CHAPTER 15

Rates and Proportions

15.1 INTRODUCTION

In this chapter and the next we want to study in more detail some of the topics dealing with counting data introduced in Chapter 6. In this chapter we want to take an epidemiological approach, studying populations by means of describing incidence and prevalence of disease. In a sense this is where statistics began: with a numerical description of the characteristics of a state, frequently involving mortality, fecundity, and morbidity. We call the occurrence of one of those outcomes an *event*. In the next chapter we deal with more recent developments, which have focused on a more detailed modeling of survival (hence also death, morbidity, and fecundity) and dealt with such data obtained in experiments rather than observational studies. An implication of the latter point is that sample sizes have been much smaller than used traditionally in the epidemiological context. For example, the evaluation of the success of heart transplants has, by necessity, been based on a relatively small set of data.

We begin the chapter with definitions of incidence and prevalence rates and discuss some problems with these "crude" rates. Two methods of standardization, direct and indirect, are then discussed and compared. In Section 15.4, a third standardization procedure is presented to adjust for varying exposure times among individuals. In Section 15.5, a brief tie-in is made to the multiple logistic procedures of Chapter 13. We close the chapter with notes, problems, and references.

15.2 RATES, INCIDENCE, AND PREVALENCE

The term *rate* refers to the amount of change occurring in a quantity with respect to time. In practice, *rate* refers to the amount of change in a variable over a specified time interval divided by the length of the time interval.

The data used in this chapter to illustrate the concepts come from the Third National Cancer Survey [National Cancer Institute, 1975]. For this reason we discuss the concepts in terms of incidence rates. The *incidence* of a disease in a fixed time interval is the number of new cases diagnosed during the time interval. The *prevalence* of a disease is the number of people with the disease at a fixed time point. For a chronic disease, incidence and prevalence may present markedly different ideas of the importance of a disease.

Consider the Third National Cancer Survey [National Cancer Institute, 1975]. This survey examined the incidence of cancer (by site) in nine areas during the time period 1969–1971.

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2

The areas were the Detroit SMSA (Standard Metropolitan Statistical Area); Pittsburgh SMSA, Atlanta SMSA, Birmingham SMSA, Dallas–Fort Worth SMSA, state of Iowa, Minneapolis–St. Paul SMSA, state of Colorado, and the San Francisco–Oakland SMSA. The information used in this chapter refers to the combined data from the Atlanta SMSA and San Francisco–Oakland SMSA. The data are abstracted from tables in the survey. Suppose that we wanted the rate for all sites (of cancer) combined. The rate per year in the 1969–1971 time interval would be simply the number of cases divided by 3, as the data were collected over a three-year interval. The rates are as follows:

$$\text{Combined area :} \quad \frac{181{,}027}{3} = 60{,}342.3$$

$$\text{Atlanta :} \quad \frac{9{,}341}{3} = 3{,}113.7$$

$$\text{San Francisco–Oakland :} \quad \frac{30{,}931}{3} = 10{,}310.3$$

Can we conclude that cancer incidence is worse in the San Francisco–Oakland area than in the Atlanta area? The answer is "yes and no." Yes, in that there are more cases to take care of in the San Francisco–Oakland area. If we are concerned about the chance of a person getting cancer, the numbers would not be meaningful. As the San Francisco–Oakland area may have a larger population, the number of cases per number of the population might be less. To make comparisons taking the population size into account, we use

$$\text{incidence per time interval} = \frac{\text{number of new cases}}{\text{total population} \times \text{time interval}} \tag{1}$$

The result of equation (1) would be quite small, so that the number of cases per 100,000 population is used to give a more convenient number. The rate per 100,000 population per year is then

$$\text{incidence per 100,000 per time interval} = \frac{\text{number of new cases}}{\text{total population} \times \text{time interval}} \times 100{,}000$$

For these data sets, the values are:

$$\text{Combined area :} \quad \frac{181{,}027 \times 100{,}000}{21{,}003{,}451 \times 3} = 287.3 \text{ new cases per } 100{,}000 \text{ per year}$$

$$\text{Atlanta :} \quad \frac{9{,}341 \times 100{,}000}{1{,}390{,}164 \times 3} = 224.0 \text{ new cases per } 100{,}000 \text{ per year}$$

$$\text{San Francisco-Oakland :} \quad \frac{30{,}931 \times 100{,}000}{3{,}109{,}519 \times 3} = 331.6 \text{ new cases per } 100{,}000 \text{ per year}$$

Even after adjusting for population size, the San Francisco–Oakland area has a higher overall rate.

Note several facts about the estimated rates. The estimates are binomial proportions times a constant (here 100,000/3). Thus, the rate has a standard error easily estimated. Let N be the total population and n the number of new cases; the rate is $n/N \times C$ ($C = 100{,}000/3$ in this example) and the standard error is estimated by

$$\sqrt{C^2 \frac{1}{N} \frac{n}{N} \left(1 - \frac{n}{N}\right)}$$

or

$$\text{standard error of rate per time interval} = C\sqrt{\frac{1}{N} \frac{n}{N} \left(1 - \frac{n}{N}\right)}$$

For example, the combined area estimate has a standard error of

$$\frac{100{,}000}{3}\sqrt{\frac{1}{21{,}003{,}451}\frac{181{,}027}{21{,}003{,}451}\left(1-\frac{181{,}027}{21{,}003{,}451}\right)}=0.67$$

As the rates are assumed to be binomial proportions, the methods of Chapter 6 may be used to get adjusted estimates or standardized estimates of proportions.

Rates computed by the foregoing methods,

$$\frac{\text{number of new cases in the interval}}{\text{population size} \times \text{time interval}}$$

are called *crude* or *total rates*. This term is used in distinction to *standardized* or *adjusted rates*, as discussed below.

Similarly, a *prevalence rate* can be defined as

$$\text{prevalence} = \frac{\text{number of cases at a point in time}}{\text{population size}}$$

Sometimes a distinction is made between *point prevalence* and *prevalence* to facilitate discussion of chronic disease such as epilepsy and a disease of shorter duration, for example, a common cold or even accidents. It is debatable whether the word *prevalence* should be used for accidents or illnesses of short duration.

