INFANT MORTALITY RATES AS ESTIMATES: THEIR

BIASES AND VARIANCES

BY

Z.W. BIRNBAUM
MONROE G. SIRKEN

TECHNICAL REPORT NO. 4
JANUARY 1981

RESEARCH SUPPORTED BY THE U.S.A. NATIONAL CENTER FOR HEALTH STATISTICS

DEPARTMENT OF STATISTICS
UNIVERSITY OF WASHINGTON
SEATTLE, WASHINGTON 98195
Abstract

To estimate for a given population and time period the probability of an infant death, i.e. death at age less than one year, it has been customary to use statistics known as Infant Mortality Rates. It is well known that these statistics may be biased and that their biases depend, among others, on changes in birthrates. This paper explores in some detail the factors which determine the biases of most frequently used infant mortality rates; furthermore, variances of these mortality rates are derived, as an aid for judging the significance of differences between such rates.

Key Words: Infant Mortality, Biased Estimates, Variances of Mortality Rates
Infant Mortality Rates as Estimates; their biases and variances

by

Z.W. Birnbaum and Monroe G. Sirken

0. Introduction.

An infant death is usually defined as the death of a live born individual occurring at the age of no more than one year. It is of obvious interest to obtain, from whatever data are available, estimates of probabilities of infant deaths for various populations and years (or other time intervals) and to compare these probabilities.

The statistics commonly used to estimate such probabilities are known as Infant Mortality Rates. It is well known (See, for example, Moriyama and Greville [2]) that these statistics may be biased and their biases depend, among others, on changes in birthrates. In this paper we explore in some detail the factors which determine the biases of infant mortality rates; furthermore, as an aid in concluding whether differences between infant mortality rates are significant, we derive expressions for their variances.

1. Some Probabilities.

We consider three consecutive calendar years, e.g. 1976, 1977, 1978, sometimes to be called the past year, the present year, and the following year. On a time axis, these years will be represented by the intervals (-1,0), (0,1), (1,2) respectively, so that the value $t = -\frac{1}{2}$, say, will correspond to a date at the middle of the past year.

*) Research supported by the U.S.A. National Center for Health Statistics.
Infant deaths in the present year (0,1) may occur to individuals born either in the past year (-1,0) or to those born in the present year (0,1). Similarly, infant deaths of those born in the present year may occur in the present or in the following year.

To define the probability of infant death in the present year (0,1), we consider two random variables: the date of birth $T$ and the life-length $X$. For an individual born at time $T$ and with life-length $X$, the date of death will be $T + X$. The random variables $T$ and $X$ may be dependent.

In the $(T, X)$-plane we consider regions, indicated on Fig. 1, which correspond to the following events:

- **G**: $0 < T < 1$ and $T + X < 1$, i.e. an individual is born in the present year and in the present year (hence at age $< 1$),

- **H**: $0 < T < 1$ and $1 - T < X < 1$, i.e. an individual is born in the present year and dies in the following year at age $< 1$.

- **K**: $0 < T < 1$ and $X > 1$, i.e. an individual is born in the present year and dies at age $> 1$,

- **L**: $-1 < T < 0$ and $T + X < 0$, i.e. an individual is born in the past year and dies in the past year (hence at age $< 1$).

- **M**: $-1 < T < 0$ and $-T < X < 1$, i.e. an individual is born in the past year and dies in the present year at age $< 1$,

- **N**: $-1 < T < 0$ and $1 < X < 1 - T$, i.e. an individual is born in the past year and dies in the present year at age $> 1$,

- **Q**: $-1 < T < 0$ and $T + X > 1$, i.e. an individual is born in the past year and dies in the following year or later.
We consider two more regions

\[ B_1 = G \cup H \cup K, \quad B_0 = L \cup M \cup N \cup Q \]

which correspond to the events:

- \( B_1 : 0 < T \leq 1, \) i.e. an individual is born in the present year,
- \( B_0 : -1 < T \leq 0, \) i.e. and individual is born in the past year.

