquad (2.7\text{-}7)$$

In this case, the interpretation of the coefficient k(W) depends on both f and n. Therefore it is not sensible to compare the degree of 6-out-of-7 majority agreement with the degree of 5-out-of-7 majority agreement or with the degree of 4-out-of-6 majority agreement. Notice that simultaneous agreement is a special case of majority agreement.

Bibliographic Notes, 2.7

With regard to the example considered in the present chapter Landis and Koch(1977b) assessed the degree of majority agreement in order to select a homogeneous subgroup of pathologists. They interpreted kappa values in an absolute manner and compared e.g. the degree of 6-out-of-7 majority agreement with the degree of 5-out-of-7 majority agreement. In my opinion, their statistical analysis has no sensible interpretation. A homogeneous subgroup of observers may be selected by using comparable kappa statistics; see sections 2.1 to 2.4 inclusive.

In order to identify subjects whom the observers find difficult to classify, $(W(i_{h1},i_{h2},\ldots,i_{hn}) - e(W))/(1 - e(W))$ may be used as a measure of agreement on the h-th subject. O'Connell and Dobson(1984) developed this idea for the weights defined in (2.7-4).

Chapter 3

AGREEMENT AMONG MANY OBSERVERS

When Subjects are Not Necessarily Judged by the Same Observers

3.1 A Clinical Diagnosis Example with Fixed Observers

Each of a random sample of N subjects is assigned to one of L categories by each available observer from a small group of n observers. Availability may depend on the point of time upon which a subject is to be judged, e.g. a patient is to be operated upon immediately and not all n judging physicians are present at that time. However, availability may also depend on skills and facilities. Conclusions drawn have reference to the n fixed observers and the population of subjects from which a random sample is taken. The only difference with the situation considered in chapter 2 is that there are missing values. The procedures treated in sections 3.1 and 3.2 are equivalent to the procedures in sections 2.1 and 2.2 if there are no missing values; this is proved in the appendix to section 3.2. In section 3.3 the observers are not considered fixed.

The statistical procedures are illustrated within the context of a study that was designed to investigate the observer variability in the assessment of pupil reaction to light in head-injured patients. These patients had been in coma for at least six hours. Each patient was examined by four physicians from a regular group of six. A more detailed description of the study can be found in the article by Van den Berge, Schouten, Boomstra, van Drunen Littel and Braakman(1979). When that study was designed the known statistical procedures demanded

a constant number of physicians per patient. The available physicians separately classified the patients according to presence(+) or absence(-) of a reaction to light of the right and left pupil. The data resulting from this classification are presented in table 3.1-1.

Table 3.1-2 shows how the patients were distributed over the categories by each physician. In all tables the physicians are ordered

TABLE 3.1-1
Separate Classification of Coma Patients According to Presence(+) or Absence(-) of Reaction of Pupils to Light

Reaction of Right Pupil							Reaction of Left Pupil						
	Physician							Physician					
Patient	5	1	3	4	2	6	Patient	5	1	3	4	2	6
1		-		-	-	-	1		-		-	-	-
2		-		-	-	-	2		-		+	-	-
3			-	-	-	-	3			-	-	-	-
4	+	-		+		-	4	+	-		-		-
5	+		+	+	+		5	+		+	+	+	
6	-		-	-	+		6	-		-	-	+	
7	+	+			-	+	7	+	+			+	+
8	+	+			-	-	8	+	+			-	-
9	+			+	+	+	9	+			+	+	+
10	+	+		+	+		10	+	+		+	+	
11		-	-	-		-	11		-	-	-		-
12	-		-	-		-	12	-		-	-		-
13	+	+	+			+	13	+	+	+			+
14	-	+	+			-	14	-	+	+			-
15	-		-		-	-	15	-		-		-	-
16	+		+		+	+	16	+		+		+	+
17	-		-		-	-	17	-		-		-	-
18	-		+		-	-	18	+		+		-	-
19	+		+		-	+	19	+		+		+	+
20	+	+	-			+	20	+	+	-			-
21	-	-	-			-	21	-	-	-			-
22	-		-		-	-	22	-		-		-	-
23	-		-		-	-	23	-		-		-	-
24			-	-	-	-	24			-	-	-	-
25			-	-	-	-	25			-	-	-	-
26	+	+	+			+	26	+	+	+			+
27	+	-	+			-	27	+	-	+			-
28	+	+			+	+	28	+	+			+	+

according to proportion of positive reactions of the right pupil. The physicians tend to distribute the patients differently over the categories. This is taken into account in the computation of chance expected proportions; see also the next section.

TABLE 3.1-2
Proportion of Positive Pupil Reactions

Physician	Pupil		Number of Patients
	Right	Left	
5	.59	.64	22
1	.57	.57	14
3	.40	.40	20
4	.33	.33	12
2	.32	.42	19
6	.32	.28	25

TABLE 3.1-3
Kappa Values Representing the Degree of Agreement in the Assessment of Reaction of Right(Above the Diagonal) and Left(Below the Diagonal) Pupil to Light in Coma Patients

Physician		5	1	3	4	2	6
5	Kappa		.23	.63	1.00	.44	.69
	s.e.		.45	.20		.25	.16
1	Kappa	.23		.42	.55	.40	.70
	s.e.	.45		.41	.63	.35	.20
3	Kappa	.75	.42		1.00	.40	.51
	s.e.	.17	.41			.34	.23
4	Kappa	.67	.55	1.00		.77	.61
	s.e.	.35	.63			.24	.68
2	Kappa	.55	.67	.63	.55		.67
	s.e.	.24	.35	.27	.31		.24
6	Kappa	.51	.55	.62	.61	1.00	
	s.e.	.18	.22	.20	.68		

For each pair of physicians table 3.1-3 shows the kappa value, separately for the right (above the diagonal) and the left (below the diagonal) pupil. Each kappa value has been computed from the judgements on those patients judged by both of the two physicians under consideration. Some kappa values have been based on a very small number of patients; $k_{1,4}$ has been based on only five patients, as can be concluded from table 3.1-1. For that reason the standard errors of the kappa statistics are very large, which implies that the kappa values heavily depend on random fluctuations. This makes these kappa values nearly useless. In the next section a better representation of the data is proposed.

Bibliographic Notes, 3.1

Van den Berge, Schouten, Boomstra, van Drunen Littel and Braakman (1979) described the study in more detail and presented the data shown in table 3.1-1.

3.2 Agreement Within a Group of Observers

The estimation procedures in this section are equivalent to those in section 2.2 if there are no missing values; this is shown in the appendix directly following this section. The notation needed in the estimation procedures is explained first. The h-th subject is judged by n_h observers and is assigned to category i by x_{hi} observers;

$$n_h = \sum_{i=1}^{L} x_{hi} .$$

The subgroup of n_h observers who judge the h-th subject is denoted by G_h. From the subjects judged by the a-th observer a proportion $m_a(i)$ is assigned to category i;

$$\sum_{i=1}^{L} m_a(i) = 1 .$$

These proportions are shown in table 3.1-2 of the preceding section. When a subject is judged by the a-th observer, $m_a(i)$ is an unbiased estimate of the probability that category i is used. As in section 2.2, in the computation of the expected proportions q(i,j) account is taken of the fact that the observers may distribute the subjects differently over the categories; see formulas (3.2-3) and (3.2-4) below.

From the $n_h(n_h - 1)$ ordered pairs of observers who judge the h-th subject there are $x_{hi}(x_{hi} - 1)$ ordered pairs of observers where both observers use category i. So the observed proportion

$$p(i,i) = \frac{1}{N} \sum_{h=1}^{N} \frac{x_{hi}(x_{hi} - 1)}{n_h(n_h - 1)} \qquad (3.2\text{-}1)$$

is an unbiased estimate of the probability that a randomly selected subject is assigned to category i by both of two observers who are taken at random from the available observers. For $i \neq j$, from the $n_h(n_h - 1)$ ordered pairs of observers who judge the h-th subject there are $x_{hi}x_{hj}$ ordered pairs of observers where the first observer uses category i and the second observer uses category j. So the observed

proportion

$$p(i,j) = p(j,i) = \frac{1}{N} \sum_{h=1}^{N} \frac{x_{hi}x_{hj}}{n_h(n_h - 1)} \quad \text{for } i \neq j \qquad (3.2\text{-}2)$$

is an unbiased estimate of the probability that a randomly selected subject is assigned to category i by the first and to category j by the second of two observers who are taken at random (and without replacement) from the available observers. The estimates in (3.2-1) and (3.2-2) are unbiased because the summands are unbiased.

In the preceding chapters we saw that it is necessary to investigate if the observed proportions p(i,j) substantially differ from the proportions that are to be expected under the null hypothesis that all judgements are independently distributed, especially on the diagonal where the agreement is perfect. Given that the h-th subject is judged by the n_h observers from G_h, under the null hypothesis of independence

$$q(h;\, i,j) = \frac{1}{n_h(n_h - 1)} \sum_{a \in G_h} \sum_{\substack{b \in G_h \\ b \neq a}} m_a(i)m_b(j) \qquad (3.2\text{-}3)$$

is an estimate of the probability that the h-th subject is assigned to category i by the first and to category j by the second of two observers who are taken at random from G_h. Thus, for the h-th subject, the chance expected proportions q(h; i,j) are based on the marginal proportions $m_a(i)$ of the judgements by those observers who actually judge this h-th subject. Observers who do not judge the h-th subject do not have any influence upon the q(h; i,j). The expected proportion

$$q(i,j) = q(j,i) = \frac{1}{N} \sum_{h=1}^{N} q(h;\, i,j) \qquad (3.2\text{-}4)$$

is an estimate of the probability under independence that a randomly selected subject is assigned to category i by the first and to category j by the second of two observers who are taken at random (and without replacement) from the available observers. If the null hypothesis of independence really holds, we may expect that p(i,j) and q(i,j) will be

about equal. According to (3.2-3) the estimates q(i,j) are based on the marginal proportions $m_a(i)$, which were shown in table 3.1-2 for the example mentioned in section 3.1.

The marginal proportion

$$p(i,+) = p(+,i) = \sum_{j=1}^{L} p(i,j) = \frac{1}{N} \sum_{h=1}^{N} \frac{x_{hi}}{n_h} \qquad (3.2\text{-}5)$$

is an unbiased estimate of the probability that a randomly selected subject is assigned to category i by an observer who is taken at random from the available observers. According to section 2.2 (the paragraph containing formula (2.2-5)) and the appendix to section 2.2 the expected proportion q(i,j) in general is not equal to p(i,+)p(+,j).

The conditional proportion

$$p(i|i) = \frac{p(i,i)}{p(i,+)} \qquad (3.2\text{-}6)$$

is an estimate of an important conditional probability: given that a randomly taken observer uses category i, there is an estimated probability p(i|i) that a second randomly taken observer also uses category i. Formulas (1.2-6), (1.2-7) and (1.2-4) in section 1.2 may be used to compute kappa as a group measure of agreement among the observers within a group. The theorem in section 1.3 then becomes:
Combining categories i and j increases kappa if and only if

$$\frac{p(i,j)}{q(i,j)} > 1 - \text{kappa} = 1 - k.$$

The formulas in section 1.4 may be used to compute intraclass kappas, the first three formulas in section 1.5 may be used to compute upsilon, and the formulas in section 1.6 may be used to compute weighted kappa.

For the data in table 3.1-1, table 3.2-1 shows the observed, expected and conditional proportions. Two random physicians disagree in 18 percent of the cases, regardless which of the two pupils is judged. If a physician observes that the left pupil of a coma patient reacts to light, there is an estimated probability .80 that a second physician also

observes a reaction. I think that the proportions shown in table 3.2-1 are easier to understand than measures of agreement like kappa.