15.3 DIRECT AND INDIRECT STANDARDIZATION

15.3.1 Problems with the Use of Crude Rates

Crude rates are useful for certain purposes. For example, the crude rates indicate the load of new cases per capita in a given area of the country. Suppose that we wished to use the cancer rates as epidemiologic indicators. The inference would be that it was likely that environmental or genetic differences were responsible for a difference, if any. There may be simpler explanations, however. Breast cancer rates would probably differ in areas that had differing gender proportions. A retirement community with an older population will tend to have a higher rate. To make fair comparisons, we often want to adjust for the differences between populations in one or more factors (covariates). One approach is to find an index that is adjusted in some fashion. We discuss two methods of adjustment in the next two sections.

15.3.2 Direct Standardization

In direct standardization we are interested in adjusting by one or more variables that are divided (or naturally fall) into discrete categories. For example, in Table 15.1 we adjust for gender and for age divided into a total of 18 categories. The idea is to find an answer to the following question: Suppose that the distribution with regard to the adjusting factors was not as observed, but rather, had been the same as this other (reference) population; what would the rate have been? In other words, we apply the risks observed in our study population to a reference population.

In symbols, the adjusting variable is broken down into I cells. In each cell we know the number of events (the numerator) n_i and the total number of individuals (the denominator) N_i:

Level of adjusting factor, i:	1	2	$\cdots$	i	$\cdots$	I
Proportion observed in study population:	$\frac{n_1}{N_1}$	$\frac{n_2}{N_2}$	$\cdots$	$\frac{n_i}{N_i}$	$\cdots$	$\frac{n_1}{N_1}$

Table 15.1 Rate for Cancer of All Sites for Blacks in the San Francisco–Oakland SMSA and Reference Population

	Study Population n_i/N_i		Reference Population M_i	
Age	Females	Males	Females	Males
<5	8/16,046	6/16,493	872,451	908,739
5–9	6/18,852	7/19,265	1,012,554	1,053,350
10–14	6/19,034	3/19,070	1,061,579	1,098,507
15–19	7/16,507	6/16,506	971,894	964,845
20–24	16/15,885	9/14,015	919,434	796,774
25–29	27/12,886	19/12,091	755,140	731,598
30–34	28/10,705	18/10,445	620,499	603,548
35–39	46/9,580	25/8,764	595,108	570,117
40–44	83/9,862	47/8,858	650,232	618,891
45–49	109/10,341	108/9,297	661,500	623,879
50–54	125/8,691	131/8,052	595,876	558,124
55–59	120/6,850	189/6,428	520,069	481,137
60–64	102/5,017	158/4,690	442,191	391,746
65–69	119/3,806	159/3,345	367,046	292,621
70–74	75/2,264	154/1,847	300,747	216,929
75–79	44/1,403	72/931	224,513	149,867
80–84	28/765	51/471	139,552	84,360
>85	25/629	26/416	96,419	51,615
Subtotal	974/169,123	1,188/160,984	10,806,804	10,196,647
Total	2,162/330,107		21,003,451	

Source: National Cancer Institute [1975].

Both numerator and denominator are presented in the table. The crude rate is estimated by

$$C\frac{\sum_{i=1}^{I} n_i}{\sum_{i=1}^{I} N_i}$$

Consider now a *standard or reference population*, which instead of having N_i persons in the ith cell has M_i.

	Reference Population					
Level of adjusting factor	1	2	⋯	i	⋯	I
Number in reference population	M_1	M_2	⋯	M_i	⋯	M_I

The question now is: If the study population has M_i instead of N_i persons in the ith cell, what would the crude rate have been? We cannot determine what the crude rate was, but we can estimate what it might have been. In the ith cell the proportion of observed deaths was n_i/N_i. If the same proportion of deaths occurred with M_i persons, we would expect

$$n_i^* = \frac{n_i}{N_i} M_i \text{ deaths}$$

Thus, if the adjusting variables had been distributed with M_i persons in the ith cell, we estimate that the data would have been:

Level of adjusting factor:	1	2	$\cdots$	i	$\cdots$	I
Expected proportion of cases:	$\frac{n_1 M_1/N_1}{M_1}$	$\frac{n_2 M_2/N_2}{M_2}$	$\cdots$	$\frac{n_i^*}{M_i}$	$\cdots$	$\frac{n_I M_I/N_I}{M_I}$

The *adjusted rate*, r, is the crude rate for this estimated standard population:

$$r = \frac{C \sum_{i=1}^{I} n_i M_i/N_i}{\sum_{i=1}^{I} M_i} = \frac{C \sum_{i=1}^{I} n_i^*}{\sum_{i=1}^{I} M_i}$$

As an example, consider the rate for cancer for all sites for blacks in the San Francisco–Oakland SMSA, adjusted for gender and age to the total combined sample of the Third Cancer Survey, as given by the 1970 census. There are two gender categories and 18 age categories, for a total of 36 cells. The cells are laid out in two columns rather than in one row of 36 cells. The data are given in Table 15.1.

The crude rate for the San Francisco–Oakland black population is

$$\frac{100{,}000}{3} \frac{974 + 1188}{169{,}123 + 160{,}984} = 218.3$$

Table 15.2 gives the values of $n_i M_i/N_i$.

The gender- and age-adjusted rate is thus

$$\frac{100{,}000}{3} \frac{193{,}499.42}{21{,}003{,}451} = 307.09$$

Note the dramatic change in the estimated rate. This occurs because the San Francisco–Oakland SMSA black population differs in its age distribution from the overall sample.

The variance is estimated by considering the denominators in the cell as fixed and using the binomial variance of the n_i's. Since the cells constitute independent samples,

$$\begin{aligned} \operatorname{var}(r) &= \operatorname{var}\left(C \sum_{i=1}^{I} \frac{n_i M_i}{N_i} \Big/ \sum_{i=1}^{I} M_i \right) \\ &= \frac{C^2}{M_\cdot^2} \sum_{i=1}^{I} \left(\frac{M_i}{N_i} \right)^2 \operatorname{var}(n_i) \end{aligned}$$

Table 15.2 Estimated Number of Cases per Cell ($n_i M_i/N_i$) if the San Francisco–Oakland Area Had the Reference Population Age and Gender Distribution

Age	Females	Males	Age	Females	Males
<5	434.97	330.59	55–59	9,110.70	14,146.69
5–9	322.26	382.74	60–64	8,990.13	13,197.41
10–14	334.64	172.81	65–69	11,476.21	13,909.34
15–19	412.14	350.73	70–74	9,962.91	18,087.20
20–24	926.09	511.66	75–79	7,041.03	11,590.14
25–29	1,582.24	1,149.65	80–84	5,107.79	9,134.52
30–34	1,622.98	1,040.10	>85	3,832.23	3,225.94
35–39	2,857.51	1,629.30			
40–44	5,472.45	3,283.80	Subtotal	85,029.16	108,470.26
45–49	6,972.58	7,247.38	Total	193,499.42	
50–54	8,570.30	9,080.26			

$$= \frac{C^2}{M_\cdot^2} \sum_{i=1}^{I} \left(\frac{M_i}{N_i}\right)^2 N_i \frac{n_i}{N_i} \left(1 - \frac{n_i}{N_i}\right)$$

$$= \frac{C^2}{M_\cdot^2} \sum_{i=1}^{I} \frac{M_i}{N_i} \frac{n_i M_i}{N_i} \left(1 - \frac{n_i}{N_i}\right)$$

where $M_\cdot = \sum_{i=1}^{I} M_i$.