A plausible definition of infant death probability for the present year \((0,1)\) may be stated in the form of a conditional probability

\[ p_1 = P\{X \leq 1 | 0 < T \leq 1\}, \]

that is the probability that an individual, born in the present year, dies at age \( \leq 1. \) One clearly has

\[ p_1 = \frac{P\{G\} + P\{H\}}{P\{B_1\}}. \]

Since we consider only births in the years \((-1,0)\) and \((0,1)\) and \( X \) is a non-negative random variable, we shall assume that

\[ P\{-1 < T \leq 1\} = P\{B_0\} + P\{B_1\} = 1 \]

As alternatives to (1.1) one may also consider one of the following definitions of the probability of infant death for the year \((0,1)\)

\[ p'_1 \quad \text{def} = \text{Prob} \{\text{an individual, alive at some time in } (0,1) \text{ will die in } (0,1) \text{ at age } \leq 1\} = \frac{P\{M\} + P\{G\}}{P\{M\} + P\{N\} + P\{Q\} + P\{B_1\}} \]

or

\[ p''_1 \quad \text{def} = \text{Prob} \{\text{an individual, born in } (-1,1), \text{ will die at age } \leq 1\} = \frac{P\{L\} + P\{M\} + P\{G\} + P\{H\}}{P\{B_0\} + P\{B_1\}} \]

\[ = P\{L\} + P\{M\} + P\{G\} + P\{H\}, \]
or

\[
\begin{align*}
(1.6) \quad p''_1 &= \text{Prob} \{\text{an individual, born in } (-1,1), \text{ will die} \\
&\text{in } (0,1) \text{ at age } \leq 1\} = \frac{P\{M\} + P\{G\}}{P\{B_0\} + P\{B_1\}} \\
&= P\{M\} + P\{G\}
\end{align*}
\]

We shall be mainly interested in (1.1), but (1.6) will be used in Section 4, to compute the variances of \(r\).

2. **Estimates, rates, and separation factors.**

From now on it will be assumed that the data available for a given population consist of

(i) all death certificates for the present year \((0,1)\) and the past year \((-1,0)\), possibly also for earlier years, but not for the following year \((1,2)\) or later years; and that these death certificates contain dates of birth and death (hence also life-lengths), and

(ii) all birth certificates for the same years, containing dates of birth.

We shall not concern ourselves with possible incompleteness or errors in these certificates -- for our purposes it will be assumed that they are complete and correct.

These data enable us to count individuals falling into some of the regions in the \((T,X)\) -- plane. Denoting by \(\#(R)\) the number of observations in any region \(R\), we see that we can obtain from the death certificates the frequencies \(\#(G), \#(L), \#(M), \#(N)\), and from the birth certificates the frequencies \(\#(B_0), \#(B_1)\), but the data will not enable us to obtain \(\#(H), \#(K)\).
By combining the information from death – and birth-certificates, we can compute

$$\#(Q) = \#(B_o) - \#(L) - \#(M) - \#(N)$$

and

$$\#(H) + \#(K) = \#(B_1) - \#(G).$$

In view of (1.3) and of (1.2), (1.4), (1.5), (1.6) the usual binomial estimates of probabilities $p_1, p'_1, p''_1, p'''_1$ are

(2.1) \[ \hat{p}_1 = \frac{\#(G) + \#(H)}{\#(B_1)} \]

(2.2) \[ \hat{p}'_1 = \frac{\#(G) + \#(M)}{\#(M) + \#(N) + \#(Q) + \#(B_1)} = \frac{\#(G) + \#(M)}{\#(B_0) - \#(L) + \#(B_1)} \]

(2.3) \[ \hat{p}''_1 = \frac{\#(G) + \#(H) + \#(L) + \#(M)}{\#(B_0) + \#(B_1)} \]

(2.4) \[ \hat{p}'''_1 = \frac{\#(G) + \#(M)}{\#(B_0) + \#(B_1)} \]

Of these four estimates, only $\hat{p}_1$ and $\hat{p}'''_1$ can be computed directly from our data.

The statistic most frequently used to describe quantitatively infant mortality in a given population for a given year is the Infant Mortality Rate (IMR) which, in its simplest form, is defined by

(2.5) \[ r = \frac{\text{number of infant deaths in the present year}}{\text{number of live births in the present year}} \]

Using our notations, this can be written in the form

(2.6) \[ r = \frac{\#(G) + \#(M)}{\#(B_1)} \]
This statistic is not equal to any of the four estimates (2.1) - (2.4).