TABLE 3.2-1
Agreement Between Two Random Physicians:
Observed and Expected (Under Independence)
Proportions of Coma Patients Classified According to
Presence(+) or Absence(-) of Pupil Reaction to Light

		Category by the Other Physician			
Category by the One Physician		Right Pupil		Left Pupil	
		-	+	-	+
-	Observed	.49	.09	.48	.09
	Expected	.33	.25	.31	.25
+	Observed	.09	.33	.09	.35
	Expected	.25	.17	.25	.19
Total		.58	.42	.56	.44
Conditional		.84	.78	.85	.80
Kappa		.62		.65	
s.e.		.11		.11	

In the rest of this section it is assumed that all N subjects are judged by all n observers, that is no judgement is missing and $n_h = n$ for all subjects. For that case it is indicated below that formulas (3.2-1) to (3.2-4) inclusive are equivalent to formulas (2.2-1) and (2.2-4) in section 2.2. Since $n_h = n$ and G_h is the whole group of n observers, $q(h; i,j) = q(i,j)$ and it easily follows that formulas (3.2-3) and (3.2-4) are equivalent to formula (2.2-4); (2.2-3) is used in (2.2-4), and $m_a(i)$ has the same meaning in sections 2.2 and 3.2. From all $Nn(n-1)$ ordered pairs of assignments, of the N subjects by the $n(n-1)$ ordered pairs of observers, there are $Nn(n-1)p(i,j)$ ordered pairs of assignments where the first assignment is to category i and the second assignment is to category j, irrespective whether formula (2.2-1) is used or formulas

(3.2-1) and (3.2-2) are used. An exact mathematical proof of the equivalence of (2.2-1) on the one hand and (3.2-1) to (3.2-2) on the other hand, however, requires the use of indicator variables and is presented in the appendix directly following the present section.

Bibliographic Notes, 3.2

Regarding the computation of the observed proportions p(i,j), formulas (3.2-1) and (3.2-2) are generalizations of formulas proposed by Schouten (1980, 1982a) for the case of varying observers as considered in the next section. Regarding the computation of the chance expected proportions q(i,j), formulas (3.2-3) and (3.2-4) have not been published elsewhere.

Uebersax(1982) also considered the case of many fixed observers and missing judgements. He proposed an efficient computation of kappa, but without mentioning the proportions p(i,j) and q(i,j). Consequently, he did not indicate the computation of weighted kappa. Moreover, his formulas are not easy to understand.

Appendix to section 3.2

In this appendix it is shown that formulas (3.2-1) and (3.2-2) in section 3.2, which formulas are used to compute the observed proportions p(i,j), are equivalent to formula (2.2-1) in section 2.2 *if there are no missing values*. First, the indicator variables x(h,a,i) are defined:

x(h,a,i) = 1 if subject h is assigned by observer a to category i
= 0 otherwise.

Notice that $x(h,a,i) = x(h,a,i)^2$, and $x(h,a,i)x(h,a,j) = 0$ if $i \neq j$.

Since $x_{hi} = \sum_{a=1}^{n} x(h,a,i)$,

and
$$x_{hi}^2 = \sum_{a=1}^{n} \sum_{b=1}^{n} x(h,a,i)x(h,b,i)$$
$$= \sum_{a=1}^{n} x(h,a,i) + \sum_{a=1}^{n} \sum_{\substack{b=1 \\ b \neq a}}^{n} x(h,a,i)x(h,b,i) ,$$

formula (3.2-1) can be rewritten as

$$p(i,i) = \frac{1}{N} \sum_{h=1}^{N} \frac{1}{n(n-1)} \sum_{a=1}^{n} \sum_{\substack{b=1 \\ b \neq a}}^{n} x(h,a,i)x(h,b,i) \qquad (3.2\text{-}1R)$$

Since
$$x_{hi}x_{hj} = \sum_{a=1}^{n} \sum_{b=1}^{n} x(h,a,i)x(h,b,j)$$
$$= \sum_{a=1}^{n} x(h,a,i)x(h,a,j) + \sum_{a=1}^{n} \sum_{\substack{b=1 \\ b \neq a}}^{n} x(h,a,i)x(h,b,j)$$
$$= \sum_{a=1}^{n} \sum_{\substack{b=1 \\ b \neq a}}^{n} x(h,a,i)x(h,b,j) \quad \text{for } i \neq j,$$

formula (3.2-2) can be rewritten as

$$p(i,j) = \frac{1}{N} \sum_{h=1}^{N} \frac{1}{n(n-1)} \sum_{a=1}^{n} \sum_{\substack{b=1 \\ b \neq a}}^{n} x(h,a,i)x(h,b,j) \qquad (3.2\text{-}2R)$$

Because of (3.2-1R), (3.2-2R) holds for $i \neq j$ and for $i=j$.

Since $p_{a,b}(i,j)$ can be written as

$$p_{a,b}(i,j) = \frac{1}{N} \sum_{h=1}^{N} x(h,a,i)x(h,b,j) ,$$

formula (2.2-1) can be rewritten as

$$p(i,j) = \frac{1}{n(n-1)} \sum_{a=1}^{n} \sum_{\substack{b=1 \\ b \neq a}}^{n} \frac{1}{N} \sum_{h=1}^{N} x(h,a,i)x(h,b,j) \qquad (2.2\text{-}1R)$$

Obviously, formulas (3.2-1R) and (3.2-2R) are equivalent to formula (2.2-1R), and so (3.2-1) and (3.2-2) are equivalent to (2.2-1).

3.3 A Clinical Diagnosis Example with Varying Observers

In this section an experiment is considered where for each subject anew a random sample of n_h observers is taken from a large population of observers; with respect to the sampling of the observers, it is assumed that the difference between sampling with replacement and sampling without replacement may be neglected. In such a case it is not necessary to apply formulas (3.2-3) and (3.2-4) in the preceding section, as can be concluded from section 2.2 (see the paragraph containing formula (2.2-5)) and from the appendix to section 2.2. A simpler procedure is considered on the next pages.

As an example table 3.3-1 shows data that were presented and analyzed in a fundamental paper by Fleiss(1971). Each of N=30 patients was classified separately by six psychiatrists into one of the following L=5 diagnostic categories:

Category 1: Depression
Category 2: Personality Disorder
Category 3: Schizophrenia
Category 4: Neurosis
Category 5: Other

For each patient anew six psychiatrists were chosen from a pool of 43 psychiatrists. In table 3.3-1 the number of psychiatrists is the same for all patients, but in the last paragraph of this section an example is given where the number of psychiatrists is not the same for all patients. For each patient table 3.3-1 shows the number of psychiatrists who assigned that patient to a certain category. In other words, table 3.3-1 shows the numbers x_{hi} defined in the preceding section; the sixth patient is assigned to category 1 by $x_{61} = 2$ psychiatrists and to category 3 by $x_{63} = 4$ psychiatrists, $n_6 = 6$. Formulas (3.2-1) and (3.2-2) can be used to estimate the probability that a random subject is assigned to category i by the first and to category j by the second of two random observers.

TABLE 3.3-1
Number of Psychiatrists by Whom a Patient is Assigned to a Certain Diagnostic Category

	Category						Category						Category				
Patient	1	2	3	4	5	Patient	1	2	3	4	5	Patient	1	2	3	4	5
1				6		11	1			5		21					6
2		3			3	12	1	1		4		22		1		5	
3		1	4		1	13			3	3		23		2		1	3
4					6	14	1			5		24	2			4	
5		3		3		15		2		3	1	25	1			4	1
6	2		4			16			5		1	26		5		1	
7			4		2	17	3			1	2	27	4				2
8	2		3	1		18	5	1				28		2		4	
9	2			4		19		2		4		29	1		5		
10					6	20	1		2		3	30					6

Category 1: Depression
Category 2: Personality Disorder
Category 3: Schizophrenia
Category 4: Neurosis
Category 5: Other

The marginal proportion

$$p(i,+) = \sum_{j=1}^{L} p(i,j) = \frac{1}{N} \sum_{h=1}^{N} \frac{x_{hi}}{n_h} = \sum_{j=1}^{L} p(j,i) = p(+,i) \qquad (3.3\text{-}1)$$

is an unbiased estimate of the probability that a random subject is assigned to category i by a random observer. The expected proportion

$$q(i,j) = p(i,+)p(+,j) \qquad (3.3\text{-}2)$$

is an estimate of the probability under independence that a random subject is assigned to category i by the first and to category j by the second of two random observers.

In section 3.2 it was explained how measures of agreement can be computed from the proportions p(i,j) and q(i,j) in a straightforward manner.

Table 3.3-2 shows the proportions $p(i,j)$, $q(i,j)$, $p(i,+)$ and $p(i|i) = p(i,i)/p(i,+)$. Given that a psychiatrist assigned a patient to category 1, there is a rather small estimated conditional probability $p(1|1) = .35$ that a second psychiatrist also assigns this patient to category 1. The same applies to category 2. One cause is that the marginal proportions $p(1,+)$ and $p(2,+)$ are small. Another, more interesting, cause is that category 4 is hard to distinguish from categories 1 and 2: $p(1,4)$ and $p(2,4)$ are relatively large and even not smaller than what is to be expected under independence. Combining

TABLE 3.3-2
Agreement Between Two Random Psychiatrists:
Observed and Expected (Under Independence) Proportions
of Patients Classified into Five Diagnostic Categories

Category by the One Psychiatrist		Category by the Other Psychiatrist					
		1	2	3	4	5	Total
1	Observed	.05	.01	.02	.04	.02	.14
	Expected	.02	.02	.02	.04	.03	
2	Observed	.01	.05	.01	.05	.02	.14
	Expected	.02	.02	.02	.04	.03	
3	Observed	.02	.01	.10	.00	.03	.17
	Expected	.02	.02	.03	.05	.04	
4	Observed	.04	.05	.00	.19	.01	.31
	Expected	.04	.04	.05	.09	.07	
5	Observed	.02	.02	.03	.01	.16	.24
	Expected	.03	.03	.04	.07	.06	
Total		.14	.14	.17	.31	.24	1.00
Conditional		.35	.35	.60	.63	.67	

Category 1: Depression
Category 2: Personality Disorder
Category 3: Schizophrenia
Category 4: Neurosis
Category 5: Other
Kappa = .43 with Standard Error .06

categories 1, 2 and 4 significantly increases kappa from .43 to .57 ($z = 2.79$, $P_1 < .003$; see section 1.9). Even if we take into account that there are ten different subsets of three categories, and that the statistical test was guided by the data, there remains a strong indication that the degree of agreement can be increased by combining categories 1, 2 and 4. According to the theorem in section 1.3, and its proof, this implies that

$$\frac{p(1,2) + p(1,4) + p(2,4)}{q(1,2) + q(1,4) + q(2,4)} ,$$

the ratio of observed to chance expected disagreement among these

TABLE 3.3-3
Agreement Between Two Random Psychiatrists:
Observed and Expected (under Independence) Proportions
of Patients Classified into Four Diagnostic Categories
where Assignments to the "Other" Category are Omitted

Category by the One Psychiatrist		Category by the Other Psychiatrist 1	2	3	4	Total
1	Observed	.09	.01	.04	.06	.20
	Expected	.04	.04	.05	.07	
2	Observed	.01	.10	.02	.07	.20
	Expected	.04	.04	.05	.08	
3	Observed	.04	.02	.17	.00	.23
	Expected	.05	.05	.05	.09	
4	Observed	.06	.07	.00	.23	.37
	Expected	.07	.08	.09	.14	
Total		.20	.20	.23	.37	1.00
Conditional		.48	.50	.74	.63	

Category 1: Depression
Category 2: Personality Disorder
Category 3: Schizophrenia
Category 4: Neurosis
Kappa = .45 with Standard Error .07

categories, is relatively high compared to the other off diagonal cells. Taking the marginal proportions into account, it is concluded that these categories are easily confused. Of course, no psychiatrist is willing to combine these three categories in his practice because the treatment is based on the diagnostic category. The practical conclusion from the analysis above is that psychiatrists must try to distinguish more clearly between categories 1, 2 and 4; discussing patients 5, 9, 12, 19, 24 and 28 may help to improve future interobserver agreement.

