empty cells. When $n_{ab} + n_{ba} = 0$ for any pair (such as categories 1 and 4 in Table 10.8), the ML fitted values for quasi-symmetry in those cells must also be zero since one of its likelihood equations is $\hat{\mu}_{ab} + \hat{\mu}_{ba} = n_{ab} + n_{ba}$. One should eliminate those cells from the fitting process to get the proper residual df value.

Under quasi-symmetry, $\hat{\tau}_{ab} = \exp(\hat{\lambda}_{aa} + \hat{\lambda}_{bb} - \hat{\lambda}_{ab} - \hat{\lambda}_{ba})$, where $\hat{\lambda}_{ab} = \hat{\lambda}_{ba}$. For categories 2 and 3 of Table 10.8, for instance, $\hat{\tau}_{23} = 10.7$.

Loglinear models directly address the association component of agreement. The quasi-symmetry model also yields information about similarity of marginal distributions. The simpler symmetry model that forces the margins to be identical fits Table 10.8 poorly ($G^2 = 39.2$, df = 5). The statistic $G^2(S\,|\,QS) = 39.2 - 1.0 = 38.2$ (df = 3) provides strong evidence of marginal heterogeneity. In Table 10.8, differences in marginal proportions are substantial in each category but the first. The marginal heterogeneity is one reason that the agreement is not stronger.

Models for agreement can take ordering of categories into account. Conditional on observer disagreement, a tendency usually remains for high (low) ratings by one observer to occur with relatively high (low) ratings by the other observer (see Problem 10.41).

### 10.5.4 Kappa Measure of Agreement

An alternative approach summarizes agreement with a single index. For nominal scales, the most popular measure is *Cohen's kappa* (Cohen 1960). It compares the probability of agreement $\Sigma_a \pi_{aa}$ to that expected if the ratings were independent, $\Sigma_a \pi_{a+} \pi_{+a}$, by

$$\kappa = \frac{\Sigma_a \pi_{aa} - \Sigma_a \pi_{a+} \pi_{+a}}{1 - \Sigma_a \pi_{a+} \pi_{+a}} .$$

The denominator equals the numerator with $\Sigma_a \pi_{aa}$ replaced by its maximum possible value of 1, corresponding to perfect agreement. Kappa equals 0 when the agreement merely equals that expected under independence. It equals 1.0 when perfect agreement occurs. The stronger the agreement, the higher is $\kappa$, for given marginal distributions. Negative values occur when agreement is weaker than expected by chance, but this rarely happens.

For multinomial sampling, the sample value $\hat{\kappa}$ has a large-sample normal distribution. Its estimated asymptotic variance (Fleiss et al. 1969) is

$$\hat{\sigma}^2(\hat{\kappa}) = \frac{1}{n}\left\{ \frac{P_o(1 - P_o)}{(1 - P_e)^2} + \frac{2(1 - P_o)[2P_oP_e - \Sigma_a p_{aa}(p_{a+} + p_{+a})]}{(1 - P_e)^3} \right.$$
$$\left. + \frac{(1 - P_o)^2\left[\Sigma_a\Sigma_b p_{ab}(p_{b+} + p_{+a})^2 - 4P_e^2\right]}{(1 - P_e)^4} \right\},$$

where $P_o = \Sigma_a p_{aa}$ and $P_e = \Sigma_a p_{a+} p_{+a}$. It is rarely plausible that agreement is no better than expected by chance. Thus, rather than testing $H_0$: $\kappa = 0$, it is more relevant to estimate strength of agreement by interval estimation of $\kappa$.

For Table 10.8, $P_o = 0.636$ and $P_e = 0.281$. Sample kappa equals $(0.636 - 0.281)/(1 - 0.281) = 0.493$. The difference between observed agreement and that expected under independence is about 50% of the maximum possible difference. The estimated standard error is 0.057, so $\kappa$ apparently falls roughly between 0.4 and 0.6, moderately strong agreement.

### 10.5.5 Weighted Kappa: Quantifying Disagreement

Kappa treats classifications as nominal. When categories are ordered, the seriousness of a disagreement depends on the difference between the ratings. For nominal classifications also, some disagreements may be considered more severe than others. The measure *weighted kappa* (Spitzer et al. 1967) uses weights $\{w_{ab}\}$ satisfying $0 \le w_{ab} \le 1$, with all $w_{aa} = 1$ and all $w_{ab} = w_{ba}$ to describe closeness of agreement. One possibility is $\{w_{ab} = 1 - |a - b|/(I - 1)\}$, for which agreement is greater for cells nearer the main diagonal. Fleiss and Cohen (1973) suggested $\{w_{ab} = 1 - (a - b)^2/(I - 1)^2\}$. The weighted agreement is $\Sigma_a \Sigma_b w_{ab} \pi_{ab}$ and weighted kappa is

$$\kappa_w = \frac{\Sigma_a \Sigma_b w_{ab} \pi_{ab} - \Sigma_a \Sigma_b w_{ab} \pi_{a+} \pi_{+b}}{1 - \Sigma_a \Sigma_b w_{ab} \pi_{a+} \pi_{+b}} .$$

Controversy surrounds the utility of kappa and weighted kappa, partly because their values depend strongly on the marginal distributions. The same diagnostic rating process can yield quite different values, depending on the proportions of cases of the various types (Problem 10.40). In summarizing a contingency table by a single number, the reduction in information can be severe. It is helpful to construct models providing more detailed investigation of the agreement and disagreement structure rather than to depend solely on a summary index.

### 10.5.6 Extensions to Multiple Observers

With several observers, ordinary loglinear models are not usually relevant. Their description of agreement and association between two observers is conditional on ratings by the others. It is more relevant to study this marginally, without conditioning on the other ratings. Hence, for $R$ observers, modelling simultaneously the pairwise agreement and association structure requires studying the $\binom{R}{2}$ pairs of two-way marginal distributions