The numerator in (2.6) keeps count of infant deaths occurring in (0,1) among those born in (-1,0) as well as among those born in (0,1), while the denominator counts only the births in (0,1) hence does not include all those at risk in (0,1). To adjust for this, it has been frequent practice to use "separation factors"

\[ f = \frac{\#(M)}{\#(M) + \#(G)} \]

(2.7)

\[ 1 - f = \frac{\#(G)}{\#(M) + \#(G)} \]

and to compute the "adjusted IMR"\(^2\)

\[ r^* = \frac{\#(G) + \#(M)}{\#(B_1)} \cdot (1 - f) + \frac{\#(G) + \#(M)}{\#(B_0)} \cdot f = \]

(2.8)

\[ = \frac{\#(G)}{\#(B_1)} + \frac{\#(M)}{\#(B_0)}. \]

The separation factor \( f \) is itself a random variable, but in practice it is mostly taken from extraneous demographic studies, so that \( f \) and \((1-f)\) in (2.8) become constant weights, independent of the death- and birth-certificates used.

3. **Asymptotic values and biases of infant mortality rates.**

Writing (2.6) in the form

\[ r = \frac{\left[ \frac{\#(G) + \#(M)}{\#(B_0) + \#(B_1)} \right]}{\#(B_1) / \left[ \frac{\#(B_0) + \#(B_1)}{\#(B_1) + \#(B_0)} \right]} \]
and keeping in mind (1.3) one sees that \( r \) tends in probability to

\[
(3.1) \quad \rho = \frac{\mathbb{P}(G) + \mathbb{P}(M)}{\mathbb{P}(B_1)}.
\]

If used to estimate the infant death probability \( p_1 \), computed according to (1.2), the IMR has therefore the asymptotic bias

\[
(3.2) \quad \rho - p_1 = \frac{\mathbb{P}(M) - \mathbb{P}(H)}{\mathbb{P}(B_1)} = b.
\]

Similarly it can be seen that \( r^* \), the adjusted IMR (2.8), tends in probability to

\[
(3.3) \quad \rho^* = \frac{\mathbb{P}(G)}{\mathbb{P}(B_0)} + \frac{\mathbb{P}(G)}{\mathbb{P}(B_1)}
\]

and, when used to estimate \( p_1 \), has the asymptotic bias

\[
(3.4) \quad \rho^* - p_1 = \frac{\mathbb{P}(M)}{\mathbb{P}(B_0)} - \frac{\mathbb{P}(H)}{\mathbb{P}(B_1)} = b^*.
\]

Either one of \( b \) or \( b^* \) may be greater than the other, since

\[
(3.5) \quad b^* - b = \mathbb{P}(M) \left[ \frac{1}{\mathbb{P}(B_0)} - \frac{1}{\mathbb{P}(B_1)} \right].
\]

As can be seen from (3.2) and (3.4), the asymptotic biases are determined only by the probabilities \( \mathbb{P}(G) \), \( \mathbb{P}(H) \), \( \mathbb{P}(M) \), and \( \mathbb{P}(B_1) \), since by (1.3) one has \( \mathbb{P}(B_0) = 1 - \mathbb{P}(B_1) \). To see what effect changes of these probabilities from one year to the next have on the biases \( b \) and \( b^* \), we write

\[
\rho_0 = \frac{\mathbb{P}(L) + \mathbb{P}(M)}{\mathbb{P}(B_0)} = \text{infant death probability in the past year}
\]
\[ P(B_1) - P(B_0) = \gamma = \text{change of birth probability} \]

\[ P_1 - P_0 = \delta = \text{change of infant death probability} \]

\[ \frac{P(H)}{P(B_1)} - \frac{P(M)}{P(B_0)} = \epsilon \]

\[ \frac{P(G)}{P(B_1)} - \frac{P(M)}{P(B_0)} = \psi \]

where \( \epsilon \) has the following intuitive meaning: \( \frac{P(H)}{P(B_1)} \) as well as \( \frac{P(M)}{P(B_0)} \) denotes the probability that an individual born in one year will die during the next year at an age of no more than one year, hence \( \epsilon \) is the change of this probability from the past to the present year. A similar interpretation can be given for \( \psi \) and one has

\[ \delta = \epsilon + \psi \]

Using the quantities defined in (3.6) and noting that by (1.3) one has

\[ P(B_0) = (1-\gamma)/2, \quad P(B_1) = (1+\gamma)/2, \]

one obtains for the biases of \( r^* \) and of \( r \) the expressions

\[ b^* = -\epsilon \]

\[ b = b^* - P(M) \cdot \frac{4\gamma}{1-\gamma^2} = - (\epsilon + P(M) \cdot \frac{4\gamma}{1-\gamma^2}) \]