Since category 5, the "Other" category, represents a very heterogeneous patient population, agreement on this category is dubious. Therefore the data were reanalyzed after assignments to category 5 were omitted. Notice that the number of psychiatrists per patient is no longer constant, and the number of patients is reduced to N=26. The resulting proportions are shown in table 3.3-3. Combining categories 1, 2 and 4 again significantly increases kappa from .45 to .66 ($z = 2.23$, $P_1 < .013$). The test result remains convincing if we take into account that there are four different subsets of three categories.

Bibliographic Notes, 3.3

The data set in table 3.3-1 was presented and analyzed in a fundamental paper by Fleiss(1971) who computed kappa and the five intraclass kappas. These data were reanalyzed by Landis and Koch(1977c), Kraemer(1980) and James(1983), but not in the way described above. The formulas for the computation of the proportions p(i,j) and q(i,j) are generalizations of formulas proposed by Schouten(1980, 1982a) who builded on Fleiss (1971); they considered a fixed number of observers per subject and in that case their definition of kappa is equivalent to the definition in the present chapter.

In order to show which categories are most easily confused by the psychiatrists, some authors previously considered the effect of combining categories. Kraemer(1980) inspected the disagreements of the paired observations in the raw data. She based her definition of the kappa coefficient on the Spearman rank correlation coefficient and the Kendall coefficient of concordance. James(1983) considered the conditional

probability that a second observer assigns a subject to category j, given that a first observer assigned the subject to category i and given that the observers disagree. In my opinion, the approach in the present section is more direct and simpler.

Landis and Koch(1977c), Fleiss and Cuzick(1979) and Fleiss(1981) considered a one-way analysis of variance model. Landis and Koch(1977c) defined indicator variables $y_{h,a,i}$, analogous to the indicator variables x(h,a,i) in the appendix to section 3.2, to indicate if the h-th subject is assigned to category i on the a-th judgement. These indicator variables are used to compute an intraclass correlation coefficient $r_{i,i}$ for each category i and an interclass correlation coefficient $r_{i,j}$ for each pair of different categories i and j. Then the overall measure of reliability r is computed as a weighted average of the L intraclass correlation coefficients. For moderately large N the kappa coefficients used by Fleiss and Cuzick(1979) and Fleiss(1981) are virtually identical to $r_{i,i}$ and r. Their kappa coefficients are identical to the kappa coefficients used in the present book if all subjects are judged by the same number of observers. They did not indicate in which way their model may be used to define weighted kappa and they did not consider the proportions p(i,j) and q(i,j).

Fleiss(1981, section 13.3) noticed that the measurement of interobserver agreement is related to the measurement of heritability in genetics. Consider N families, where there are n_h children in the h-th family. The agreement between children within a family, regarding a qualitative property, may reflect the heritability of the property; of course, depending on the property under consideration, education may play an important role.

3.4 A Property of the Parameter Upsilon

In this section it is first shown that the population parameter upsilon equals an intraclass correlation coefficient; see also the bibliographic notes to section 1.5. Then some results from psychometric test theory are treated.

Consider a population of observers and a population of subjects. From these populations random samples can be taken. Both populations are assumed to be infinite, that is there is a probability zero of taking the same observer or the same subject twice; so it does not matter whether sampling is with or without replacement. The argumentation below is of some practical importance if the two populations are not very small. For the sake of clarity: below the random variable y_{ha} refers to the populations of subjects and observers; in the present section samples are not considered.

Let y_{ha} denote the quantitative judgement on subject h by observer a and let s_h denote the expected judgement on subject h by a randomly selected observer; so $o_{ha} = y_{ha} - s_h$ has expectation zero. Notice that $y_{ha} = s_h + o_{ha}$. Given that a particular subject h (considered fixed in this sentence) is judged separately by two randomly selected observers a and b, the resulting judgements y_{ha} and y_{hb} are conditionally independent and so $o_{ha} = y_{ha} - s_h$ and $o_{hb} = y_{hb} - s_h$ are conditionally independent.

The squared difference $(y_{ha} - y_{hb})^2$ of the judgements y_{ha} and y_{hb} on the same randomly selected subject by two randomly selected observers has expectation

$$\begin{aligned} \delta(y) &= \mathrm{var}(y_{ha}) + \mathrm{var}(y_{hb}) - 2\mathrm{cov}(y_{ha}, y_{hb}) \\ &= 2\mathrm{var}(y_{ha}) - 2\mathrm{cov}(s_h + o_{ha}\ ,\ s_h + o_{hb}) \\ &= 2\mathrm{var}(y_{ha}) - 2\mathrm{var}(s_h). \end{aligned} \tag{3.4-1}$$

In contrast with the notation used in section 1.5, now var(.) and

cov(.,.) denote the population variance and the population covariance respectively. Under the null hypothesis of independence between the judgements y_{ha} and y_{hb} on the same randomly selected subject by two randomly selected observers, the squared difference $(y_{ha} - y_{hb})^2$ has expectation

$$\gamma(y) = var(y_{ha}) + var(y_{hb}) = 2var(y_{ha}). \qquad (3.4\text{-}2)$$

The parameter upsilon is

$$\upsilon(y) = \frac{\gamma(y) - \delta(y)}{\gamma(y)} = \frac{cov(y_{ha},y_{hb})}{var(y_{ha})} = \rho(y_{ha},y_{hb}), \qquad (3.4\text{-}3)$$

the Pearson product moment correlation between the judgements on the same randomly selected subject by two randomly selected observers, which is one of the many intraclass correlation coefficients. The coefficient upsilon can also be written as

$$\upsilon(y) = \frac{var(s_h)}{var(y_{ha})} , \qquad (3.4\text{-}4)$$

the ratio of the variance due to differences among subjects and the total variance.

Now, let (y_{ha},y'_{ha}) denote the bivariate judgement on subject h by observer a regarding the variables y and y', e.g. y and y' represent presence or absence of two different symptoms assessed in a patient. For the Pearson correlation $\rho(y_{ha},y'_{hb})$ between the judgements y_{ha} and y'_{hb} by two randomly selected observers who separately judge the same randomly selected subject, we have the upper bound

$$\rho(y_{ha},y'_{hb}) \leq \sqrt{(\upsilon(y)\upsilon(y'))}. \qquad (3.4\text{-}5)$$

Writing $y_{ha} = s_h + o_{ha}$ and $y'_{hb} = s'_h + o'_{hb}$, (3.4-5) directly follows from

$$cov(y_{ha},y'_{hb}) = cov(s_h + o_{ha}, s'_h + o'_{hb}) =$$

$$= \text{cov}(s_h, s'_h) \leq \sqrt{(\text{var}(s_h)\text{var}(s'_h))},$$

where it is assumed that, given that subject h is judged, o_{ha} and o'_{hb} are conditionally independent.

In formula (3.4-5) the index b may be replaced by the index a if we are prepared to assume that o_{ha} and o'_{ha} are conditionally independent given that subject h is judged. This may be quite unrealistic because it means that knowledge regarding the one variable never results in a prejudice that may influence the judgement regarding the other variable.

Bibliographic Notes, 3.4

Equality (3.4-4) and inequality (3.4-5) are well-known results from psychometric test theory; see e.g. Lord and Novick(1968, p.70). With respect to kappa, and especially regarding the case of L=2 categories, the reader is referred to the very interesting paper by Kraemer(1979); see also section 1.8.

In spite of equality (3.4-3) the coefficient upsilon and the intraclass correlation coefficient are estimated in a different way; see also the bibliographic notes to section 1.5.

Chapter 4
SAMPLING THEORY

4.1 General Remarks

In a certain sense the present chapter is a continuation of section 1.9. Some remarks made in that section will be shortly repeated below.

When the kappa statistic k is computed from the proportions p(i,j) and q(i,j), the corresponding kappa parameter κ is implicitly defined as a function of the probabilities that are estimated by the p(i,j) and q(i,j). Inferences about the parameter κ can be based on the statistic k and its estimated standard error. For the sake of clarity: in the present chapter statistical independence between judgements by observers is not assumed, unless clearly stated, and marginal proportions are not considered fixed.

However large the sample of subjects may be, it may happen that all subjects are assigned to the same category by all observers, in which case o = e = 1 and so the kappa statistic k cannot be computed; see also section 1.2 (the paragraph between formulas (1.2-4) and (1.2-5)). It is possible to define k = 0 because o = e, but it is also possible to define k = 1 because o = 1. In my opinion, the most practical solution is to consider the distribution of kappa given that at least two different categories are used. The problem demands a thorough theoretical examination, but that is not within the scope of this work.

When the sample of subjects is large, the kappa statistic k (that

can be written as a smooth function of multinomial proportions; see section 4.3) approximately follows a normal distribution with mean κ and standard deviation (the standard error of the kappa statistic) that is to be estimated using one of the four methods described in the next sections. As far as I know, the statistical performance of all four methods have not yet been compared in a simulation study regarding the standard error of kappa. So the following remarks are somewhat speculative. The method in section 4.2 is very simple but probably results in a poor estimate of the standard error. The delta method, the jackknife and the bootstrap, described in sections 4.3 to 4.5 inclusive, probably all result in satisfactory estimates of the standard error, but the jackknife is much simpler to apply than the other two methods when there are more than two observers.

The remarks in this section equally apply to kappa, upsilon and weighted kappa.

Bibliographic Notes, 4.1

Early in their invited expository paper Efron and Gong(1983) consider the standard error of a sample average $\bar{x} = (x_1 + x_2 + \ldots + x_N)/N$. In that case the standard jackknife technique results in the usual unbiased estimate

$$s_u^2(\bar{x}) = \frac{1}{N(N-1)} \sum_{h=1}^{N} (x_h - \bar{x})^2 \qquad (4.1\text{-}1)$$

of the variance of $\bar{x}$, whereas the delta and bootstrap method result in the maximum likelihood (assuming a normal distribution) estimate

$$s_{ML}^2(\bar{x}) = \frac{1}{N^2} \sum_{h=1}^{N} (x_h - \bar{x})^2 \qquad (4.1\text{-}2)$$

of the variance of $\bar{x}$.

Efron(1982, p.21) shows that, in general, the jackknife estimator of the variance has an expectation that is slightly greater than the true variance, whereas the delta method estimator of the variance may have a severe downward bias.

4.2 A Very Simple Method

Compared to the other methods, the quick method described in the present section probably results in a less satisfactory estimate of the standard error. Sampling fluctuations in the chance proportion of agreement e, or in the chance degree of agreement e(w), are neglected. As a consequence, the true standard error may be seriously overestimated. For the sake of clarity: that sampling fluctuations in e or e(w) are neglected certainly does not mean that the conditional distribution given the marginal distributions is considered. The conditional distribution of kappa or weighted kappa is rather complicated; Everitt(1968) and Hubert(1978) considered the conditional distribution under independence.

