Twin Research and Human Genetics
Volume 16 ▌ Number 6 ▌ pp. 1107–1111 ▌ © The Authors 2013 ▌ doi:10.1017/thg.2013.65

Statistical Analyses of Monozygotic and Dizygotic Twinning Rates

Johan Fellman
Folkhälsan Institute of Genetics, Population Genetics Unit, Helsinki, Finland Hanken School of Economics, Helsinki, Finland

The French mathematician Bertillon reasoned that the number of dizygotic (DZ) pairs would equal twice the number of twin pairs of unlike sexes. The remaining twin pairs in a sample would presumably be monozygotic (MZ). Weinberg restated this idea and the calculation has come to be known as Weinberg's differential rule (WDR). The keystone of WDR is that DZ twin pairs should be equally likely to be of the same or the opposite sex. Although the probability of a male birth is greater than .5, the reliability of WDR's assumptions has never been conclusively verified or rejected. Let the probability for an opposite-sex (OS) twin maternity be p_O, for a same-sex (SS) twin maternity p_S and, consequently, the probability for other maternities $1 - p_S - p_O$. The parameter estimates $\hat{p}_O$ and $\hat{p}_S$ are relative frequencies. Applying WDR, the MZ rate is $m = p_S - p_O$ and the DZ rate is $d = 2p_O$, but the estimates $\hat{m}$ and $\hat{d}$ are not relative frequencies. The maximum likelihood estimators $\hat{p}_S$ and $\hat{p}_O$ are unbiased, efficient, and asymptotically normal. The linear transformations $\hat{m} = \hat{p}_S - \hat{p}_O$ and $\hat{d} = 2\hat{p}_O$ are efficient and asymptotically normal. If WDR holds they are also unbiased. For the tests of a set of m and d rates, contingency tables cannot be used. Alternative tests are presented and the models are applied on published data.

■ **Keywords:** maximum likelihood estimation, parameter transformation, Poisson distribution, confidence interval, Åland Islands

The French mathematician Bertillon (1874) reasoned that the number of dizygotic (DZ) pairs would equal twice the number of twin pairs of unlike sexes. The remaining twin pairs in a sample would presumably be monozygotic (MZ). Weinberg (1902) restated this idea and the calculation has come to be known as Weinberg's differential rule (WDR). The basis of WDR is that DZ twin pairs should be equally likely to be of the same or the opposite sex. The small excess of males at birth (about 106:100) is not usually thought to affect the validity of the theory (Bulmer, 1970). Departure of the sex ratio from 100 in either direction reduces the expected proportion of pairs of opposite sex, and thus the estimated number of DZ twin pairs of the same sex, in turn leading to an overestimation of the number of MZ twin pairs (Allen, 1981; Boklage, 1985; Bulmer, 1970; Little & Thompson, 1988).

Methods

Estimation of MZ and DZ Twinning Rates

The estimation is usually based on WDR, assuming that the probabilities of a male birth and female birth are equal and that the sexes within a twin set are independent. Although the probability of a male birth is greater than .5, the reliability of the WDR's assumptions has never been conclusively verified or rejected (Fellman & Eriksson, 2006). According to WDR, the total number of DZ twin maternities is twice the number of twin maternities with opposite-sex (OS) twin sets. The number of MZ twin sets is the difference between the number of same-sex (SS) and OS twin sets.

Let the total number of maternities be n, the number of SS twin sets be n_S, and the number of OS twin sets be n_O. One can assume that the distribution of n_S, n_O, and $n - n_s - n_o$ has a trinomial distribution. Let the probability for an OS twin maternity be p_O, for an SS twin maternity p_S and, consequently, the probability for other maternities $1 - p_S - p_O$. Applying WDR, the MZ rate is $m = p_S - p_O$ and the DZ rate is $d = 2p_O$. According to the maximum likelihood theory, the estimators of transformed parameters are the corresponding transformations of the estimators of the initial parameters. Assuming that WDR holds, we can

RECEIVED 20 June 2013; ACCEPTED 22 August 2013. First published online 24 September 2013.