Some immediate conclusions from these expressions are:

- \( b = b^* \) if and only if \( \gamma = 0 \),
- \( b > b^* \) (\( b < b^* \)) if and only if \( \gamma < 0 \) (\( \gamma > 0 \)),
- \( b^* = 0 \iff \epsilon = 0 \)
- \( \gamma = \delta = 0 \) is not sufficient to assure that either \( b = 0 \) or that \( b^* = 0 \),

the value of \( \gamma \) has no effect on \( b^* \), which depends on changes in birth probability only insofar as they affect \( \epsilon \).
4. Variances of $r$ and $r^*$

The frequencies of events in the numerators of (2.6) and (2.8) and of those in the denominators in these expressions come from different sources: $\#(G)$ and $\#(M)$ are obtained from death certificates, while $\#(B_0)$ and $\#(B_1)$ are taken from birth certificates. We write (2.6) in the form

$$ r = \frac{\#(G) + \#(M)}{\#(B_0) + \#(B_1)} \cdot \frac{\#(B_0) + +(B_1)}{\#(B_1)} $$

The first factor is the binomial estimate of the probability $p_1'''$ given by (1.6), while the second factor is a number computed from birth certificates.

If one considers $\#(B_0)$ and $\#(B_1)$ as numbers of individuals at risk, among whom the events $G$ and $M$ occur, then one has

$$ \text{var}(r) = \frac{p_1'''(1-p_1''')}{\#(B_0) \cdot \#(B_1)} \cdot \frac{\[(\#(B_0) + \#(B_1))^2 - \#(B_1)^2\]}{\#(B_1)^2} $$

(4.1)

To derive the variance of $r^*$, we observe that $\#(G)$ is the number of infant deaths occurring in $(0,1)$ among those born in $(0,1)$, while $\#(M)$ is the number of such deaths among those born in $(-1,0)$. The random variables $\#(G)$ and $\#(M)$ are therefore independent. Using the notations

$$ p_{11} = P(G)/P(B_1), \quad p_{10} = P(M)/P(B_0) $$

we have therefore, from (2.8)

$$ \text{var}(r^*) = \frac{p_{11}(1-p_{11})}{\#(B_1)} + \frac{p_{10}(1-p_{10})}{\#(B_0)} $$

(4.2)
In practice one will have to replace in (4.1) and (4.2) the probabilities \( p_{1}''', p_{11} \) and \( p_{10} \) by their binomial estimates
\[
\hat{p}_{1}''' = \frac{\#(G) + \#(M)}{\#(B_{0}) + \#(B_{1})}, \quad \hat{p}_{11} = \frac{\#(G)}{\#(B_{1})}, \quad \hat{p}_{10} = \frac{\#(M)}{\#(B_{0})}
\]
to obtain approximate values of \( \text{var}(r) \) and \( \text{var}(r^*) \).

5. Some conclusions and an illustration.

Our discussion shows that neither the unadjusted IMR \( r \) nor the adjusted IMR \( r^* \) is, in general, a consistent estimate of \( p_{1} \) or, as can be seen by similar arguments, or any of the other probabilities (1.4) – (1.6). Assumptions which would assure consistency, such as assuming \( \epsilon = 0 \), are close to assuming periodicity from year to year.

The difficulty in assuring consistency is due to the fact that in \( r \) or \( r^* \) one does not use data corresponding to region \( H \) in Fig. 1., that is information from death certificates for the following year. By using this information when it is available and defining the infant mortality rate as the statistic
\[
(5.1) \quad r^{**} = \frac{\#(G) + \#(H)}{\#(B_{1})}
\]
one would avoid this difficulty, since (7.1) is the usual binomial estimate of \( p_{1} \) mentioned already in (2.1).

The main reason why infant mortality rates are used instead of the unbiased estimate of \( p_{1} \) given by (5.1) appears to be that mortality data are usually published in a form which yields for each year the total number of infant deaths occurring in that year, i.e. \( \#(M) \rightarrow \#(r) \), but does not allow to separate that total into \( \#(M) \) and \( \#(G) \). Shryock and Siegel [3] illustrate several computations on data for Venezuela, one of the countries for which there are available mortality data grouped by the year of death and the year of birth.
Using the same source [4] for the years 1957 to 1963, we obtained the entries in the first four lines of Table 1. We shall not concern ourselves with possible sources of error which may be inherent in the original demographic data, but use those data only to carry out the calculations presented in the remainder of Table 1.