In the case of two fixed observers o is a binomial proportion with estimated variance $s^2(o) = o(1 - o)/N$, and the variance of kappa may be estimated by

$$s^2(k) = \frac{s^2(o)}{(1 - e)^2} = \frac{o(1 - o)}{N(1 - e)^2} \qquad (4.2\text{-}1)$$

The estimated standard error s(k), which is the square root of $s^2(k)$, may be used to compute an approximate confidence interval. In order to obtain an estimate of the variance of kappa under the null hypothesis of independence, o is replaced by e in the most right expression in (4.2-1); the resulting estimate is

$$s_0^2(k) = \frac{e(1 - e)}{N(1 - e)^2} = \frac{e}{N(1 - e)} \qquad (4.2\text{-}2)$$

The estimated standard error $s_0(k)$ is the square root of $s_0^2(k)$. The ratio $k/s_0(k)$ may be referred to tables of the standard normal distribution to approximately test the null hypothesis of independence between the judgements by the two observers.

The observed degree of agreement o(w) is an average. In the case of two fixed observers the variance of o(w) may be estimated by $s^2(o(w)) = (1/N)\sum_i\sum_j p(i,j)(w(i,j) - o(w))^2$, and the variance of weighted kappa may be estimated by

$$s^2(k(w)) = \frac{s^2(o(w))}{(1 - e(w))^2}$$

$$= \frac{1}{N(1 - e(w))^2} \sum_{i=1}^{L} \sum_{j=1}^{L} p(i,j)(w(i,j) - o(w))^2 \qquad (4.2\text{-}3)$$

It is easy to see that (4.2-1) is a special case of (4.2-3). In order to obtain an estimate of the variance of weighted kappa under the null hypothesis of independence, on the right hand side in (4.2-3) o(w) is replaced by e(w) and p(i,j) is replaced by q(i,j); the resulting estimate is

$$s_0^2(k(w)) = \frac{1}{N(1 - e(w))^2} \sum_{i=1}^{L} \sum_{j=1}^{L} q(i,j)(w(i,j) - e(w))^2 \qquad (4.2\text{-}4)$$

Of course, (4.2-2) is a special case of (4.2-4).

Bibliographic Notes, 4.2

Cicchetti and Fleiss(1977) investigated the performance of (4.2-3) under the null hypothesis of independence when the linear disagreement weights $v(i,j) = |i - j|$ are used (disagreement weights have been discussed in section 1.6 of this book). They concluded that application of (4.2-3) may result in a serious overestimation of the true variance; see also the bibliographic notes to the next section.

The method described above has been applied by Cohen(1960, 1968) in the case of two fixed observers, his formulas are equivalent to the four formulas above, and by Scott(1955), Fleiss(1971) and Schouten(1980) in the case of varying observers.

4.3 The Delta Method

In this section the delta method is described for the special case of a multinomial distribution. A more general description has been presented by e.g. Bishop, Fienberg and Holland(1975, chapter 14). This method has been applied by Fleiss, Cohen and Everitt(1969) and by Goodman and Kruskal(1972).

Each of a random sample of N subjects is assigned to one of ν possible response patterns. With respect to the examples in chapters 2 and 3, tables 2.1-1, 3.1-1 and 3.3-1 show the response pattern for each patient. When each subject is judged by each of n fixed observers, there are $\nu = L^n$ possible patterns; if there may be missing values, there are $\nu = (L+1)^n$ possible patterns. The expression for ν is more complicated in the case of a varying group of observers, especially if the number of observers is not the same for all subjects, but ν is a finite number on condition that there is an upper bound for the number of observers per subject. The parameter π_i denotes the probability that a randomly selected subject is assigned to the i-th response pattern; $\pi_1 + \pi_2 + \ldots + \pi_\nu = 1$. The variable p_i represents the proportion of subjects that is assigned to the i-th response pattern; $p_1 + p_2 + \ldots + p_\nu = 1$. Notice that Np_i is the number of subjects assigned to the i-th response category. The numbers $Np_1, Np_2, \ldots, Np_\nu$ follow a multinomial distribution with parameters $N, \pi_1, \pi_2, \ldots, \pi_\nu$. The random variable p_i has expectation π_i and variance

$$\mathrm{var}(p_i) = (\pi_i - \pi_i^2)/N. \qquad (4.3\text{-}1)$$

The covariance of p_i and p_j is

$$\mathrm{cov}(p_i, p_j) = -\pi_i \pi_j / N \quad \text{for } i \neq j. \qquad (4.3\text{-}2)$$

Let Σ denote the ν by ν covariance matrix of $p_1, p_2, \ldots, p_\nu$. If $f(\pi) = f(\pi_1, \pi_2, \ldots, \pi_\nu)$ is a continuous function with continuous first partial derivatives $f_i'(\pi) = \partial f(\pi)/\partial \pi_i$, then $f(p) = f(p_1, p_2, \ldots, p_\nu)$

asymptotically ($N \to \infty$) follows a normal distribution with mean $f(\pi)$ and variance $\varphi^T \Sigma \varphi$, where φ is the column vector with components $f_1'(\pi)$, $f_2'(\pi)$, ..., $f_\nu'(\pi)$ and φ^T is the transpose of φ. The quantity $\varphi^T \Sigma \varphi$ can be written as

$$\varphi^T \Sigma \varphi = \frac{1}{N} \{ \sum_{i=1}^{\nu} \pi_i (f_i'(\pi))^2 - (\sum_{i=1}^{\nu} \pi_i f_i'(\pi))^2 \}$$
$$= \frac{1}{N} \sum_{i=1}^{\nu} \pi_i (f_i'(\pi) - \tau_f)^2 \qquad (4.3\text{-}3)$$

where $\tau_f = \sum_{i=1}^{\nu} \pi_i f_i'(\pi)$.

In large samples of subjects an estimate of the variance of $f(p)$ may be taken as

$$s^2(f(p)) = \frac{1}{N} \sum_{i=1}^{\nu} p_i (f_i'(p) - t_f)^2 \qquad (4.3\text{-}4)$$

where $t_f = \sum_{i=1}^{\nu} p_i f_i'(p)$

When there are two functions f and g, in large samples of subjects an estimate of the covariance of $f(p)$ and $g(p)$ may be taken as

$$s_{1,2}(f(p),g(p)) = \frac{1}{N} \sum_{i=1}^{\nu} p_i (f_i'(p) - t_f)(g_i'(p) - t_g) \qquad (4.3\text{-}5)$$

and the test statistic

$$z = \frac{f(p) - g(p)}{\sqrt{(s^2(f(p)) + s^2(g(p)) - 2s_{1,2}(f(p),g(p)))}} \qquad (4.3\text{-}6)$$

may be referred to tables of the standard normal distribution in order to assess the statistical significance of the difference $f(p) - g(p)$. Since kappa can be written as

$$k = 1 - \frac{1 - o}{1 - e}, \qquad (4.3\text{-}7)$$

it is easy to see that

$$\frac{\partial k}{\partial p_i} = \frac{(1 - e)(\partial o/\partial p_i) - (1 - o)(\partial e/\partial p_i)}{(1 - e)^2} \qquad (4.3\text{-}8)$$

Since p_i is a discrete variable, the notation in (4.3-8) is rather loose, and

$\frac{\partial k}{\partial p_i}$ should be read as $\left| \frac{\partial k(\pi)}{\partial \pi_i} \right|_{\pi=p}$.

For convenience the loose notation is used below and in the appendix.

In the case of two fixed observers the proportions $p(i,j)$ take the place of the p_i. Accurate application of the chain rule gives

$$\frac{\partial o}{\partial p(i,j)} = \delta_{ij} = \begin{cases} 1 & \text{if } i=j \\ 0 & \text{if } i\neq j \end{cases} \qquad (4.3\text{-}9)$$

$$\frac{\partial e}{\partial p(i,j)} = m_2(i) + m_1(j) \qquad (4.3\text{-}10)$$

where $m_a(i)$ denotes the proportion of subjects assigned to category i by the a-th observer, and finally

$$\frac{\partial k}{\partial p(i,j)} = \frac{d(i,j)}{(1 - e)^2} \ , \qquad (4.3\text{-}11)$$

$$\text{where} \quad d(i,j) = (1 - e)\delta_{ij} - (1 - o)(m_2(i) + m_1(j)). \qquad (4.3\text{-}12)$$

The mean $\bar{d}$ of the $d(i,j)$ can be written as

$$\begin{aligned} \bar{d} &= \sum_{i=1}^{L} \sum_{j=1}^{L} p(i,j)d(i,j) \\ &= (1\text{-}e) \sum_{i=1}^{L} \sum_{j=1}^{L} p(i,j)\delta_{ij} - (1\text{-}o) \sum_{i=1}^{L} \sum_{j=1}^{L} p(i,j)(m_2(i) + m_1(j)) \\ &= (1\text{-}e) \sum_{i=1}^{L} p(i,i) - (1\text{-}o) \sum_{i=1}^{L} m_1(i)m_2(i) - (1\text{-}o) \sum_{j=1}^{L} m_2(j)m_1(j) \\ &= (1\text{-}e)o - (1\text{-}o)e - (1\text{-}o)e = oe - 2e + o \ . \end{aligned} \qquad (4.3\text{-}13)$$

From the foregoing it follows that, for large N, an estimate of the variance of kappa may be taken as

$$s^2(k) = \frac{1}{N(1 - e)^4} \sum_{i=1}^{L} \sum_{j=1}^{L} p(i,j)(d(i,j) - \bar{d})^2 . \qquad (4.3\text{-}14)$$

In order to obtain an estimate $s_0^2(k)$ of the variance of kappa under the null hypothesis of independence, in the foregoing formulas o is replaced by e and p(i,j) is replaced by q(i,j); the resulting formula is

$$s_0^2(k) = \frac{1}{N(1 - e)^2} \sum_{i=1}^{L} \sum_{j=1}^{L} q(i,j)(\delta_{ij} - m_2(i) - m_1(j) + e)^2 . \qquad (4.3\text{-}15)$$

The variance of weighted kappa, in the case of two fixed observers, is presented below without derivation. For the a-th observer the statistics

$$w_a(j) = \sum_{i=1}^{L} m_a(i)w(i,j) \qquad (4.3\text{-}16)$$

are defined. It is easy to see that

$$\sum_{j=1}^{L} m_1(j)w_2(j) = \sum_{j=1}^{L} m_2(j)w_1(j) = e(w).$$

The statistic

$$d(w;i,j) = (1 - e(w))w(i,j) - (1 - o(w))(w_2(i) + w_1(j)) \qquad (4.3\text{-}17)$$

is a generalization of (4.3-12), and

$$\bar{d}(w) = \sum_{i=1}^{L} \sum_{j=1}^{L} p(i,j)d(w;i,j)$$
$$= o(w)e(w) - 2e(w) + o(w) \qquad (4.3\text{-}18)$$

is a generalization of (4.3-13). In large samples of subjects an estimate of the variance of weighted kappa may be taken as

$$s^2(k(w)) = \frac{1}{N(1 - e(w))^4} \sum_{i=1}^{L} \sum_{j=1}^{L} p(i,j)(d(w;i,j) - \bar{d}(w))^2, \quad (4.3\text{-}19)$$

and an estimate of the covariance between two weighted kappa statistics $k(w_1)$ and $k(w_2)$ may be taken as

$$s_{1,2}(k(w_1),k(w_2)) = \quad (4.3\text{-}20)$$

$$= \frac{\sum_{i=1}^{L} \sum_{j=1}^{L} p(i,j)(d(w_1;i,j) - \bar{d}(w_1))(d(w_2;i,j) - \bar{d}(w_2))}{N(1 - e(w_1))^2(1 - e(w_2))^2} .$$

Under the null hypothesis of independence an estimate of the variance of weighted kappa may be taken as

$$s_0^2(k(w)) = \quad (4.3\text{-}21)$$

$$= \frac{1}{N(1 - e(w))^2} \sum_{i=1}^{L} \sum_{j=1}^{L} q(i,j)(w(i,j) - w_2(i) - w_1(j) + e(w))^2 .$$

In the appendix I derive formulas to estimate the variance of kappa and weighted kappa in the case of varying observers, under dependence as well as under independence. Such formulas make it possible to compare the delta method with other methods.

