Motivating Example

This example of a study that requires a test of agreement comes from the SAS Help. Medical researchers are interested in evaluating the efficacy of a new treatment for a skin condition. Dermatologists from participating clinics were trained to conduct the study and to evaluate the condition. After the training, two dermatologists examined patients with the skin condition from a pilot study and rated the same patients. It is desired to know whether or not the dermatologists are rating the patients in the same way. This can be tested with a test of agreement.

In the original data, the dermatologists had rated each patient as either terrible, poor, marginal, or clear. To use McNemar’s test with this example, the data had to fit a 2 by 2 table. To make this adjustment, the categories terrible and poor were merged into one category called bad, and the categories marginal and clear were merged into one category called good. Figure 1 shows the data in a two-way table.

![Figure 1: Number of patients whose skin condition was rated as bad and good by each dermatologist](image)

McNemar’s Test of Agreement

One approach to this problem is to use McNemar’s test of Agreement (SAS Help). This test uses the null hypothesis that $p_{1.} = p_{.1}$ versus the alternative that $p_{1.} \neq p_{.1}$. In the motivating example, this means that the null hypothesis is that the proportion of patients rated as bad by the first dermatologist is equal to the proportion of patients rated as bad by the second dermatologist. In other words, the null hypothesis is that there is agreement (although agreement in this context actually includes the possibility that there is no relationship at all) and the alternative is that there is disagreement.

McNemar’s test is performed using the test statistic:

$$Q = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$
The test statistic, $Q$, follows a chi-square distribution with one degree of freedom.

**Cohen’s Kappa**

Another, very different, approach to testing for agreement in the use of Cohen’s Kappa (SAS Help). Instead of simply performing a hypothesis test, Cohen found a statistic that estimates the degree of agreement in the data on a scale from -1 to 1. Positive values of kappa mean that there is agreement; negative values mean that there is active disagreement. If kappa is equal to zero, that means that there is no relation at all. In practice, since no relation is not agreement, it could be considered the same as disagreement. However, since there is a difference between two things being independent and two things having an inverse relationship, no relation will be considered to be different than disagreement in this paper.

Cohen’s Kappa is estimated using:

$$\hat{\kappa} = \frac{(P_o - P_e)}{(1 - P_e)}$$

where:

$$P_o = \Sigma p_{ii}$$
$$P_e = \Sigma p_{i.}p_{.i}$$

$P_o$ is equal to the observed proportion of agreement, and $P_e$ is equal to the expected proportion of agreement (Cohen, 1960). For $\hat{\kappa}$ to equal one, all of the observations must be on the diagonal, which is perfect agreement. For $\hat{\kappa}$ to equal zero, the observed counts must all be equal to the expected counts, which is the definition of independence. For $\hat{\kappa}$ to be negative, the observed agreement must be less than the expected agreement. Although this would happen about half of the time when kappa really does equal 0, in practice, it is unusual.

$\hat{\kappa}$ is asymptotically normal, which allows hypothesis testing to be done with a z-test. The null hypothesis is that kappa equals zero, or there is no relationship. Either a one or two sided alternative may be used. The two sided alternative is that kappa does not equal zero, or there is either agreement or disagreement; the one sided alternative is that kappa is greater than 0, or there is agreement. Under the null hypothesis:

$$\text{Var}(\hat{\kappa}) = \frac{[P_e + P_e^2 - \Sigma p_{ii}(p_{i.} + p_{.i})]}{(1 - P_e)^2 n}$$

**Differences**

The biggest differences between McNemar’s test and Cohen’s kappa are the null hypotheses and the amount of data that is actually being used.

McNemar’s test only uses the disagreement cells of the table, whereas Cohen’s kappa uses both the agreement and disagreement cells. As a result, the two methods can potentially give very different conclusions. From one point of view, using more data seems like it has to be
better, and sometimes it is. For example, if a sample shows perfect disagreement that happens to be balanced, McNemar’s test will fail to reject $H_0$, and a careless researcher might determine that there is agreement. In the same case, Cohen’s kappa would be at or near negative one, which would clearly show that there is strong disagreement. However, if the counts in the agreement cells are high enough, then $\kappa$ will be significantly larger than zero no matter what the counts in the disagreement cells are. If a researcher wants to detect even a small amount of disagreement when the agreement cell counts are high, then McNemar’s test may be more powerful.