ADDRESS FOR CORRESPONDENCE: Johan Fellman, Folkhälsan Institute of Genetics, Population Genetics Unit, POB 211, FI-00251 Helsinki, Finland. E-mail: fellman@hanken.fi

estimate p_O and p_S and use the formulae $\hat{m} = \hat{p}_S - \hat{p}_O$ and $\hat{d} = 2\hat{p}_O$. This method does not solve if the proposed estimators $\hat{m}$ and $\hat{d}$ are unbiased estimators of the MZ and DZ rates, respectively. The unbiasedness depends on the reliability of Weinberg's assumptions. The likelihood function is

$$L(p_S, p_O) = (1 - p_S - p_O)^{n-n_S-n_O} p_S^{n_S} p_O^{n_O}. \quad (1)$$

The log-likelihood function is

$$l(p_S, p_O) = (n - n_S - n_O)\ln(1 - p_S - p_O) + n_S \ln(p_S) + n_O \ln(p_O).$$

After differentiation, we obtain

$$\frac{\partial l}{\partial p_S} = -\frac{n - n_S - n_O}{1 - p_S - p_O} + \frac{n_S}{p_S} \quad \text{and}$$
$$\frac{\partial l}{\partial p_O} = -\frac{n - n_S - n_O}{1 - p_S - p_O} + \frac{n_O}{p_O}. \quad (2)$$

The maximum of $l(p_S, p_O)$ yields the estimators $\hat{p}_S = \frac{n_S}{n}$ and $\hat{p}_O = \frac{n_O}{n}$, being unbiased, efficient, and asymptotically normal. To obtain the covariance matrix, we differentiate the formulae in (2) once more and obtain

$$\frac{\partial^2 l}{\partial p_S^2} = -\frac{n - n_S - n_O}{(1 - p_S - p_O)^2} - \frac{n_S}{p_S^2},$$
$$\frac{\partial^2 l}{\partial p_O^2} = -\frac{n - n_S - n_O}{(1 - p_S - p_O)^2} - \frac{n_O}{p_O^2}, \quad \text{and}$$
$$\frac{\partial^2 l}{\partial p_O \partial p_S} = -\frac{n - n_S - n_O}{(1 - p_S - p_O)^2}.$$

Now, $E(n_S) = np_S$ and $E(n_O) = np_O$ and, consequently,

$$-E\left(\frac{\partial^2 l}{\partial p_S^2}\right) = \frac{n(1 - p_O)}{p_S(1 - p_S - p_O)},$$

$$-E\left(\frac{\partial^2 l}{\partial p_O^2}\right) = \frac{n(1 - p_S)}{p_O(1 - p_S - p_O)},$$

and

$$-E\left(\frac{\partial^2 l}{\partial p_O \partial p_S}\right) = \frac{n}{(1 - p_S - p_O)}.$$

The information matrix for the estimators $\hat{p}_S$ and $\hat{p}_O$ is

$$\mathbf{I}(\hat{p}_S, \hat{p}_O) = \begin{bmatrix} \frac{n(1 - p_O)}{p_S(1 - p_S - p_O)} & \frac{n}{(1 - p_S - p_O)} \\ \frac{n}{(1 - p_S - p_O)} & \frac{n(1 - p_S)}{p_O(1 - p_S - p_O)} \end{bmatrix} \quad (3)$$

and the covariance matrix is

$$\mathbf{C}(\hat{p}_S, \hat{p}_O) = \mathbf{I}^{-1}(\hat{p}_S, \hat{p}_O) = \frac{1}{n}\begin{bmatrix} p_S(1 - p_S) & -p_S p_O \\ -p_S p_O & p_O(1 - p_O) \end{bmatrix}. \quad (4)$$

The estimators $\hat{p}_S$ and $\hat{p}_O$ have the variances $\mathrm{Var}(\hat{p}_S) = \frac{p_S(1-p_S)}{n}$ and $\mathrm{Var}(\hat{p}_O) = \frac{p_O(1-p_O)}{n}$ and the correlation coefficient $\mathrm{Cor}(\hat{p}_S, \hat{p}_O) = -\sqrt{\frac{p_S p_O}{(1-p_S)(1-p_O)}}$.

The linear transformations $\hat{m} = \hat{p}_S - \hat{p}_O$ and $\hat{d} = 2\hat{p}_O$ are efficient and asymptotically normal and, if WDR holds, also unbiased. The estimators of these parameters are $\hat{m} = \hat{p}_S - \hat{p}_O = \frac{n_S - n_O}{n}$ and $\hat{d} = 2\hat{p}_O = \frac{2n_O}{n}$. The covariance matrix is

$$\mathbf{C}(\hat{m}, \hat{d}) = \begin{bmatrix} 1 & -1 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} \frac{p_S(1 - p_S)}{n} & -\frac{p_S p_O}{n} \\ -\frac{p_S p_O}{n} & \frac{p_O(1 - p_O)}{n} \end{bmatrix}\begin{bmatrix} 1 & 0 \\ -1 & 2 \end{bmatrix} =$$
$$= \begin{bmatrix} \frac{p_S + p_O - (p_S - p_O)^2}{n} & -\frac{2p_O(1 + p_S - p_O)}{n} \\ -\frac{2p_O(1 + p_S - p_O)}{n} & \frac{4p_O(1 - p_O)}{n} \end{bmatrix}. \quad (5)$$