Bibliographic Notes, 4.3

The delta method has been described in several books; see e.g. Bishop, Fienberg and Holland(1975, section 14.6). This method should be applied in a careful way in order to avoid walking into the trap described by Goodman and Kruskal(1972, section 6) in a cautionary note. Their note may be summarized as follows. As an example consider a binomial trial. The proportions of successes and failures are denoted by p_1 and p_2; $p_1 + p_2 = 1$. The functions $f(p_1,p_2) = p_1/p_2$ and $g(p_1,p_2) = p_1/(1 - p_1)$ are numerically equal, but their partial derivatives are completely different; $f(p_1,p_2) = g(p_1,p_2)$, but $\partial f(p_1,p_2)/\partial p_2 = -p_1/p_2^2$ whereas $\partial g(p_1,p_2)/\partial p_2 = 0$. "Which way an expression is written makes no difference (except for convenience of computation) in the final asymptotic variance, *provided that* the same symbolic functional form is used throughout in finding derivatives. If not, incorrect results may be obtained."

Fleiss(1982) proved that $t_f = 0$ on condition that the function f is (re)written in such a way that $f(p_1,p_2,\ldots,p_v) = f(cp_1,cp_2,\ldots,cp_v)$ for all nonzero c, but this does not always simplify application of the delta method. Notice that, regarding the functions in the preceding paragraph and assuming that $0 < p_1 < 1$, $f(p_1,p_2) = p_1/p_2 = (cp_1)/(cp_2) = f(cp_1,cp_2)$ for all nonzero c, whereas $g(p_1,p_2) = p_1/(1 - p_1) \neq 2p_1/(1 - 2p_1) = g(2p_1,2p_2)$. In my experience, it often is convenient to write the function under consideration in such a way that partial derivatives are most easily obtained.

Two Fixed Observers

For the case of two fixed observers Fleiss, Cohen and Everitt (1969) derived the approximate large sample variances of kappa and weighted kappa. Formulas (4.3-14), (4.3-15), (4.3-19) and (4.3-21) are equivalent to their formulas, although my notation is different. Whether the marginal totals in the two-way table are fixed or variable, the same formula may be used to estimate the variance of weighted kappa under independence; see Hubert(1978). Cicchetti and Fleiss(1977) investigated the performance of (4.2-3), although they divided by N-1 instead of N, and (4.3-21) under the null hypothesis of independence

when the linear disagreement weights $v(i,j) = |i - j|$ are used (disagreement weights have been discussed in section 1.6 of this book). They concluded that (4.3-21) is preferable to (4.2-3). Cicchetti and Fleiss (1977) and Cicchetti(1981) further concluded that $N \geq 2L^2$ is required for a valid application of (4.3-21) when the linear disagreement weights $v(i,j) = |i - j|$ are used. Fleiss and Cicchetti(1978) concluded that $N \geq 16L^2$ is required for a valid application of (4.3-19) to obtain a confidence interval for weighted kappa when the linear disagreement weights $v(i,j) = |i - j|$ are used; for L=2 this means a minimum sample size $16L^2 = 64$, while for L=5 the minimum sample size is $16L^2 = 400$.

Many Fixed Observers

In the case of more than two fixed observers, variance formulas are available if there are no missing values. Davies and Fleiss(1982) presented formulas to estimate the variance under independence if there are only two categories. Schouten(1982b) presented formulas to estimate the variance of weighted kappa under dependence and under independence.

Varying Observers

For the case of varying observers, but with a constant number of observers per subject, Fleiss, Nee and Landis(1979) derived the estimated variance of kappa under independence, and Schouten(1982a) derived the estimated variance of weighted kappa under dependence and under independence. For the case of two categories Fleiss and Cuzick (1979) derived formulas to estimate the variance of kappa under independence when the number of observers is not the same for all subjects; since they used an analysis of variance model, their definition of kappa is different from my definition in section 3.3. The approach by Fleiss and Cuzick(1979) has also been used by Fleiss(1981, section 13.2). In the appendix after this chapter I derive formulas to estimate the variance of kappa and weighted kappa in the case of varying observers, under dependence as well as under independence, where the number of observers is not required to be constant.

4.4 The Grouped Jackknife

In applying the standard jackknife, that was described in section 1.9, N times a single subject is deleted from the sample of N subjects. In applying the grouped jackknife several times a group of subjects is deleted from the sample. The grouped jackknife requires considerably less computational effort, but may result in less stable variance estimates.

When the number of subjects in the sample can be written as $N = GH$ for integers G and H, the sample of subjects can be divided into G consecutive groups of H subjects. Let y be the statistic (such as kappa) computed from the complete sample of N subjects and let $y_{(-g)}$ be the value of the statistic when the g-th group of H subjects is deleted from the sample. The g-th pseudovalue is computed as

$$y^{(g)} = Gy - (G-1)y_{(-g)} \qquad (4.4\text{-}1)$$

The jackknife estimate

$$y^{(\cdot)} = \frac{1}{G} \sum_{g=1}^{G} y^{(g)} \qquad (4.4\text{-}2)$$

has a slightly smaller bias than y. Although the true varianees $\sigma^2(y^{(\cdot)})$ and $\sigma^2(y)$ are not exactly equal, the same estimate, namely

$$s^2(y^{(\cdot)}) = s^2(y) = \frac{1}{G(G-1)} \sum_{g=1}^{G} (y^{(g)} - y^{(\cdot)})^2 \qquad (4.4\text{-}3)$$

is used to estimate both $\sigma^2(y^{(\cdot)})$ and $\sigma^2(y)$.

Bibliographic Notes, 4.4

Let μ denote the population parameter that is estimated by y and $y^{(\cdot)}$. If the group size H is very large, then $(y^{(\cdot)} - \mu)/s(y^{(\cdot)})$ approximately follows the Student distribution with G-1 degrees of freedom, even if G is small; see Miller(1974) and references therein.

4.5 The Bootstrap

The bootstrap makes full use of the inherent variation in a random sample. It is a good method to estimate the variance of a statistic, on condition that a reliable pseudo random number generator is used and a sufficiently large number of bootstrap samples is taken; the meaning of this sentence becomes clear below.

The statistic y is computed for a sample of N subjects and the variance of y is to be estimated. All methods make the assumption that the sample of N subjects reflects the true probability distribution, but the bootstrap uses this assumption in the most direct way. The sample of N subjects is treated as a population from which random samples can be taken, in order to investigate sampling fluctuations. A *bootstrap sample* consists of N subjects chosen at random and *with replacement* from the original N subjects; so some of the original N subjects are chosen more than once. A pseudo random number r between 0 and 1 may be generated to choose a subject: the h-th subject is chosen if $h-1 \leq Nr < h$. In this way B bootstrap samples are taken, where e.g. B = 200 or B = 1000; it is not quite clear how large B must be. From the b-th bootstrap sample the *bootstrap replication* $y_{(b)}$ is computed. An estimate of the variance of y may be taken as

$$s^2(y) = \frac{1}{B-1} \sum_{b=1}^{B} (y_{(b)} - y_{(.)})^2 \qquad (4.5\text{-}1)$$

$$\text{with } y_{(.)} = \frac{1}{B} \sum_{b=1}^{B} y_{(b)}$$

The bootstrap estimate of the variance is obtained if B becomes infinitely large.

Bibliographic Notes, 4.5

Efron(1979) invented the bootstrap. Efron and Gong(1983) recommended the bootstrap if one had the disposal of sufficient computational power. In their opinion the "jackknife is the method of choice if one does not want to do the bootstrap computations." Diaconis and Efron(1983) explained the bootstrap to non-statisticians as a very general tool for investigating sampling fluctuations.

Reliable pseudo random number generators can be found in Van Es, Gill and Van Putten(1983), and in Wichmann and Hill(1982). The first of these papers contains the warning that "the random number generators in manufacturers' statistical packages are sometimes of surprisingly bad quality."

4.6 Future Research: Simulation Experiments

It needs to be investigated whether large or small differences in statistical performance exist between the methods described in the previous sections. The final choice will also depend on the conveniences and inconveniences of a method with respect to application and implementation in a computer program.

It is not difficult to simulate certain multinomial distributions representing the behaviour of two or more fixed or varying observers. In simulation experiments attention may be paid to the following things.

i) the average estimated standard error of (weighted) kappa compared to the true standard error
ii) the coefficient of variation of the estimated standard error
iii) the bias in (weighted) kappa
iv) the coverage probability of one-sided and two-sided confidence intervals; in general the distribution of (weighted) kappa will be markedly skewed.
v) the test size when comparing two dependent or independent statistics
vi) the test size when testing the null hypothesis of independence; this point is related to (iv).

To avoid situations wherein an agreement statistic cannot be computed, it was decided in section 4.1 to consider the distribution given that at least two categories are used in the whole experiment. When it happens, in the course of a simulation study, that only one category is used, the corresponding sample should be removed. Notice, however, that the probability distribution of (weighted) kappa also depends on this decision.

Suppose that all subjects but subject h_0 are assigned to the same category by all observers. Since the statistic kappa cannot be computed if subject h_0 is deleted from the sample of N subjects, the standard jackknife technique does not yield an estimate of the standard error of kappa, as was also noted in section 1.9. When this happens in a simulation study, I propose to take the jackknife variance estimate equal to its most important competitor the delta variance estimate. This

proposal ensures that differences found between the distributions of these two variance estimates are not caused by the occurrence of an irregular event, while the distribution of the delta variance estimate remains unaltered.

APPENDIX

The Delta Estimate of the Variance in the Case of Varying Observers

A1 Introduction

In this appendix the delta method is applied to obtain an estimate of the variance of kappa and weighted kappa in the case of varying observers, where the number of observers per subject is considered a random variable. For each of the N subjects anew a random sample of observers is taken from a large population of observers; in the mathematical model the population of observers is assumed to be infinitely large. The notation is summarized below.

From the n_h observers who judge the h-th subject x_{hi} observers use category i. As was explained in section 4.3, it is necessary to assume that

$$n_h = \sum_{i=1}^{L} x_{hi} \quad \text{is smaller than a certain upper bound.}$$

As was explained in sections 3.2 and 3.3, the observed proportion p(i,j) is the average proportion of ordered pairs of observers where the first observer uses category i and the second observer uses category j:

$$p(i,i) = \frac{1}{N} \sum_{h=1}^{N} \frac{x_{hi}(x_{hi} - 1)}{n_h(n_h - 1)} , \qquad \text{(A1-1)}$$

$$p(i,j) = p(j,i) = \frac{1}{N} \sum_{h=1}^{N} \frac{x_{hi}x_{hj}}{n_h(n_h - 1)} \quad \text{for } i \neq j . \qquad \text{(A1-2)}$$

The marginal proportion $p(i,+)$ is the average proportion of observers who use category i:

$$p(i,+) = p(+,i) = \sum_{j=1}^{L} p(i,j) = \frac{1}{N} \sum_{h=1}^{N} \frac{x_{hi}}{n_h} \quad . \qquad \text{(A1-3)}$$

The chance proportion $q(i,j)$ is computed as

$$q(i,j) = q(j,i) = p(i,+)p(+,j) \quad . \qquad \text{(A1-4)}$$

A2 Variance of Weighted Kappa in the Case of Varying Observers

With reference to the agreement weights w(i,j), the observed and chance degree of agreement and weighted kappa are computed according to the following formulas:

$$o(w) = \sum_{i=1}^{L} \sum_{j=1}^{L} p(i,j)w(i,j) \ , \qquad \text{(A2-1)}$$

$$e(w) = \sum_{i=1}^{L} \sum_{j=1}^{L} q(i,j)w(i,j) \ , \qquad \text{(A2-2)}$$

$$k(w) = \frac{o(w) - e(w)}{1 - e(w)} \ . \qquad \text{(A2-3)}$$

Below explicit formulas are presented to estimate the variance of weighted kappa when the number of subjects is large. In section A4 a mathematical derivation of these formulas is given.
For category i define the statistic

$$\overline{w}(i) = \sum_{j=1}^{L} p(+,j)w(i,j) = \sum_{j=1}^{L} p(j,+)w(j,i) \qquad \text{(A2-4)}$$

and note that $\sum_{i=1}^{L} p(i,+)\overline{w}(i) = \sum_{i=1}^{L} p(+,i)\overline{w}(i) = e(w)$.

