The next example shows that stratification can even reverse the association.

**Example of Simpson's paradox:**

This example is extracted from an article written by Morrell (1999). He discusses data collected in a South African longitudinal study of growth of children, referred to as the Birth to Ten study (BTT).

<u>Extract</u>: This study commenced in the greater Johannesburg/Soweto metropolitan area of South Africa during 1990. A birth cohort was formed from all singleton births during a seven-week period between April and June 1990 to women with permanent addresses within a defined area. Identification of children born during this seven-week period and living in the defined areas took place throughout the first year of the study, by the end of which 4029 births had been enrolled. The BTT study collected prenatal, birth, and early development information on these children. The aim of the study was to identify factors related to the emergence of cardiovascular disease risk factors in children living in an urban environment in South Africa. In 1995, when the children were five years old, the children and care-givers were invited to attend interviews. Detailed questionnaires were completed that included questions about living conditions within the child's home, the child's exposure to tobacco smoke, and additional health-related issues. The five-year sample consisted of 964 children. Unfortunately, there was a great deal of missing data in the baseline group, especially on the variables reported below.

If the five-year sample is to be used to draw conclusions about the entire birth cohort, the five-year group should have characteristics similar to those who were not traced from the initial group. Thus, the five-year group was compared to those who did not participate in the five-year interview on a number of factors. One of the factors was a variable that determined whether the mother had medical aid (which is similar to health insurance) at the time of the birth of the child.

Table 2.3 shows that 11.1% of those in the five-year cohort had medical aid, whereas 16.6% of those who were not traced had medical. This difference is statistically significant ($p$-value $= 0.007$). The subjects in the BTT study were also classified by their racial group. In this article we consider only white and black participants. Tables 2.4 and 2.5 show the distribution of the medical aid variable broken down by race (two strata). For whites, 83.3% of those in the five-year cohort had medical aid, whereas 82.5% of those who did not participate in the five-year tests had medical aid. For blacks, the corresponding percentages are 8.9% and 8.7%. This shows that even though overall a significantly smaller percentage of the five-year cohort had medical aid, when the race of the subjects is taken into account, the association is reversed. Furthermore, there is negligible evidence of any difference between the percentages when stratified by race; $p$-value $= 0.945$ and 0.891 for whites and blacks, respectively. The $(\mathrm{MH})^2$ statistic (page 42), which pools effects across the two race tables, has a value of 0.0025 with $p$-value = 0.9599.

Table 2.3: *Number and percentages of mothers with medical aid*

| | Children Not Traced | Five-Year Group | Total |
|---|---|---|---|
| Had Medical Aid | 195 (16.61%) | 46 (11.06%) | 241 |
| No Medical Aid | 979 (83.39%) | 370 (88.94%) | 1349 |
| Total | 1174 (100%) | 416 (100%) | 1590 |

Table 2.4: *Number and percentages of mothers with medical aid (white)*

| | Children Not Traced | Five-Year Group | Total |
|---|---|---|---|
| Had Medical Aid | 104 (82.54%) | 10 (83.33%) | 114 |
| No Medical Aid | 22 (17.46%) | 2 (16.67%) | 24 |
| Total | 126 (100%) | 12 (100%) | 138 |

Table 2.5: *Number and percentages of mothers with medical aid (black)*

| | Children Not Traced | Five-Year Group | Total |
|---|---|---|---|
| Had Medical Aid | 91 (8.68%) | 36 (8.91%) | 127 |
| No Medical Aid | 957 (91.32%) | 368 (91.09%) | 1325 |
| Total | 1048 (100%) | 404 (100%) | 1452 |

This reversal, and elimination, of association is easily explained. Whites tend to have more access to medical aid than do black South Africans (83% and 8.7%, respectively). In addition, many more blacks were originally included in the BTT study than whites (1452 blacks, 138 whites). Consequently, when the race groups were combined, a relatively small percentage (241/1590 = 15.16%) of the subjects have access to medical aid. At the five-year follow-up, very few whites agree to attend the interviews (12/138 = 8.67% of those with data on the medical aid variable). Possibly whites felt they had little to gain from participating in the study, while a larger proportion of blacks (404/1452 = 27.82% of those with data on the medical aid variable) continue into the five-year study. The blacks may have valued the medical checkup and screening provided to children in the study as a replacement for (or in addition to) a regular medical screening.

The data contained in the above tables are found in the S data frame BirthtoTen and in BirthtoTen.xls.

| Key to variables in BirthtoTen data: | | |
|---|---|---|
| Medical.Aid | 0 = No, | 1 = Yes |
| Traced | 0 = No, | 1 = Five-Year Cohort |
| Race | 1 = White, | 2 = Black |

The S function `crosstabs` produces contingency tables. The S function `mantelhae.test` conducts the $(\mathrm{MH})^2$ test when pooling $2 \times 2$ tables. The following S code produces the foregoing results:

```
> crosstabs(~ Medical.Aid + Traced,data=BirthtoTen)
> crosstabs(~ Medical.Aid + Traced,data=BirthtoTen,
                                   subset=Race==1)
> crosstabs(~ Medical.Aid+Traced,data=BirthtoTen,
                                   subset=Race==2)
> mantelhaen.test(BirthtoTen$Medical.Aid,BirthtoTen$Traced,
                                    BirthtoTen$Race)
```

## 2.3 Exercises

A. *Applications*

2.1 Use only hand-held calculator. No need for computer.

(a) Calculate the following table and sketch the Kaplan-Meier (K-M) estimate of survival for the data set $y$: 1, $1^+$, 2, 4, 4, $4^+$, 6, 9. ("+" denotes censored observation.) The s.e.$(\widehat{S}(t))$ is computed using Greenwood's formula (2.3) for the estimated (asymptotic) variance of the K-M curve at time $t$.

| $y_{(i)}$ | $d_i$ | $n_i$ | $\widehat{p}_i$ | $\widehat{S}(y_{(i)}) = \widehat{P}(T > y_{(i)})$ | s.e.$(\widehat{S}(y_{(i)}))$ |
|---|---|---|---|---|---|
| 0 | | | | | |
| 1 | | | | | |
| $1^+$ | | | | | |
| 2 | | | | | |
| 4 | | | | | |
| $4^+$ | | | | | |
| 6 | | | | | |
| 9 | | | | | |