We rewrite (5) by using the parameters m and d and obtain

$$\mathbf{C}(\hat{m}, \hat{d}) = \frac{1}{n}\begin{bmatrix} m(1 - m) + d & -d(1 + m) \\ -d(1 + m) & d(1 - d) + d \end{bmatrix}. \quad (6)$$

We see that

$$\mathrm{Var}(\hat{m}) = \frac{m(1 - m) + d}{n},$$
$$\mathrm{Var}(\hat{d}) = \frac{d(1 - d) + d}{n}, \quad \text{and}$$
$$\mathrm{Cov}(\hat{m}, \hat{d}) = -\frac{d(1 + m)}{n}. \quad (7)$$

Bulmer (1970) presented slightly different approximate variance formulae. He assumed that the numbers of SS and OS twin sets are Poisson distributed but ignored that the numbers of OS and SS twin sets are correlated. His formulae are $\mathrm{Var}(\hat{m}) = \frac{m+d}{n}$ and $\mathrm{Var}(\hat{d}) = \frac{2d}{n}$, yielding figures that, compared with ours, are slightly too large.

If $\hat{m}$ and $\hat{d}$ are considered to be relative frequencies, one obtains the formulae $\mathrm{Var}(\hat{m}) = \frac{m(1-m)}{n}$ and $\mathrm{Var}(\hat{d}) = \frac{d(1-d)}{n}$, which are incorrect. The differences between the correct and incorrect formulae are positive and of the same order of magnitude (n^{-1}) as the correct ones and, consequently, the simple formulae should not be used, not even when the data sets are large. In addition, the erroneous variances are too small and may yield quite misleading test results and confidence intervals (CIs). For $\hat{d}$, the ratio between the correct variance estimate and the erroneous one

https://doi.org/10.1017/thg.2013.65 Published online by Cambridge University Press

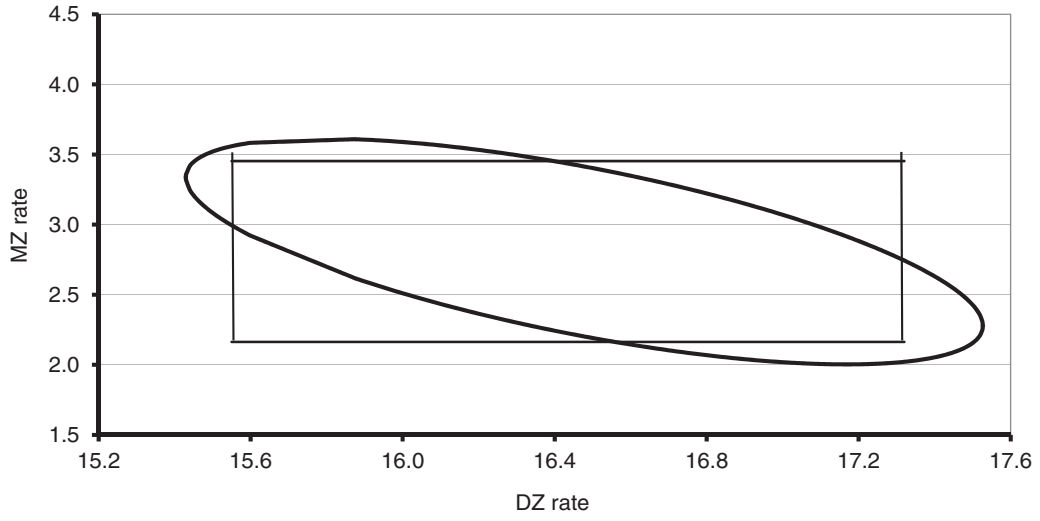

FIGURE 1

Confidence regions for MZ and DZ rates for Åland–Åboland , 1653–1949. Using SE($\hat{m}$) and SE($\hat{d}$), we can construct the 95% CIs for the MZ rate and the DZ rate. For the Åland–Åboland data, we get the interval (2.16, 3.45) for *m* and (15.64, 17.32) for *d*, and the rectangle is constructed according to the individual CIs for *m* and *d* (Fellman & Eriksson, 2006).

is $\frac{d(1-d)+d}{d(1-d)} = \frac{2-d}{1-d} \approx 2$, and for $\hat{m}$ it is $\frac{m(1-m)+d}{m(1-m)} \approx \frac{m+d}{m}$. The latter ratio is approximately the ratio between the total and the MZ twinning rate and can be considerably larger than the ratio for $\hat{d}$.