For the h-th subject in the sample, define the statistics

$$o'_h = \frac{1}{n_h(n_h - 1)} \sum_{i=1}^{L} x_{hi} \left(\sum_{j=1}^{L} x_{hj} w(i,j) - w(i,i) \right) \ , \qquad \text{(A2-5)}$$

$$e'_h = \frac{2}{n_h} \sum_{i=1}^{L} x_{hi} \overline{w}(i) \ , \qquad \text{(A2-6)}$$

$$d_h = d_h(w) = (1 - e(w))o'_h - (1 - o(w))e'_h \ . \qquad \text{(A2-7)}$$

From the formulas

$$\frac{1}{N} \sum_{h=1}^{N} o'_h = \sum_{i=1}^{L} \sum_{j=1}^{L} p(i,j)w(i,j) = o(w) \ , \qquad \text{(A2-8)}$$

$$\frac{1}{N}\sum_{h=1}^{N} e'_h = \sum_{i=1}^{L} 2p(i,+)\bar{w}(i) = 2e(w) , \qquad \text{(A2-9)}$$

it follows that

$$\bar{d}(w) = \frac{1}{N}\sum_{h=1}^{N} d_h(w) = (1 - e(w))o(w) - 2(1 - o(w))e(w) \qquad \text{(A2-10)}$$

$$= o(w)e(w) - 2e(w) + o(w) .$$

For large N an estimate of the variance of weighted kappa may be taken as

$$s^2(k(w)) = \frac{1}{N^2(1 - e(w))^4}\sum_{h=1}^{N} (d_h(w) - \bar{d}(w))^2 . \qquad \text{(A2-11)}$$

If two sets of agreement weights, say $w_1(i,j)$ and $w_2(i,j)$, are used, the chi square test statistic on one degree of freedom

$$\text{chi square} = \frac{(k(w_1) - k(w_2))^2}{s^2(k(w_1)) + s^2(k(w_2)) - 2s_{1,2}(k(w_1),k(w_2))} \qquad \text{(A2-12)}$$

may be computed to compare $k(w_1)$ and $k(w_2)$. In the last formula

$$s_{1,2}(k(w_1),k(w_2)) = \frac{\sum_{h=1}^{N} (d_h(w_1) - \bar{d}(w_1))(d_h(w_2) - \bar{d}(w_2))}{N^2(1 - e(w_1))^2(1 - e(w_2))^2} \qquad \text{(A2-13)}$$

is the estimated covariance of $k(w_1)$ and $k(w_2)$.

Under the null hypothesis that assignments by different observers judging the same subject are independently distributed, for large N an estimate of the variance of weighted kappa may be taken as

$$s_0^2(k(w)) = \frac{2n_0}{N(1 - e(w))^2}\sum_{i=1}^{L}\sum_{j=1}^{L} q(i,j)(w(i,j) - \bar{w}(i) - \bar{w}(j) + e(w))^2 \qquad \text{(A2-14)}$$

$$\text{with } n_0 = \frac{1}{N}\sum_{h=1}^{N} \frac{1}{n_h(n_h - 1)}$$

Warning

The asymptotic variance formulas should be applied with caution. In many practical situations the number of possible patterns $(x_{h1}, x_{h2}, \ldots, x_{hL})$ will be rather large and many patterns will not even be observed, especially if the number of observers varies from subject to subject (if possible, the number of observers should be kept constant). As a consequence, only a very rough estimate of the true variance is obtained, and in many situations the estimated variance is only an indication of the order of magnitude of the true variance.

Bibliographic Notes, A2

When the number of observers per subject is constant, formulas (A2-11) and (A2-14) reduce to formulas (13) and (14) in Schouten(1982a); see also the bibliographic notes to sections 4.3, A3 and A4.

A3 Variance of Kappa in the Case of Varying Observers

The observed and chance proportion of agreement and kappa are computed according to the following formulas:

$$o = \sum_{i=1}^{L} p(i,i) \ , \tag{A3-1}$$

$$e = \sum_{i=1}^{L} q(i,i) \ , \tag{A3-2}$$

$$k = \frac{o - e}{1 - e} \ . \tag{A3-3}$$

Below explicit formulas are presented to estimate the variance of kappa when the number of subjects is large. These formulas are special cases of the formulas in section A2. All variance formulas are derived in section A4.

For the h-th subject in the sample, define the statistics

$$o'_h = \frac{1}{n_h(n_h - 1)} \sum_{i=1}^{L} x_{hi}(x_{hi} - 1) \ , \tag{A3-4}$$

$$e'_h = \frac{2}{n_h} \sum_{i=1}^{L} x_{hi} p(i,+) \ , \tag{A3-5}$$

$$d_h = (1 - e)o'_h - (1 - o)e'_h \ , \tag{A3-6}$$

$$\bar{d} = \frac{1}{N} \sum_{h=1}^{N} d_h = oe - 2e + o \ . \tag{A3-7}$$

For large N an estimate of the variance of kappa may be taken as

$$s^2(k) = \frac{1}{N^2(1 - e)^4} \sum_{h=1}^{N} (d_h - \bar{d})^2 \ . \tag{A3-8}$$

Under the null hypothesis that assignments by different observers judging the same subject are independently distributed, for large N an estimate of the variance of kappa may be taken as

$$s_0^2(k) = \frac{2n_0}{N(1-e)^2} \times \left(e + e^2 - 2\sum_{i=1}^{L} p(i,+)^3\right) \qquad \text{(A3-9)}$$

$$\text{with} \quad n_0 = \frac{1}{N}\sum_{h=1}^{N} \frac{1}{n_h(n_h - 1)} \quad .$$

When there are only two categories, formula (A3-9) reduces to

$$s_0^2(k) = 2n_0/N \quad \text{if } L=2. \qquad \text{(A3-10)}$$

It is surprising that the last expression does not depend on the marginal proportion $p(1,+)$.

Bibliographic Notes, A3

Fleiss, Nee and Landis(1979) directly derived the variance of kappa under independence for the case of a constant number of observers per subject, without first deriving the variance under dependence; see also the bibliographic notes to section 4.3. When the number of observers per subject is constant, formulas (A3-9) and (A3-10) reduce to formulas (12) and (13) in Fleiss, Nee and Landis(1979).

A4 Derivation of Variance Formulas

Let $p(x) = p(x_1,x_2,\ldots,x_L)$ denote the proportion of subjects assigned to category 1 by x_1 observers, to category 2 by x_2 observers, ..., and to category L by x_L observers; $\sum_x p(x) = 1$. Since the number of observers judging a subject, that is

$$n = \sum_{i=1}^{L} x_i \ , \tag{A4-1}$$

is not fixed beforehand, it is not the same for all patterns $x = (x_1,x_2,\ldots,x_L)$ with $p(x) = p(x_1,x_2,\ldots,x_L) > 0$.

The first three formulas in section A1 can be rewritten as

$$p(i,i) = \sum_x p(x) \frac{x_i(x_i - 1)}{n(n - 1)} \ , \tag{A4-2}$$

$$p(i,j) = \sum_x p(x) \frac{x_i x_j}{n(n - 1)} \qquad \text{for } i \neq j, \tag{A4-3}$$

$$p(i,+) = p(+,i) = \sum_x p(x) \frac{x_i}{n} \ . \tag{A4-4}$$

Since $p(x)$ is a multinomial proportion, the delta method may be applied. The weighted kappa statistic $k(w)$ asymptotically follows a normal distribution. According to section 4.3, in a large sample of subjects an estimate of the variance of weighted kappa may be taken as

$$s^2(k(w)) = \frac{1}{N} \sum_x p(x)\left(\frac{\partial k(w)}{\partial p(x)} - t \right)^2 \tag{A4-5}$$

$$\text{with} \quad t = \sum_x p(x) \frac{\partial k(w)}{\partial p(x)} \ .$$

Since $k(w) = 1 - \frac{1 - o(w)}{1 - e(w)}$, we have

$$\frac{\partial k(w)}{\partial p(x)} = \frac{(1 - e(w)) \frac{\partial o(w)}{\partial p(x)} - (1 - o(w)) \frac{\partial e(w)}{\partial p(x)}}{(1 - e(w))^2} \ . \tag{A4-6}$$

Since the observed degree of agreement can be rewritten as

$$o(w) = \sum_{x} \frac{p(x)}{n(n-1)} \sum_{i=1}^{L} x_i \Big(\sum_{j=1}^{L} x_j w(i,j) - w(i,i) \Big),$$

it is immediately clear that

$$o'(x,w) = \frac{\partial o(w)}{\partial p(x)} = \frac{1}{n(n-1)} \sum_{i=1}^{L} x_i \Big(\sum_{j=1}^{L} x_j w(i,j) - w(i,i) \Big). \qquad (A4\text{-}7)$$

Since $\frac{\partial p(i,+)}{\partial p(x)} = \frac{\partial p(+,i)}{\partial p(x)} = \frac{x_i}{n}$, we further have

$$e'(x,w) = \frac{\partial e(w)}{\partial p(x)} = \sum_{i=1}^{L} \sum_{j=1}^{L} \Big(\frac{x_i}{n} p(+,j) w(i,j) + p(i,+) \frac{x_j}{n} w(i,j) \Big)$$

$$= \frac{2}{n} \sum_{i=1}^{L} x_i \overline{w}(i), \qquad (A4\text{-}8)$$

where it is used that $w(i,j) = w(j,i)$ and $p(i,+) = p(+,i)$.
Define further

$$d(x,w) = (1 - e(w))o'(x,w) - (1 - o(w))e'(x,w). \qquad (A4\text{-}9)$$

Notice that, analogous to (A2-10),

$$\overline{d}(w) = \sum_{x} p(x) d(x,w) = o(w)e(w) - 2e(w) + o(w). \qquad (A4\text{-}10)$$

From the formulas (A4-5) to (A4-10) inclusive it follows that

$$\frac{\partial k(w)}{\partial p(x)} = \frac{d(x,w)}{(1 - e(w))^2} \quad \text{and} \quad t = \frac{\overline{d}(w)}{(1 - e(w))^2}. \qquad (A4\text{-}11)$$

For large N an estimate of the variance of weighted kappa may be taken as

$$s^2(k(w)) = \frac{1}{N(1 - e(w))^4} \sum_{x} p(x)(d(x,w) - \overline{d}(w))^2. \qquad (A4\text{-}12)$$

When, for the h-th subject in the sample, $x_{h1} = x_1$, $x_{h2} = x_2$, ..., $x_{hL} = x_L$, it is easy to see that $o'_h = o'(x,w)$, $e'_h = e'(x,w)$ and $d_h(w) = d(x,w)$. It follows that formulas (A2-11) and (A4-12) are equivalent. The derivation of the covariance formula is analogous to the derivation of the variance formula.