If n_i/N_i is small, then $1 - n_i/N_i \doteq 1$ and

$$\text{var}(r) \doteq \frac{C^2}{M_\cdot^2} \sum_{i=1}^{I} \frac{M_i}{N_i} \left(\frac{n_i M_i}{N_i}\right) \tag{2}$$

We use this to compute a 95% confidence interval for the adjusted rate computed above. Using equation (2), the standard error is

$$\text{SE}(r) = \frac{C}{M_\cdot} \sqrt{\sum_{i=1}^{I} \frac{M_i}{N_i} \left(\frac{n_i M_i}{N_i}\right)}$$

$$= \frac{100{,}000}{3} \frac{1}{21{,}003{,}451} \left(\frac{872{,}451}{16{,}046} 434.97 + \cdots\right)^{1/2}$$

$$= 7.02$$

The quantity r is approximately normally distributed, so that the interval is

$$307.09 \pm 1.96 \times 7.02 \quad \text{or} \quad (293.3, 320.8)$$

If adjusted rates are estimated for two different populations, say r_1 and r_2, with standard errors $\text{SE}(r_1)$ and $\text{SE}(r_2)$, respectively, equality of the adjusted rates may be tested by using

$$z = \frac{r_1 - r_2}{\sqrt{\text{SE}(r_1)^2 + \text{SE}(r_2)^2}}$$

The $N(0,1)$ critical values are used, as z is approximately $N(0,1)$ under the null hypothesis of equal rates.

15.3.3 Indirect Standardization

In indirect standardization, the procedure of direct standardization is used in the opposite direction. That is, we ask the question: What would the mortality rate have been for the study population if it had the same rates as the population reference? That is, we apply the observed risks in the reference population to the study population.

Let m_i be the number of deaths in the reference population in the ith cell. The data are:

Level of adjusting factor:	1	2	$\cdots$	i	$\cdots$	I
Observed proportion in reference population:	$\frac{m_1}{M_1}$	$\frac{m_2}{M_2}$	$\cdots$	$\frac{m_i}{M_i}$	$\cdots$	$\frac{m_I}{M_I}$

where both numerator and denominators are presented in the table. Also,

Level of adjusting factor:	1	2	$\cdots$	i	$\cdots$	I
Denominators in study population:	N_1	N_2	$\cdots$	N_i	$\cdots$	N_I

The estimate of the rate the study population would have experienced is (analogous to the argument in Section 15.3.2)

$$r_{\text{REF}} = \frac{C \sum_{i=1}^{I} N_i (m_i / M_i)}{\sum_{i=1}^{I} N_i}$$

The crude rate for the study population is

$$r_{\text{STUDY}} = \frac{C \sum_{i=1}^{I} n_i}{\sum_{i=1}^{I} N_i}$$

where n_i is the observed number of cases in the study population at level i. Usually, there is not much interest in comparing the values r_{REF} and r_{STUDY} as such, because the distribution of the study population with regard to the adjusting factors is not a distribution of much interest. For this reason, attention is usually focused on the *standardized mortality ratio* (SMR), when death rates are considered, or the *standardized incidence ratio* (SIR), defined to be

$$\text{standardized ratio} = s = \frac{r_{\text{STUDY}}}{r_{\text{REF}}} = \frac{\sum_{i=1}^{I} n_i}{\sum_{i=1}^{I} N_i m_i / M_i} \tag{3}$$

The main advantage of the indirect standardization is that the SMR involves only the total number of events, so you do not need to know in which cells the deaths occur for the study population. An alternative way of thinking of the SMR is that it is the observed number of deaths in the study population divided by the expected number if the cell-specific rates of the reference population held.

As an example, let us compute the SIR of cancer in black males in the Third Cancer Survey, using white males of the same study as the reference population and adjusting for age. The data are presented in Table 15.3. The standardized incidence ratio is

$$s = \frac{8793}{7474.16} = 1.17645 = 1.18$$

One reasonable question to ask is whether this ratio is significantly different from 1. An approximate variance can be derived as follows:

$$s = \frac{O}{E} \qquad \text{where} \quad O = \sum_{i=1}^{I} n_i = n. \quad \text{and} \quad E = \sum_{i=1}^{I} N_i \left(\frac{m_i}{M_i} \right)$$

The variance of s is estimated by

$$\text{var}(s) = \frac{\text{var}(O) + s^2 \text{ var}(E)}{E^2} \tag{4}$$

The basic "trick" is to (1) assume that the number of cases in a particular cell follows a Poisson distribution and (2) to note that the sum of independent Poisson random variables is Poisson. Using these two facts yields

$$\text{var}(O) \doteq \sum_{i=1}^{I} n_i = n \tag{5}$$

Table 15.3 Cancer of All Areas Combined, Number of Cases, Black and White Males by Age and Number Eligible by Age