Testing

According to (7), the estimators are negatively correlated and the correlation is usually rather strong. Therefore, the simultaneous testing of both MZ and DZ rates is to be preferred and a joint confidence region is better than two separate CIs. The information matrix is the inverse of the covariance matrix formula (6), which is

$$\mathbf{I}(\hat{m}, \hat{d}) = n \begin{bmatrix} \frac{d(1-d)+d}{d(2m-3md-2m^2+d-d^2)} & \frac{d(1+m)}{d(2m-3md-2m^2+d-d^2)} \\ \frac{d(1+m)}{d(2m-3md-2m^2+d-d^2)} & \frac{m(1-m)+d}{d(2m-3md-2m^2+d-d^2)} \end{bmatrix} = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix}.$$

We construct the test statistic

$$\chi^2 = \left[((\hat{m}-m)(\hat{d}-d)\right] \mathbf{I}\left(\hat{m}, \hat{d}\right) \begin{bmatrix} \hat{m}-m \\ \hat{d}-d \end{bmatrix} = I_{11}\left(\hat{m}-m\right)^2 + 2I_{12}\left(\hat{m}-m\right)\left(\hat{d}-d\right) + I_{22}\left(\hat{d}-d\right)^2, \quad (8)$$

which is approximately χ^2 distributed with 2 degrees of freedom. If we want to test the given hypothetical rates m and d, we observe $\hat{m}$ and $\hat{d}$ and use the formula (8). It can also be used for the construction of a simultaneous confidence region for m and d. We observe that $P(I_{11}(\hat{m}-m)^2 + 2I_{12}(\hat{m}-m)(\hat{d}-d) + I_{22}(\hat{d}-d)^2 \leq \chi^2_\alpha) \geq 1-\alpha$ is a confidence region for the parameters (m, d) with the confidence level $1-\alpha$. The obtained region is an ellipse in the (m, d) plane (cf. Figure 1).

For tests of a set of MZ and DZ rates, contingence tables cannot be used. An approximate test of a set of MZ or DZ rates can be performed in the following way. Consider MZ rates. Under the null hypothesis that the MZ rates are constant, being m for $(t = 1, ..., T)$, $\mathrm{Var}(\hat{m}_t) = \frac{m(1-m)+d}{n_t}$, where n_t is the sample size for $t = 1, ..., T$. Define the weighted mean $\bar{m} = \sum_t \frac{n_t \hat{m}_t}{\sum n_t}$. Under the null hypothesis, $\bar{m}$ is the most efficient estimate of m. Consequently, $\chi^2 = \frac{1}{c}\sum_{t=1}^{T} n_t (m_t - \bar{m})^2$, where $c = m(1-m)+d$, is asymptotically χ^2 distributed with $T-1$ degrees of freedom. An approximate χ^2 test can be obtained if one uses $\hat{c} = \hat{m}(1-\hat{m}) + \hat{d}$ as an estimate of c. For $\hat{d}$ one obtains a similar formula, but now $c = d(1-d)+d$.

In numerical applications, the rates are usually given in per mille. This means that the theoretical variances should be scaled with a million and the estimators and standard errors with 1,000.

Results

Eriksson (1973, pp. 32 and 44) gave twinning data for the Åland Islands for 1653–1949 and for the neighboring archipelago of Åboland for 1655–1949.

Confidence Region

The total number of maternities is 178,055, the observed number of OS twin maternities is 1,467, and the observed number of SS twin maternities is 1,967. Using these data, we obtain (in per mille), $\hat{p}_O = 8.24$, $\hat{p}_S = 11.05$, $\hat{m} = 2.81$,

https://doi.org/10.1017/thg.2013.65 Published online by Cambridge University Press