If $w(i,j) = 0$ for $i \neq j$, and $w(i,i) = 1$, we have $\overline{w}(i) = p(i,+)$ and formulas (A3-4) to (A3-8) inclusive follow from the formulas (A2-5) to (A2-11) inclusive.

Below the variance under independence is derived from the preceding formulas. The proportion $p(x)$ in (A4-12) can be written as

$$p(x) = p(x_1,x_2,\ldots,x_L) = p(n)p(x_1,x_2,\ldots,x_L \mid n) = p(x|n), \qquad \text{(A4-13)}$$

where $p(n)$ denotes the proportion of subjects judged by $n = \sum_{i=1}^{L} x_i$ observers; the conditional proportion $p(x|n)$ is defined as

$$p(x|n) = \frac{p(x)}{p(n)} \quad \text{if} \quad n = \sum_{i=1}^{L} x_i$$

$$= 0 \quad \text{otherwise.}$$

Formula (A4-12) can be rewritten as

$$N(1 - e(w))^4 s^2(k(w)) = \sum_n \sum_x p(n)p(x|n) \times \qquad \text{(A4-14)}$$

$$\times \{(1 - e(w))o'(x,w) - (1 - o(w))e'(x,w) - (o(w)e(w) - 2e(w) + o(w))\}^2$$

In order to obtain the estimate $s_0^2(k(w))$ of the variance of weighted kappa under the null hypothesis of independence, estimates in (A4-14) are replaced by the corresponding estimates under independence as indicated below. The conditional proportion

$p(x|n) = p(x_1,x_2,\ldots,x_L|n)$ is replaced by $p(i_1,+)p(i_2,+)\cdots p(i_n,+)$,

which is the estimate of the probability under independence that a randomly selected subject is assigned to category i_1 by the first, to category i_2 by the second, ..., and to category i_n by the last of n randomly selected observers: at the same time

$$\sum_n \sum_x \text{ is replaced by } \sum_n \sum_{i_1=1}^{L} \sum_{i_2=1}^{L} \cdots \sum_{i_n=1}^{L} :$$

summation over all possible patterns $(n,i_1,i_2,\ldots,i_n)$. Further,

$o(w)$ is replaced by $e(w)$,

$o'(x,w)$ is replaced by $\frac{1}{n(n-1)} \sum_{a=1}^{n} \sum_{\substack{b=1 \\ b \neq a}}^{n} w(i_a,i_b)$,

$e'(x,w)$ is replaced by $\frac{2}{n} \sum_{a=1}^{n} \overline{w}(i_a)$.

So the estimate $s_0^2(k(w))$ of the variance of weighted kappa under independence can be written, after multiplication by $N(1 - e(w))^2$, as

$$N(1 - e(w))^2 s_0^2(k(w)) =$$

$$= \sum_{n} p(n) \sum_{i_1=1}^{L} \sum_{i_2=1}^{L} \cdots \sum_{i_n=1}^{L} p(i_1,+)p(i_2,+)\cdots p(i_n,+) \times \tag{A4-15}$$

$$\times \{ \frac{1}{n(n-1)} \sum_{a=1}^{n} \sum_{\substack{b=1 \\ b \neq a}}^{n} w(i_a,i_b) - \frac{2}{n} \sum_{a=1}^{n} \overline{w}(i_a) + e(w) \}^2 .$$

Since we have the equality

$$\frac{2}{n} \sum_{a=1}^{n} \overline{w}(i_a) = \frac{1}{n(n-1)} \sum_{a=1}^{n} \sum_{\substack{b=1 \\ b \neq a}}^{n} (\overline{w}(i_a) + \overline{w}(i_b)) ,$$

and the triviality $e(w) = \frac{1}{n(n-1)} \sum_{a=1}^{n} \sum_{\substack{b=1 \\ b \neq a}}^{n} e(w)$,

formula (A4-15) can be rewritten as

$$N(1 - e(w))^2 s_0^2(k(w)) =$$

$$= \sum_{n} p(n) \sum_{i_1=1}^{L} \sum_{i_2=1}^{L} \cdots \sum_{i_n=1}^{L} p(i_1,+)p(i_2,+)\cdots p(i_n,+) \times \tag{A4-16}$$

$$\times \{ \frac{1}{n(n-1)} \sum_{a=1}^{n} \sum_{\substack{b=1 \\ b \neq a}}^{n} (w(i_a,i_b) - \overline{w}(i_a) - \overline{w}(i_b) + e(w)) \}^2 =$$

$$= \sum_n p(n) \sum_{i_1=1}^{L} \sum_{i_2=1}^{L} \cdots \sum_{i_n=1}^{L} p(i_1,+)p(i_2,+)\cdots p(i_n,+) \times$$

$$\times \{ \frac{2}{n(n-1)} \sum_{a=1}^{n-1} \sum_{b=a+1}^{n} (w(i_a,i_b) - \bar{w}(i_a) - \bar{w}(i_b) + e(w)) \}^2 .$$

Since, for instance,

$$\sum_{i_1=1}^{L} \sum_{i_2=1}^{L} \sum_{i_3=1}^{L} p(i_1,+)p(i_2,+)p(i_3,+) \times$$

$$\times \{ w(i_1,i_2)-\bar{w}(i_1)-\bar{w}(i_2)+e(w) \} \{ w(i_1,i_3)-\bar{w}(i_1)-\bar{w}(i_3)+e(w) \} = 0 ,$$

and

$$\sum_{i_1=1}^{L} \sum_{i_2=1}^{L} \sum_{i_3=1}^{L} \sum_{i_4=1}^{L} p(i_1,+)p(i_2,+)p(i_3,+)p(i_4,+) \times$$

$$\times \{ w(i_1,i_2)-\bar{w}(i_1)-\bar{w}(i_2)+e(w) \} \{ w(i_3,i_4)-\bar{w}(i_3)-\bar{w}(i_4)+e(w) \} = 0 ,$$

it follows that

$$N(1 - e(w))^2 s_0^2(k(w)) =$$

$$= \sum_n p(n) \sum_{i_1=1}^{L} \sum_{i_2=1}^{L} \cdots \sum_{i_n=1}^{L} p(i_1,+)p(i_2,+)\cdots p(i_n,+) \times \qquad \text{(A4-17)}$$

$$\times \{ \frac{4}{n^2(n-1)^2} \sum_{a=1}^{n-1} \sum_{b=a+1}^{n} (w(i_a,i_b) - \bar{w}(i_a) - \bar{w}(i_b) + e(w))^2 \}$$

$$= \sum_n \frac{2p(n)}{n(n-1)} \sum_{i=1}^{L} \sum_{j=1}^{L} p(i,+)p(j,+)(w(i,j) - \bar{w}(i) - \bar{w}(j) + e(w))^2 .$$

It is easy to see that the last formula is equivalent to (A2-14) in section A2; since $Np(n)$ is the number of subjects judged by n observers, $\sum_n (Np(n))/(n(n-1))$ equals Nn_0.

The equality

$$\sum_{i=1}^{L}\sum_{j=1}^{L} p(i,+)p(j,+)(w(i,j) - \overline{w}(i) - \overline{w}(j) + e(w))^2 = \tag{A4-18}$$

$$= \sum_{i=1}^{L}\sum_{j=1}^{L} p(i,+)p(j,+)w(i,j)^2 - 2\sum_{i=1}^{L} p(i,+)\overline{w}(i)^2 + e(w)^2$$

is used to prove that, if $w(i,j) = 0$ for $i \neq j$ and $w(i,i) = 1$,

$$N(1 - e)^2 s_0^2(k) = \tag{A4-19}$$

$$= \sum_{n} \frac{2p(n)}{n(n-1)} \times \left(e + e^2 - 2\sum_{i=1}^{L} p(i,+)^3 \right) ,$$

and this formula is equivalent to formula (A3-9) in section A3. Formula (A3-10) can be derived from formula (A3-9) by first proving that, for L=2, the chance proportion of agreement e can be written as $e = 1 - 2p(1,+)(1 - p(1,+))$.

Bibliographic Notes, A4

The derivation of (A2-11) is analogous to the derivation in Schouten (1982a) for the case of a constant number of observers per subject, but the derivation of (A2-14) is more elegant than the corresponding derivation in Schouten(1982a).

REFERENCES

♀ Immediately after each reference it is indicated where that reference is mentioned in this book.

Bartko JJ and Carpenter WT (1976) On the methods and theory of reliability. The Journal of Nervous and Mental Disease 163, 307-317. ♀ sections 1.1 and 1.4

Bishop YMM, Fienberg SE and Holland PW (1975) Discrete Multivariate Analysis. Massachusetts: MIT Press. ♀ section 4.3

Cicchetti DV (1981) Testing the normal approximation and minimal sample size requirements of weighted kappa when the number of categories is large. Applied Psychological Measurement 5, 101-104. ♀ section 4.3

Cicchetti DV and Fleiss JL (1977) Comparison of the null distributions of weighted kappa and the C ordinal statistic. Applied Psychological Measurement 1, 195-201. ♀ sections 1.9, 4.2 and 4.3

Cicchetti DV, Lee C, Fontana AF and Dowds BN (1978) A computer program for assessing specific category rater agreement for qualitative data. Educational and Psychological Measurement 38, 805-813. ♀ section 1.4

Cohen J (1960) A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20, 37-46. ♀ preface and sections 1.1, 1.2, 1.5 and 4.2

Cohen J (1968) Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin 70, 213-220. ♀ preface and sections 1.3, 1.5, 1.6 and 4.2

Conger AJ (1980) Integration and generalization of kappas for multiple raters. Psychological Bulletin 88, 322-328. ♀ section 2.2

Connell FA and Koepsell TD (1985) Measures of gain in certainty from a diagnostic test. American Journal of Epidemiology 121, 744-753. ♀ section 2.5

Cox DR and Hinkley DV (1974) Theoretical Statistics. London: Chapman and Hall. ♀ section 2.2

Davies M and Fleiss JL (1982) Measuring agreement for multinomial data. Biometrics 38, 1047-1051. ☞ section 4.3

Department of Clinical Epidemiology and Biostatistics, McMaster University (1980) Clinical Disagreement. Canadian Medical Association Journal 123, 499-504(I. How often it occurs and why), 613-617(II. How to avoid it and how to learn from one's mistakes). ☞ section 1.1

Department of Clinical Epidemiology and Biostatistics, McMaster University (1983) Interpretation of diagnostic data. Canadian Medical Association Journal 129, 429-432(1. How to do it with pictures), 559-565(2. How to do it with a simple table (part A)), 705-710(3. How to do it with a simple table (part B)), 832-835(4. How to do it with a more complex table), 947-954(5. How to do it with simple maths), 1093-1099(6. How to do it with more complex maths); with a correction in 1984, vol.130, p.106. ☞ section 2.5

Diaconis P and Efron B (1983) Computer-intensive methods in statistics. Scientific American 248(5), 116-130. ☞ section 4.5

Efron B (1979) Bootstrap methods: another look at the jackknife. Annals of Statistics 7, 1-26. ☞ section 4.5

Efron B (1982) The Jackknife, the Bootstrap and Other Resampling Plans. Philadelphia: Society for Industrial and Applied Mathematics. ☞ sections 1.9 and 4.1

Efron B and Gong G (1983) A leisurely look at the bootstrap, the jackknife, and cross-validation. The American Statistician 37, 36-48. ☞ sections 1.9, 4.1 and 4.5