Both of these tests are designed to determine whether or not there is agreement. However, null hypotheses are basically opposites of each other and, perhaps more importantly, the way that independence, or no relationship, is classified is completely different. If the hypotheses are stated purely in terms of agreement or disagreement, then the hypotheses for McNemar’s test are:

- $H_0$: agreement
- $H_1$: disagreement

The hypotheses for the test of Cohen’s kappa (one sided) are:

- $H_0$: disagreement
- $H_1$: agreement

This alone makes it important to determine which test is most appropriate prior to analyzing the data, because, with a large enough sample size, both tests could easily be significant. However, agreement and disagreement are not necessarily the only possibility. It is also possible that the two people, methods, etc. being compared are completely independent of each other and show neither a disposition to agree nor a disposition to disagree. For each test, the possibility of independence is lumped in with the null hypothesis. In McNemar’s test, independence will result in the disagreement cells being approximately equal, and the null hypothesis will fail to be rejected. Cohen’s kappa has a $\kappa$ near zero when there is independence, which will also result in the null hypothesis not being rejected. Because of this, the researcher must answer a couple questions. What should the test try to detect: agreement or disagreement? How should independence be handled: as agreement or as disagreement? Also, when interpreting the results, it is important to remember failing to reject the null hypothesis is not the same thing as accepting the null hypothesis.

**Results from Motivating Example**

The data from the motivating example was analyzed in SAS using the agree option in proc freq. The SAS output in figure 2 is the results of McNemar’s test. Figures 3 and 4 show the estimate of kappa and the results of the test of significance.
**McNemar’s Test**

<table>
<thead>
<tr>
<th>Statistic (S)</th>
<th>2.1304</th>
</tr>
</thead>
<tbody>
<tr>
<td>DF</td>
<td>1</td>
</tr>
<tr>
<td>Pr&gt; S</td>
<td>0.1444</td>
</tr>
</tbody>
</table>

Figure 2: Results of McNemar’s test

**Simple Kappa Coefficient**

| Kappa     | 0.4773 |
| ASE       | 0.0925 |
| 95% Lower Conf Limit | 0.2960 |
| 95% Upper Conf Limit | 0.6585 |

Figure 3: κ̂ and 95% confidence interval

**Test of H0: Kappa = 0**

| ASE under H0 | 0.1052 |
| Z            | 4.5350 |
| One-sided Pr> Z | <.0001 |
| Two-sided Pr> | <.0001 |

Figure 4: Test of significance for Cohen’s kappa

McNemar’s test gives a p-value of 0.1444 (figure 2), which is not significant. For that test, the null hypothesis is not rejected, there is insufficient evidence of disagreement between the two dermatologists.

The estimate of Cohen’s kappa is .4773 (figure 3), which indicates that there is moderate agreement. The one-sided p-value for the test of kappa equals zero is less than 0.0001 (figure 4). The test for Cohen’s kappa shows that there is significant agreement. This results confirms the result from McNemar’s test and goes one step further, showing that the dermatologists do tend to agree rather than the possibility that McNemar’s test could have be non-significant due to independence of the dermatologists ratings of skin condition.

**Other Tests of Agreement**

This example was only for a 2 by 2 table, but there are other tests of agreement for more complex data (SAS Help):

- Bowker’s Test of Symmetry is similar to McNemar’s test, but it is not restricted to 2 by 2 tables
- The weighted Kappa Coefficient is similar to the simple kappa coefficient discussed in this paper, but weights are used to account for differences between categories (eg: the difference between very poor and poor, versus the difference between poor and good).
- The overall kappa coefficient is used with multiple strata, assuming that there is a common kappa among the strata.
• Cochran’s Q test is used with 2 by 2 by 2... by 2 tables.

References


Appendix

```
Data SkinCondition;
    input derm1 $ derm2 $ count;
    datalines;
    bad bad29
    bad good 15
    good good36
    good bad 8
;
Proc freq data=SkinCondition order=data;
    Weight count;
    tables derm1*derm2 / agree noprint;
    Test kappa;
    run;
```