Age	Black Males		White Males			
	n_1	N_1	m_1	M_1	$\frac{N_i m_i}{M_i}$	$\left(\frac{N_i}{M_i}\right)^2 m_i$
<5	45	120,122	450	773,459	69.89	10.85
5–9	34	130,379	329	907,543	47.26	6.79
10–14	39	134,313	300	949,669	42.43	6.00
15–19	45	112,969	434	837,614	58.53	7.89
20–24	49	86,689	657	694,670	81.99	10.23
25–29	63	71,348	688	647,304	75.83	8.36
30–34	84	57,844	724	533,856	78.45	8.50
35–39	129	54,752	1,097	505,434	118.83	12.87
40–44	318	57,070	2,027	552,780	209.27	21.61
45–49	582	56,153	3,947	559,241	396.31	39.79
50–54	818	48,753	6,040	503,163	585.23	56.71
55–59	1,170	42,580	8,711	432,982	856.65	84.24
60–64	1,291	33,892	10,966	352,315	1,054.91	101.48
65–69	1,367	27,239	11,913	261,067	1,242.97	129.69
70–74	1,266	17,891	11,735	196,291	1,069.59	97.49
75–79	788	9,827	10,546	138,532	748.10	53.07
80–84	461	4,995	6,643	78,044	425.17	27.21
>85	244	3,850	3,799	46,766	312.75	25.75
Total	8,793	1,070,700	81,006	8,970,730	7,474.16	708.53

and

$$\operatorname{var}(E) \doteq \operatorname{var}\left(\sum_{i=1}^{I} \frac{N_i}{M_i} m_i\right)$$

$$= \sum_{i=1}^{I} \left(\frac{N_i}{M_i}\right)^2 m_i \tag{6}$$

The variance of s is estimated by using equations (4), (5), and (6):

$$\operatorname{var}(s) = \frac{n_\cdot + s^2 \sum (N_i/M_i)^2 m_i}{E^2}$$

A test of the hypothesis that the population value of s is 1 is obtained from

$$z = \frac{s - 1}{\sqrt{\operatorname{var}(s)}}$$

and $N(0, 1)$ critical values.

For the example,

$$\sum_{i=1}^{I} n_i = n_\cdot = 8793$$

$$E = \sum_{i=1}^{I} \frac{N_i}{M_i} m_i = 7474.16$$

$$\text{var}(E) \doteq \sum_{i=1}^{I} \left(\frac{N_i}{M_i}\right)^2 m_i = 708.53$$

$$\text{var}(s) \doteq \frac{8793 + (1.17645)^2 \times 708.53}{(7474.16)^2} = 0.000174957$$

From this and a standard error of $s \doteq 0.013$, the ratio is significantly different from one using

$$z = \frac{s - 1}{\text{SE}(s)} = \frac{0.17645}{0.013227} = 13.2$$

and $N(0, 1)$ critical values.

If the reference population is much larger than the study population, $\text{var}(E)$ will be much less than $\text{var}(O)$ and you may approximate $\text{var}(s)$ by $\text{var}(O)/E^2$.

15.3.4 Drawbacks to Using Standardized Rates

Any time a complex situation is summarized in one or a few numbers, considerable information is lost. There is always a danger that the lost information is crucial for understanding the situation under study. For example, two populations may have almost the same standardized rates but may differ greatly within the different cells; one population has much larger values in one subset of the cells and the reverse situation in another subset of cells. Even when the standardized rates differ, it is not clear if the difference is somewhat uniform across cells or results mostly from one or a few cells with much larger differences.

The moral of the story is that whenever possible, the rates in the cells used in standardization should be examined individually in addition to working with the standardized rates.

15.4 HAZARD RATES: WHEN SUBJECTS DIFFER IN EXPOSURE TIME

In the rates computed above, each person was exposed (eligible for cancer incidence) over the same length of time (three years, 1969–1971). (This is not quite true, as there is some population mobility, births, and deaths. The assumption that each person was exposed for three years is valid to a high degree of approximation.) There are other circumstances where people are observed for varying lengths of time. This happens, for example, when patients are recruited sequentially as they appear at a medical care facility. One approach would be to restrict the analysis to those who had been observed for at least some fixed amount of time (e.g., for one year). If large numbers of persons are not observed, this approach is wasteful by throwing away valuable and needed information. This section presents an approach that allows the rates to use all the available information if certain assumptions are satisfied.

Suppose that we observe subjects over time and look for an event that occurs only once. For definiteness, we speak about observing people where the event is death. Assume that over the time interval observed, if a subject has survived to some time t_0, the probability of death in a short interval from t_0 to t_1 is almost $\lambda(t_1 - t_0)$. The quantity λ is called the *hazard rate, force of mortality*, or *instantaneous death rate*. The units of λ are deaths per time unit.

How would we estimate λ from data in a real-life situation? Suppose that we have n individuals and begin observing the ith person at time B_i. If the person dies, let the time of death be D_i. Let the time of last contact be C_i for those people who are still alive. Thus, the time we are observing each person at risk of death is

$$O_i = \begin{cases} C_i - B_i & \text{if the subject is alive} \\ D_i - B_i & \text{if the subject is dead} \end{cases}$$

An unbiased estimate of λ is

$$\text{estimated hazard rate} = \widehat{\lambda} = \frac{\text{number of observed deaths}}{\sum_{i=1}^{n} O_i} = \frac{L}{\sum_{i=1}^{n} O_i} \tag{7}$$

As in the earlier sections of this chapter, $\widehat{\lambda}$ is often normalized to have different units. For example, suppose that $\widehat{\lambda}$ is in deaths per day of observation. That is, suppose that O_i is measured in days. To convert to deaths per 100 observation years, we use

$$\widehat{\lambda} \times 365 \frac{\text{days}}{\text{year}} \times 100$$

As an example, consider the paper by Clark et al. [1971]. This paper discusses the prognosis of patients who have undergone cardiac (heart) transplantation. They present data on 20 transplanted patients. These data are presented in Table 15.4. To estimate the deaths per year of exposure, we have

$$\frac{12 \text{ deaths}}{3599 \text{ exposure days}} \frac{365 \text{ days}}{\text{year}} = 1.22 \frac{\text{deaths}}{\text{exposure year}}$$

To compute the variance and standard error of the observed hazard rate, we again assume that L in equation (7) has a Poisson distribution. So conditional on the total observation period, the variability of the estimated hazard rate is proportional to the variance of L, which is estimated by L itself. Let

$$\widehat{\lambda} = \frac{CL}{\sum_{i=1}^{n} O_i}$$

where C is a constant that standardizes the hazard rate appropriately.

Table 15.4 Stanford Heart Transplant Data

i	Date of Transplantation	Date of Death	Time at Risk in Days (*if alive)[a]
1	1/6/68	1/21/68	15
2	5/2/68	5/5/68	3
3	8/22/68	10/7/68	46
4	8/31/68	—	608*
5	9/9/68	1/14/68	127
6	10/5/68	12/5/68	61
7	10/26/68	—	552*
8	11/20/68	12/14/68	24
9	11/22/68	8/30/69	281
10	2/8/69	—	447*
11	2/15/69	2/25/69	10
12	3/29/69	5/7/69	39
13	4/13/69	—	383*
14	5/22/69	—	344*
15	7/16/69	11/29/69	136
16	8/16/69	8/17/69	1
17	9/3/69	—	240*
18	9/14/69	11/13/69	60
19	1/3/70	—	118*
20	1/16/70	—	104*

[a] Total exposure days = 3599, $L = 12$.