The following observations on the contents of our table may be of interest:

a) The rates $r$ and $r^*$ change from year to year in the same direction as $\hat{P}_1$, and the annual changes of $r^*$ are larger than those of $r$.

b) The biases $b$ and $b^*$ may go in opposite directions (as in 1958), and the bias $b^*$ of the adjusted IMR is in four out of five years absolutely greater than that of the conventional IMR.

c) The standard deviations of $\hat{P}_1$, $r$ and $r^*$ for each year are nearly equal.

d) $b$ and $b^*$ are often high multiples of $\sigma(\hat{P}_1)$.

e) Using $r^*$, one would conclude that infant mortality has substantially increased from 1958 to 1959, while $\hat{P}_1$ did not show a significant increase.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(#) (B)</td>
<td>284,080</td>
<td>291,747</td>
<td>324,739</td>
<td>335,199</td>
<td>344,989</td>
<td>341,324</td>
<td>353,546</td>
</tr>
<tr>
<td>(#) (G)</td>
<td>13,866</td>
<td>11,059</td>
<td>14,152</td>
<td>13,592</td>
<td>13,812</td>
<td>12,183</td>
<td>12,997</td>
</tr>
<tr>
<td>(#) (H)</td>
<td>5,516</td>
<td>5,615</td>
<td>4,628</td>
<td>4,434</td>
<td>3,847</td>
<td>13,953</td>
<td>13,953</td>
</tr>
<tr>
<td>(#) (M)</td>
<td>4,823</td>
<td>5,516</td>
<td>5,615</td>
<td>4,628</td>
<td>4,434</td>
<td>3,847</td>
<td>3,953</td>
</tr>
<tr>
<td>(\hat{\beta}_1)</td>
<td>.06823</td>
<td>.05715</td>
<td>.05783</td>
<td>.05330</td>
<td>.05119</td>
<td>.04727</td>
<td></td>
</tr>
<tr>
<td>annual change of (\hat{\beta}_1)</td>
<td>-.00515</td>
<td>.00068</td>
<td>-.00453</td>
<td>-.00211</td>
<td>-.00392</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(r)</td>
<td>.06579</td>
<td>.05681</td>
<td>.06087</td>
<td>.05387</td>
<td>.05289</td>
<td>.04696</td>
<td></td>
</tr>
<tr>
<td>annual change of (r)</td>
<td>-.00899</td>
<td>.00406</td>
<td>-.00700</td>
<td>-.0098</td>
<td>-.00593</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(r^*)</td>
<td>.05732</td>
<td>.06283</td>
<td>.05444</td>
<td>.05315</td>
<td>.04684</td>
<td>.04834</td>
<td></td>
</tr>
<tr>
<td>annual change of (r^*)</td>
<td>.00511</td>
<td>-.00839</td>
<td>-.00129</td>
<td>-.00631</td>
<td>.00150</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(b)</td>
<td>-.00244</td>
<td>-.00034</td>
<td>.00304</td>
<td>.00057</td>
<td>.00170</td>
<td>-.00031</td>
<td></td>
</tr>
<tr>
<td>(b^*)</td>
<td>.00017</td>
<td>.00500</td>
<td>.00114</td>
<td>.00196</td>
<td>-.00043</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\sigma(\hat{\beta}_1))</td>
<td>.00047</td>
<td>.00043</td>
<td>.00041</td>
<td>.00039</td>
<td>.00038</td>
<td>.00036</td>
<td></td>
</tr>
<tr>
<td>(\sigma(r))</td>
<td>.00043</td>
<td>.00043</td>
<td>.00039</td>
<td>.00039</td>
<td>.00038</td>
<td>.00038</td>
<td></td>
</tr>
<tr>
<td>(\sigma(r^*))</td>
<td>.00044</td>
<td>.00044</td>
<td>.00040</td>
<td>.00039</td>
<td>.00036</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(b/\sigma(\hat{\beta}_1))</td>
<td>-5.19</td>
<td>-.79</td>
<td>7.41</td>
<td>1.46</td>
<td>4.47</td>
<td>-.86</td>
<td></td>
</tr>
<tr>
<td>(b\hat{y}\sigma(\hat{\beta}_1))</td>
<td>.40</td>
<td>12.2</td>
<td>2.92</td>
<td>5.16</td>
<td>-1.19</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>