Everitt BS (1968) Moments of the statistics kappa and weighted kappa. The British Journal of Mathematical and Statistical Psychology 21, 97-103. ☞ section 4.2

Fidler V and Nagelkerke NJD (1985) Agreement on a two-point scale. Accepted for publication in Statistica Neerlandica. ☞ section 2.4

Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychological Bulletin 76, 378-382. ☞ preface and sections 1.4, 1.7, 3.3 and 4.2

Fleiss JL (1981) Statistical Methods for Rates and Proportions. New York: Wiley, 2nd edition. ☞ sections 1.2, 1.4, 1.9, 3.3 and 4.3

Fleiss JL (1982) A simplification of the classic large-sample standard error of a function of multinomial proportions. The American Statistician 36, 377-378. ⚲ section 4.3

Fleiss JL and Cicchetti DV (1978) Inference about weighted kappa in the non-null case. Applied Psychological Measurement 2, 113-117. ⚲ section 4.3

Fleiss JL and Cohen J (1973) The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and Psychological Measurement 33, 613-619. ⚲ section 1.5

Fleiss JL, Cohen J and Everitt BS (1969) Large sample standard errors of kappa and weighted kappa. Psychological Bulletin 72, 323-327. ⚲ section 4.3

Fleiss JL and Cuzick J (1979) The reliability of dichotomous judgements: unequal numbers of judges per subject. Applied Psychological Measurement 3, 537-542. ⚲ sections 3.3 and 4.3

Fleiss JL and Davies M (1982) Jackknifing functions of multinomial frequencies, with an application to a measure of concordance. American Journal of Epidemiology 115, 841-845. ⚲ section 1.9

Fleiss JL, Nee JCM and Landis JR (1979) Large sample variance of kappa in the case of different sets of raters. Psychological Bulletin 86, 947-977. ⚲ sections 4.3 and A3

Goodman LA and Kruskal WH (1972) Measures of association for cross classifications, IV: Simplification of asymptotic variances. Journal of the American Statistical Association 67, 415-421. ⚲ section 4.3

Hall JN (1974) Inter-rater reliability of ward rating scales. British Journal of Psychiatry 125, 248-255. ⚲ section 1.6

Helzer JE et al.(1977) Reliability of psychiatric diagnosis. Archives of General Psychiatry 34, 129-133(I. A methodological review), 136-141 (II. The test/retest reliability of diagnostic classification). ⚲ section 1.1

Hildebrand DK, Laing JD and Rosenthal H (1977) Prediction Analysis of Cross Classifications. New York: Wiley. ⚲ section 1.4

Holman CDJ, James IR, Heenan PJ, Matz LR, Blackwell JB, Kelsall GRH, Singh A, and Ten Seldam REJ (1982) An improved method of analysis

of observer variation between pathologists. Histopathology 6, 581-589. ☞ section 1.4

Holmquist ND, McMahan CA and Williams OD (1967) Variability in classification of carcinoma in situ of the uterine cervix. Archives of Pathology 84, 334-345. ☞ sections 1.2 and 2.1

Hubert LJ (1977) Kappa revisited. Psychological Bulletin 84, 289-297. ☞ sections 2.2 and 2.3

Hubert LJ (1978) A general formula for the variance of Cohen's weighted kappa. Psychological Bulletin 85, 183-184. ☞ sections 4.2 and 4.3

James IR (1983) Analysis of nonagreements among multiple raters. Biometrics 39, 651-657. ☞ section 3.3

Koran LM (1975) The reliability of clinical methods, data and judgements. The New England Journal of Medicine 293, 642-646(Part I), 695-701 (Part II). ☞ section 1.1

Koran LM (1976) Increasing the reliability of clinical data and judgements. Annals of Clinical Research 8, 69-73. ☞ section 1.1

Kraemer HC (1979) Ramifications of a population model for κ as a coefficient of reliability. Psychometrika 44, 461-472. ☞ sections 1.8 and 3.4

Kraemer HC (1980) Extensions of the kappa coefficient. Biometrics 36, 207-216. ☞ sections 1.7, 1.9 and 3.3

Kraemer HC (1982) Estimating false alarms and missed events from interobserver agreement: Comment on Kaye. Psychological Bulletin 92, 749-754. ☞ section 1.8

Krippendorff K (1970) Bivariate agreement coefficients for reliability of data. In EF Borgatta and GW Bohrnstedt (eds.) Social Methodology 1970. San Francisco: Jossey-Bass, 139-150. ☞ section 1.5

Landis JR and Koch GG (1975) A review of statistical methods in the analysis of data arising from observer reliability studies. Statistica Neerlandica 29, 101-123(Part I), 151-161(Part II).

Landis JR and Koch GG (1977a) The measurement of observer agreement for categorical data. Biometrics 33, 159-174. ☞ section 1.2

Landis JR and Koch GG (1977b) An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple

observers. Biometrics 33, 363-374. ☞ section 2.7

Landis JR and Koch GG (1977c) A one-way components of variance model for categorical data. Biometrics 33, 671-679. ☞ sections 1.4 and 3.3

Light RJ (1971) Measures of response agreement for qualitative data: Some generalizations and alternatives. Psychological Bulletin 76, 365-377. ☞ section 2.2

Lord FM and Novick MR (1968) Statistical Theories of Mental Test Scores. Reading, Massachusetts: Addison-Wesley. ☞ section 3.4

Miller RG (1974) The jackknife - a review. Biometrika 61, 1-15. ☞ section 4.4

Musch DC, Landis JR, Higgins ITT, Gilson JC and Jones RN (1984) An application of kappa-type analyses to interobserver variation in classifying chest radiographs for pneumoconiosis. Statistics in Medicine 3, 73-83. ☞ section 1.9

O'Connell DL and Dobson AJ (1984) General observer-agreement measures on individual subjects and groups of subjects. Biometrics 40, 973-983. ☞ section 2.7

Parr WC (1983) A note on the jackknife, the bootstrap and the delta estimators of bias and variance. Biometrika 70, 719-722. ☞ section 1.9

Parr WC and Tolley HD (1982) Jackknifing in categorical data analysis. The Australian Journal of Statistics 24, 67-79. ☞ section 1.9

Popping R (1983a) Traces of agreement: On the dot-product as a coefficient of agreement. Quality and Quantity 17, 1-18. ☞ section 1.2

Popping R (1983b) Overeenstemmingsmaten voor Nominale Data. Ph.D. thesis, University of Groningen. ☞ section 1.2

Popping R (1984) Traces of agreement: On some agreement indices for open-ended questions. Quality and Quantity 18, 147-158. ☞ section 1.2

Popping R (1985) On agreement indices for nominal data. In WE Saris and IN Gallhofer (eds.) Sociometric Research 1985. In press. ☞ section 1.2

Quenouille MH (1956) Notes on bias in estimation. Biometrika 43, 353-360. ☞ section 1.9

Rogot E and Goldberg ID (1966) A proposed index for measuring agreement in test-retest studies. Journal of Chronic Diseases 19,

991-1006. ☞ section 1.1

Sandifer MG, Fleiss JL and Green LM (1968) Sample selection by diagnosis in clinical drug evaluations. Psychopharmacologia 13, 118-128. ☞ section 2.6

Schechter MT and Sheps SB (1985) Diagnostic testing revisited: pathways through uncertainty. Canadian Medical Association Journal 132, 755-760. ☞ section 2.5

Schouten HJA (1980) Measuring pairwise agreement among many observers. Biometrical Journal 22, 497-504. ☞ sections 1.4, 3.2, 3.3 and 4.2

Schouten HJA (1982a) Measuring pairwise agreement among many observers. II. Some improvements and additions. Biometrical Journal 24, 431-435. ☞ sections 1.4, 3.2, 3.3, 4.3, A2 and A4

Schouten HJA (1982b) Measuring pairwise interobserver agreement when all subjects are judged by the same observers. Statistica Neerlandica 36, 45-61. ☞ sections 1.7, 2.2, 2.3, 2.4 and 4.3

Scott WA (1955) Reliability of content analysis: The case of nominal scale coding. Public Opinion Quarterly 19, 321-325. ☞ preface and sections 1.2 and 4.2

Spodick DH (1975) On experts and expertise: the effect of variability in observer performance. The American Journal of Cardiology 36, 592-596. ☞ section 1.1

Steel RGD and Torrie JH (1980) Principles and Procedures of Statistics. Tokyo: McGraw-Hill Kogakusha. ☞ section 1.9

Tukey JW (1958) Bias and confidence in not quite large samples. The Annals of Mathematical Statistics 29, 614. ☞ section 1.9

Uebersax JS (1982) A generalized kappa coefficient. Educational and Psychological Measurement 42, 181-183. ☞ section 3.2

Van den Berge JH, Schouten HJA, Boomstra S, van Drunen Littel S and Braakman R (1979) Interobserver agreement in assessment of ocular signs in coma. Journal of Neurology, Neurosurgery and Psychiatry 42, 1163-1168. ☞ section 3.1

Van Es AJ, Gill RD and van Putten C (1983) Random number generators for a pocket calculator. Statistica Neerlandica 37, 95-102. ☞ section 4.5

Walter SD (1984) Measuring the reliability of clinical data: the case for using three observers. Revue d'Epidemiologie et de Santé Publique 32, 206-211. ℔ section 2.5

Wichmann BA and Hill ID (1982) Algorithm AS183: An efficient and portable pseudo-random number generator. Applied Statistics 31, 188-190; with a correction in 1984, vol.33, p.123. ℔ section 4.5

Wulff HR (1981) Rational Diagnosis and Treatment. Oxford: Blackwell. ℔ section 1.1

CURRICULUM VITAE

Hubert J.A. Schouten werd op 31 juli 1947 te Utrecht geboren. In 1966 behaalde hij het diploma h.b.s.-B aan het Sint Bonifatiuslyceum in Utrecht. Daarna volgde hij een studie wis- en natuurkunde aan de Rijksuniversiteit Utrecht. In zijn derde studiejaar hielp hij als student-assistent de eerstejaars met wiskundevraagstukken. In zijn vijfde en laatste studiejaar gaf hij les in wiskunde aan twee brugklassen van het Sint Vituscollege in Bussum. In 1971 studeerde hij af met hoofdvak wiskunde, in het bijzonder kansrekening en wiskundige statistiek bij professor G.J. Leppink, en bijvak theoretische natuurkunde. Na zijn studie werkte hij twee jaar bij de Afdeling Bedrijfspsychologie en -sociologie van de Koninklijke Luchtmacht. Vanaf 1973 werkt hij in het Instituut voor Biostatistica binnen de Faculteit der Geneeskunde van de Erasmus Universiteit Rotterdam.

NAWOORD

Al in 1976 legde de neurochirurg R. Braakman mij het statistische probleem voor. Hij beschreef enkele onderzoeken die waren opgezet om de betrouwbaarheid van bepaalde medische beoordelingen te analyseren en zonodig te verbeteren. Mijn promotor R. van Strik wees mij op enkele relevante artikelen waaruit ik concludeerde dat de bestaande statistische methoden onvoldoende waren uitgewerkt, en ik begon te denken aan mogelijke uitbreidingen. Via omwegen, langs doodlopende steegjes, over een brug van kennis uit het proefschrift van R. Popping, onder een viaduct van heldere en opbouwende kritiek van mijn beide promotoren W. Molenaar en R. van Strik, in warme zonneschijn en in louterend onweer van vele kanten, kwam na heel wat onderbrekingen dit proefschrift tot stand.

Het
uur
beslaat
een
ruime
tijd.

Stamelen,
consequent
handelen,
oreren,
uiteenzetten,
tot
eindelijk
nieuwwaardigheid.