Then the standard error of $\widehat{\lambda}$, $\mathrm{SE}(\widehat{\lambda})$, is approximately

$$\mathrm{SE}(\widehat{\lambda}) \doteq \frac{C}{\sum_{i=1}^{n} O_i}\sqrt{L}$$

A confidence interval for λ can be constructed by using confidence limits (L_1, L_2) for $E(L)$ as described in Note 6.8:

$$\text{confidence interval for } \lambda = \left(\frac{CL_1}{\sum_{i=1}^{n} O_i}, \frac{CL_2}{\sum_{i=1}^{n} O_i}\right)$$

For the example, a 95% confidence interval for the number of deaths is (6.2–21.0). A 95% confidence interval for the hazard rate is then

$$\left(\frac{6.2}{3599} \times 365, \frac{21.0}{3599} \times 365\right) = (0.63, 2.13)$$

Note that this assumes a constant hazard rate from day of transplant; this assumption is suspect. In Chapter 16 some other approaches to analyzing such data are given.

As a second more complicated illustration, consider the work of Bruce et al. [1976]. This study analyzed the experience of the Cardiopulmonary Research Institute (CAPRI) in Seattle, Washington. The program provided medically supervised exercise programs for diseased subjects. Over 50% of the participants dropped out of the program. As the subjects who continued participation and those who dropped out had similar characteristics, it was decided to compare the mortality rates for men to see if the training prevented mortality. It was recognized that subjects might drop out because of factors relating to disease, and the inference would be weak in the event of an observed difference.

The interest of this example is in the appropriate method of calculating the rates. All subjects, *including the dropouts*, enter into the computation of the mortality for active participants! The reason for this is that had they died during training, they would have been counted as active participant deaths. Thus, training must be credited with the exposure time or observed time when the dropouts were in training. For those who did not die and dropped out, the date of last contact *as an active participant* was the date at which the subjects left the training program. (Topics related to this are dealt with in Chapter 16).

In summary, to compute the mortality rates for active participants, all subjects have an observation time. The times are:

1. O_i = (time of death − time of enrollment) for those who died as active participants
2. O_i = (time of last contact − time of enrollment) for those in the program at last contact
3. O_i = (time of dropping the program−time of enrollment) for those who dropped whether or not a subsequent death was observed

The rate $\widehat{\lambda}_A$ for active participants is then computed as

$$\widehat{\lambda}_A = \frac{\text{number of deaths observed during training}}{\sum_{\text{all individuals}} O_i} = \frac{L_A}{\sum O_i}$$

To estimate the rate for dropouts, only those who drop out have time at risk of dying as a dropout. For those who have died, the time observed is

$$O_i' = (\text{time of death} - \text{time the subject dropped out})$$

For those alive at the last contact,

$$O_i' = (\text{time of last contact} - \text{time the subject dropped out})$$

The hazard rate for the dropouts, $\widehat{\lambda}_D$, is

$$\widehat{\lambda}_D = \frac{\text{number of deaths observed during dropout period}}{\sum_{\text{dropouts}} O_i'} = \frac{L_D}{\sum O_i'}$$

The paper reports rates of 2.7 deaths per 100 person-years for the active participants based on 16 deaths. The mortality rate for dropouts was 4.7 based on 34 deaths.

Are the rates statistically different at a 5% significance level? For a Poisson variable, L, the variance equals the expected number of observations and is thus estimated by the value of the variable itself. The rates $\widehat{\lambda}$ are of the form

$$\widehat{\lambda} = CL \qquad (L \text{ the number of events})$$

Thus, $\text{var}(\widehat{\lambda}) = C^2 \text{ var}(L) \doteq C^2 L = \widehat{\lambda}^2/L$.

To compare the two rates,

$$\text{var}(\widehat{\lambda}_A - \widehat{\lambda}_D) = \text{var}(\widehat{\lambda}_A) + \text{var}(\widehat{\lambda}_D) = \frac{\widehat{\lambda}_A^2}{L_A} + \frac{\widehat{\lambda}_D^2}{L_D}$$

The approximation is good for large L.

An approximate normal test for the equality of the rates is

$$z = \frac{\widehat{\lambda}_A - \widehat{\lambda}_D}{\sqrt{\widehat{\lambda}_A^2/L_A + \widehat{\lambda}_D^2/L_D}}$$

For the example, $L_A = 16, \widehat{\lambda}_A = 2.7$, and $L_D = 34, \widehat{\lambda}_D = 4.7$, so that

$$\begin{aligned} z &= \frac{2.7 - 4.7}{\sqrt{(2.7)^2/16 + (4.7)^2/34}} \\ &= -1.90 \end{aligned}$$

Thus, the difference between the two groups was not statistically significant at the 5% level.

15.5 MULTIPLE LOGISTIC MODEL FOR ESTIMATED RISK AND ADJUSTED RATES

In Chapter 13 the linear discriminant model or multiple logistic model was used to estimate the probability of an event as a function of covariates, $X_1, \ldots, X_n$. Suppose that we want a direct adjusted rate, where $X_1(i), \ldots, X_n(i)$ was the covariate value at the midpoints of the ith cell. For the study population, let p_i be the adjusted probability of an event at $X_1(i), \ldots, X_n(i)$. An adjusted estimate of the probability of an event is

$$\widehat{p} = \frac{\sum_{i=1}^{I} M_i p_i}{\sum_{i=1}^{I} M_i}$$

where M_i is the number of reference population subjects in the ith cell. This equation can be written as

$$\widehat{p} = \sum_{i=1}^{I} \left(\frac{M_i}{M.} p_i \right)$$

where $M. = \sum_{i=1}^{I} M_i$.

If the study population is small, it is better to estimate the p_i using the approach of Chapter 13 rather than the direct standardization approach of Section 15.3. This will usually be the case when there are several covariates with many possible values.

NOTES

15.1 More Than One Event per Subject

In some studies, each person may experience more than one event: for example, seizures in epileptic patients. In this case, each person could contribute more than once to the numerator in the calculation of a rate. In addition, exposure time or observed time would continue beyond an event, as the person is still at risk for another event. You need to check in this case that there are not people with "too many" events; that is, events "cluster" in a small subset of the population. A preliminary test for clustering may then be called for. This is a complicated topic. See Kalbfleisch and Prentice [2002] for references. One possible way of circumventing the problem is to record the time to the second or kth event. This builds a certain robustness into the data, but of course, makes it not possible to investigate the clustering, which may be of primary interest.

15.2 Standardization with Varying Observation Time

It is possible to compute standardized rates when the study population has the rate in each cell determined by the method of Section 15.4; that is, people are observed for varying lengths of time. In this note we discuss only the method for direct standardization.

Suppose that in each of the i cells, the rates in the study population is computed as CL_i/O_i, where C is a constant, L_i the number of events, and O_i the sum of the times observed for subjects in that cell. The adjusted rate is

$$\frac{\sum_{i=1}^{I} (M_i/L_i)\, O_i}{\sum_{i=1}^{I} M_i} = \frac{C \sum_{i=1}^{I} M_i \widehat{\lambda}_i}{M.} \qquad \text{where} \quad \widehat{\lambda}_i = \frac{L_i}{O_i}$$

The standard error is estimated to be

$$\frac{C}{M.} \sqrt{\sum_{i=1}^{I} \left(\frac{M_i}{O_i} \right) L_i}$$

15.3 Incidence, Prevalence, and Time

The *incidence* of a disease is the rate at which new cases appear; the *prevalence* is the proportion of the population that has the disease. When a disease is in a steady state, these are related via the average duration of disease:

$$\text{prevalence} = \text{incidence} \times \text{duration}$$

That is, if you catch a cold twice per year and each cold lasts a week, you will spend two weeks per year with a cold, so $2/52$ of the population should have a cold at any given time.

This equation breaks down if the disease lasts for all or most of your life and does not describe transient epidemics.

15.4 Sources of Demographic and Natural Data

There are many government sources of data in all of the Western countries. Governments of European countries, Canada, and the United States regularly publish vital statistics data as well as results of population surveys such as the Third National Cancer Survey [National Cancer Institute, 1975]. In the United States, the National Center for Health Statistics (*http://www.cdc.gov/nhcs*) publishes more than 20 series of monographs dealing with a variety of topics. For example, Series 20 provides natural data on mortality; Series 21, on natality, marriage, and divorce. These reports are obtainable from the U.S. government.

15.5 Binomial Assumptions

There is some question whether the binomial assumptions (see Chapter 6) always hold. There may be "extrabinomial" variation. In this case, standard errors will tend to be underestimated and sample size estimates will be too low, particularly in the case of dependent Bernoulli trials. Such data are not easy to analyze; sometimes a logarithmic transformation is used to stabilize the variance.

PROBLEMS

15.1 This problem will give practice by asking you to carry out analyses similar to the ones in each of the sections. The numbers from the National Cancer Institute [1975] for lung cancer cases for white males in the Pittsburgh and Detroit SMSAs are given in Table 15.5.

Table 15.5 Lung Cancer Cases by Age for White Males in the Detroit and Pittsburgh SMSAs

	Detroit		Pittsburgh	
Age	Cases	Population Size	Cases	Population Size
<5	0	149,814	0	82,242
5–9	0	175,924	0	99,975
10–14	2	189,589	1	113,146
15–19	0	156,910	0	100,139
20–24	5	113,003	0	68,062
25–29	1	113,919	0	61,254
30–34	10	92,212	7	53,289
35–39	24	90,395	21	55,604
40–44	101	108,709	56	70,832
45–49	198	110,436	148	74,781
50–54	343	98,756	249	72,247
55–59	461	82,758	368	64,114
60–64	532	63,642	470	50,592
65–69	572	47,713	414	36,087
70–74	473	35,248	330	26,840
75–79	365	25,094	259	19,492
80–84	133	12,577	105	10,987
>85	51	6,425	52	6,353
Total	3271	1,673,124	2480	1,066,036

(a) Carry out the analyses of Section 15.2 for these SMSAs.

(b) Calculate the direct and indirect standardized rates for lung cancer for white males adjusted for age. Let the Detroit SMSA be the study population and the Pittsburgh SMSA be the reference population.

(c) Compare the rates obtained in part (b) with those obtained in part (a).

15.2 **(a)** Calculate crude rates and standardized cancer rates for the white males of Table 15.5 using black males of Table 15.3 as the reference population.

(b) Calculate the standard error of the indirect standardized mortality rate and test whether it is different from 1.

(c) Compare the standardized mortality rates for blacks and whites.

15.3 The data in Table 15.6 represent the mortality experience for farmers in England and Wales 1949–1953 as compared with national mortality statistics.

Table 15.6 Mortality Experience Data for Problem 15.3

Age	National Mortality (1949–1953) Rate per 100,000/Year	Population of Farmers (1951 Census)	Deaths in 1949–1953
20–24	129.8	8,481	87
25–34	152.5	39,729	289
35–44	280.4	65,700	733
45–54	816.2	73,376	1,998
55–64	2,312.4	58,226	4,571

(a) Calculate the crude mortality rates.

(b) Calculate the standardized mortality rates.

(c) Test the significance of the standardized mortality rates.

(d) Construct a 95% confidence interval for the standardized mortality rates.

(e) What are the units for the ratios calculated in parts (a) and (b)?

15.4 Problems for discussion and thought:

(a) Direct and indirect standardization permit comparison of rates in two populations. Describe in what way this can also be accomplished by multiway contingency tables.

(b) For calculating standard errors of rates, we assumed that events were binomially (or Poisson) distributed. State the assumption of the binomial distribution in terms of, say, the event "death from cancer" for a specified population. Which of the assumptions is likely to be valid? Which is not likely to be invalid?

(c) Continuing from part (b), we calculate standard errors of rates that are population based; hence the rates are not samples. Why calculate standard errors anyway, and do significance testing?

15.5 This problem deals with a study reported in Bunker et al. [1969]. Halothane, an anesthetic agent, was introduced in 1956. Its early safety record was good, but reports of massive hepatic damage and death began to appear. In 1963, a Subcommittee on the National Halothane Study was appointed. Two prominent statisticians, Frederick Mosteller and Lincoln Moses, were members of the committee. The committee designed a large cooperative retrospective study, ultimately involving 34 institutions

Table 15.7 Mortality Data for Problem 15.5

	Number of Operations			Number of Deaths		
Physical Status	Total	Halothane	Cyclopropane	Total	Halothane	Cyclopropane
Unknown	69,239	23,684	10,147	1,378	419	297
1	185,919	65,936	27,444	445	125	91
2	104,286	36,842	14,097	1,856	560	361
3	29,491	8,918	3,814	2,135	617	403
4	3,419	1,170	681	590	182	127
5	21,797	6,579	7,423	314	74	101
6	11,112	2,632	3,814	1,392	287	476
7	2,137	439	749	673	111	253
Total	427,400	146,200	68,169	8,783	2,375	2,109

that completed the study. "The primary objective of the study was to compare halothane with other general anesthetics as to incidence of fatal massive hepatic necrosis within six weeks of anesthesia." A four-year period, 1959–1962, was chosen for the study. One categorization of the patients was by physical status at the time of the operation. Physical status varies from good (category 1) to moribund (category 7). Another categorization was by mortality level of the surgical procedure, having values of low, middle, high. The data in Table 15.7 deal with middle-level mortality surgery and two of the five anesthetic agents studied, the total number of administrations, and the number of patients dying within six weeks of the operation.

(a) Calculate the crude death rates per 100,000 per year for total, halothane, and cyclopropane. Are the crude rates for halothane and cyclopropane significantly different?

(b) By direct standardization (relative to the total), calculate standardized death rates for halothane and cyclopropane. Are the standardized rates significantly different?

(c) Calculate the standardized mortality rates for halothane and cyclopropane and test the significance of the difference.

(d) The calculations of the standard errors of the standardized rates depend on certain assumptions. Which assumptions are likely not to be valid in this example?

15.6 In 1980, 45 SIDS (sudden infant death syndrome) deaths were observed in King County. There were 15,000 births.

(a) Calculate the SIDS rate per 100,000 births.

(b) Construct a 95% confidence interval on the SIDS rate per 100,000 using the Poisson approximation to the binomial.

(c) Using the normal approximation to the Poisson, set up the 95% limits.

(d) Use the square root transformation for a Poisson random variable to generate a third set of 95% confidence intervals. Are the intervals comparable?

(e) The SIDS rate in 1970 in King County is stated to be 250 per 100,000. Someone wants to compare this 1970 rate with the 1980 rate and carries out a test of two proportions, $p_1 = 300$ per 100,000 and $p_2 = 250$ per 100,000, using the binomial distributions with $N_1 = N_2 = 100{,}000$. The large-sample normal approximation is used. What part of the Z-statistic: $(p_1 - p_2)$/standard error$(p_1 - p_2)$ will be right? What part will be wrong? Why?

Table 15.8 Heart Disease Data for Problem 15.7

Gender	Age	Epileptics: Person-Years at Risk	New and Nonfatal IHD Cases	Incidence in General Population per 100,000/year
Male	30–39	354	2	76
	40–49	303	2	430
	50–59	209	3	1291
	60–69	143	4	2166
	70+	136	4	1857
Female	30–39	534	0	9
	40–49	363	1	77
	50–59	218	3	319
	60–69	192	4	930
	70+	210	2	1087

15.7 Annegers et al. [1976] investigated ischemic heart disease (IHD) in patients with epilepsy. The hypothesis of interest was whether patients with epilepsy, particularly those on long-term anticonvulsant medication, were at less than expected risk of ischemic heart disease. The study dealt with 516 cases of epilepsy; exposure time was measured from time of diagnosis of epilepsy to time of death or time last seen alive.

(a) For males aged 60 to 69, the number of years at risk was 161 person-years. In this time interval, four IHD deaths were observed. Calculate the hazard rate for this age group in units of 100,000 persons/year.

(b) Construct a 95% confidence interval.

(c) The expected hazard rate in the general population is 1464 per 100,000 persons/year. How many deaths would you have expected in the age group 60 to 69 on the basis of the 161 person-years experience?

(d) Do the number of observed and expected deaths differ significantly?

(e) The raw data for the incidence of ischemic heart disease are given in Table 15.8. Calculate the expected number of deaths for males and the expected number of deaths for females by summing the expected numbers in the age categories (for each gender separately). Treat the total observed as a Poisson random variable and set up 95% confidence intervals. Do these include the expected number of deaths? State your conclusion.

(f) Derive a formula for an indirect standardization of these data (see Note 15.2) and apply it to these data.

15.8 A random sample of 100 subjects from a population is divided into two age groups, and for each age group the number of cases of a certain disease is determined. A reference population of 2000 persons has the following age distribution:

	Sample		Reference Population
Age	Total Number	Number of Cases	Total Number
1	80	8	1000
2	20	8	1000

(a) What is the crude case rate per 1000 population for the sample?

(b) What is the standard error of the crude case rate?

(c) What is the age-adjusted case rate per 1000 population using direct standardization and the reference population above?

(d) How would you test the hypothesis that the case rate at age 1 is not significantly different from the case rate at age 2?

15.9 The data in Table 15.9 come from a paper by Friis et al. [1981]. The mortality among male Hispanics and non-Hispanics was as shown.

Table 15.9 Mortality Data for Problem 15.9

	Hispanic Males		Non-Hispanic Males	
Age	Number	Number of Deaths	Number	Number of Deaths
0–4	11,089	0	51,250	0
5–14	18,634	0	120,301	0
15–24	10,409	0	144,363	2
25–34	16,269	2	136,808	9
35–44	11,050	0	106,492	46
45–54	6,368	7	91,513	214
55–64	3,228	8	70,950	357
65–74	1,302	12	34,834	478
75+	1,104	27	16,223	814
Total	79,453	56	772,734	1,920

(a) Calculate the crude death rate among Hispanic males.

(b) Calculate the crude death rate among non-Hispanic males.

(c) Compare parts (a) and (b) using an appropriate test.

(d) Calculate the SMR using non-Hispanic males as the reference population.

(e) Test the significance of the SMR as compared with a ratio of 1. Interpret your results.

15.10 The data in Table 15.10, abstracted from National Center for Health Statistics [1976], deal with the mortality experience in poverty and nonpoverty areas of New York and Seattle.

(a) Using New York City as the "standard population," calculate the standardized mortality rates for Seattle taking into account race and poverty area.

(b) Estimate the variance of this quantity and calculate 99% confidence limits.

(c) Calculate the standardized death rate per 100,000 population.

Table 15.10 Mortality Data for Problem 15.10

		New York City		Seattle	
Area	Race	Population	Death Rate per 1000	Population	Death Rate per 1000
Poverty	White	974,462	9.9	29,016	22.9
	All others	1,057,125	8.5	14,972	12.5
Nonpoverty	White	5,074,379	11.6	434,854	11.7
	All other	788,897	6.4	51,989	6.5

(d) Interpret your results.

(e) Why would you caution a reviewer of your analysis about the interpretation?

15.11 In a paper by Foy et al. [1983] the risk of getting *Mycoplasma* pneumonia in a two-year interval was determined on the basis of an extended survey of schoolchildren. Of interest was whether children previously exposed to *Mycoplasma pneumoniae* had a smaller risk of recurrence. In the five- to nine-year age group, the following data were obtained:

	Exposed Previously	**Not Exposed Previously**
Person-years at risk	680	134
Number with *Mycoplasma* pneumonia	7	8

(a) Calculate 95% confidence intervals for the infection rate per 100 person-years for each of the two groups.

(b) Test the significance of the difference between the infection rates.

***(c)** A statistician is asked to calculate the study size needed for a new prospective study between the two groups. He assumes that $\alpha = 0.05$, $\beta = 0.20$, and a two-tailed, two-sample test. He derives the formula

$$\lambda_2 = \sqrt{\lambda_1} - \frac{2.8}{\sqrt{n}}$$

where λ_i is the two-year infection rate for group i and n is the number of persons per group. He used the fact that the square root transformation of a Poisson random variable stabilizes the variance (see Section 10.6). Derive the formula and calculate the infection rate in group 2, λ_2 for $\lambda_1 = 10$ or 6, and sample sizes of 20, 40, 60, 80, and 100.

15.12 In a classic paper dealing with mortality among women first employed before 1930 in the U.S. radium dial–painting industry, Polednak et al. [1978] investigated 21 malignant neoplasms among a cohort of 634 women employed between 1915 and 1929. The five highest mortality rates (observed divided by expected deaths) are listed in Table 15.11.

(a) Test which ratios are significantly different from 1.

(b) Assuming that the causes of death were selected without a particular reason, adjust the observed p-values using an appropriate multiple-comparison procedure.

(c) The painters had contact with the radium through the licking of the radium-coated paintbrush to make a fine point with which to paint the dial. On the basis of this

Table 15.11 Mortality Data for Problem 15.12

Ranked Cause of Death	Observed Number	Expected Number	Ratio
Bone cancer	22	0.27	81.79
Larynx	1	0.09	11.13
Other sites	18	2.51	7.16
Brain and CNS	3	0.97	3.09
Buccal cavity, pharynx	1	0.47	2.15

information, would you have "preselected" certain malignant neoplasms? If so, how would you "adjust" the observed p-value?

15.13 Consider the data in Table 15.12 (from Janerich et al. [1974]) listing the frequency of infants with Simian creases by gender and maternal smoking status.

Table 15.12 Influence of Smoking on Development of Simian Creases

Gender of Infant	Maternal Smoking	Birthweight Interval (lb)			
		<6	6–6.99	7–7.99	≥8
Female	No	2/45	5/156	9/242	11/216
	Yes	4/48	8/107	6/110	3/44
Male	No	5/40	5/109	23/265	18/278
	Yes	10/55	6/84	10/106	6/74

(a) These data can be analyzed by the multidimensional contingency table approach of Chapter 7. However, we can also treat it as a problem in standardization. Describe how indirect standardization can be carried out using the total sample as the reference population, to compare "risk" of Simian creases in smokers and nonsmokers adjusted for birthweight and gender of the infants.

(b) Carry out the indirect standardization procedure and compare the standardized rates for smokers and nonsmokers. State your conclusions.

(c) Carry out the logistic model analysis of Chapter 7.

***15.14** Show that the variance of the standardized mortality ratio, equation (3), is approximately equal to equation (4).

REFERENCES

Annegers, J. F., Elveback, L. R., Labarthe, D. R., and Hauser, W. A. [1976]. Ischemic heart disease in patients with epilepsy. *Epilepsia*, **17**: 11–14.

Bruce, E., Frederick, R., Bruce, R., and Fisher, L. D. [1976]. Comparison of active participants and dropouts in CAPRI cardiopulmonary rehabilitation programs. *American Journal of Cardiology*, **37**: 53–60.

Bunker, J. P., Forest, W. H., Jr., Mosteller, F., and Vandam, L. D. [1969]. *The National Halothane Study: A Study of the Possible Association between Halothane Anesthesia and Postoperative Hepatic Necrosis*. National Institute of Health/National Institute of Several Medical Sciences, Bethesda, MD.

Clark, D. A., Stinson, E. B., Griepp, R. B., Schroeder, J. S., Shumway, N. E., and Harrison, D. C. [1971]. Cardiac transplantation: VI. Prognosis of patients selected for cardiac transplantation. *Annals of Internal Medicine*, **75**: 15–21. Used with permission.

Foy, H. M., Kenny, G. E., Cooney, M. K., Allan, I. D., and van Belle, G. [1983]. Naturally acquired immunity to mycoplasma pneumonia infections. *Journal of Infectious Diseases*, **147**: 967–973. Used with permission from University of Chicago Press.

Friis, R., Nanjundappa, G., Prendergast, J. J., Jr., and Welsh, M. [1981]. Coronary heart disease mortality and risk among hispanics and non-hispanics in Orange County, CA. *Public Health Reports*, **96**: 418–422.

Janerich, D. T., Skalko, R. G., and Porter, I. H. (eds.) [1974]. *Congenital Defects: New Directions in Research*. Academic Press, New York.

Kalbfleisch, J. D., and Prentice, R. L. [2002]. *The Statistical Analysis of Failure Time Data*, 2nd ed. Wiley, New York.

National Cancer Institute [1975]. *Third National Cancer Survey: Incidence Data*. Monograph 41. DHEW Publication (NIH) 75–787. U.S. Government Printing Office, Washington, DC.

National Center for Health Statistics [1976]. *Selected Vital and Health Statistics in Poverty and Non-poverty Areas of 19 Large Cities: United States, 1969–1971*. Series 21, No. 26. U.S. Government Printing Office, Washington, DC.

Polednak, A. P., Stehney, A. F., and Rowland, R. E. [1978]. Mortality among women first employed before 1930 in the U.S. radium dial-painting industry. *American Journal of Epidemiology*, **107**: 179–195.