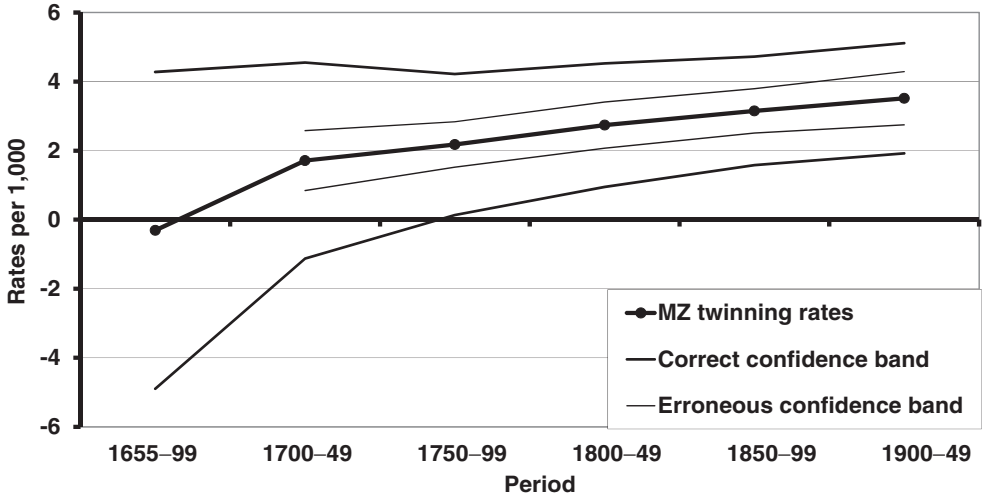

FIGURE 2

Estimated MZ twinning rates, including the correct (95%) confidence band for Åland–Åboland, 1653–1949, according to the data in Eriksson (1973). To emphasize the use of the correct SE formulae, the erroneous confidence band is also included in the figure. The temporal variation is statistically insignificant ($\chi^2 = 5.92$ with 5 degrees of freedom; $p > .05$) (cf. Fellman & Eriksson, 2006).

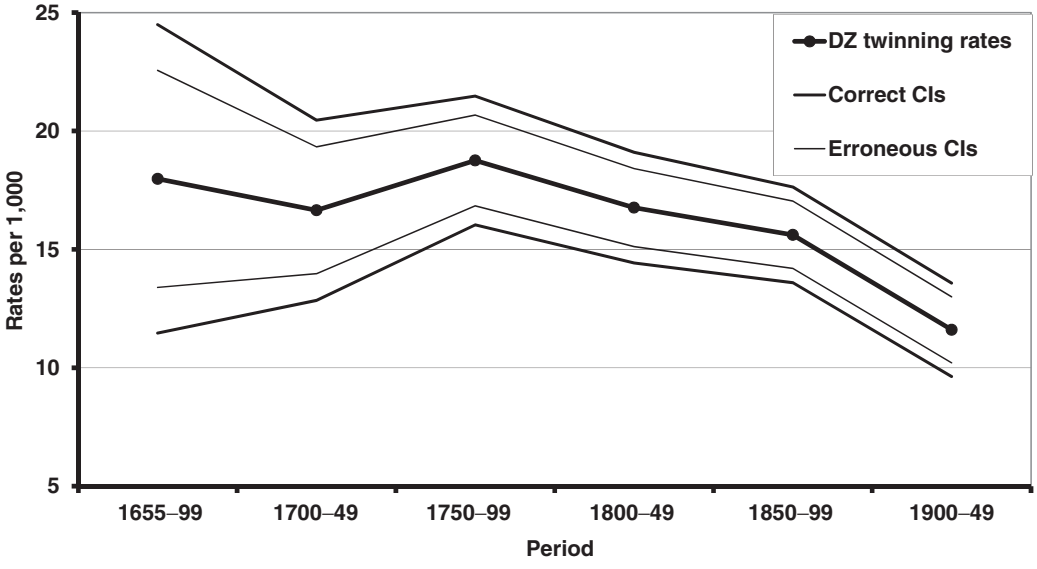

FIGURE 3

Estimated DZ twinning rates, including the correct (95%) confidence band for Åland–Åboland, 1653–1949, according to the data in Eriksson (1973). To emphasize the use of the correct SE formulae, the erroneous confidence band is also included in the figure. The temporal variation is statistically significant ($\chi^2 = 26.14$ with 5 degrees of freedom; $p < .001$), caused by a decreasing trend (Fellman & Eriksson, 2006).

and $\hat{d} = 16.48$. We obtain

$$\mathbf{C}(\hat{m}, \hat{d}) = \begin{bmatrix} 0.1083 & -0.0928 \\ -0.0928 & 0.1836 \end{bmatrix} \quad \text{and}$$

$$\mathbf{I}\left(\hat{m}, \hat{d}\right) = \begin{bmatrix} 16.2993 & 8.2404 \\ 8.2404 & 9.6138 \end{bmatrix}. \qquad (9)$$

Hence, $\mathrm{Var}(\hat{m}) = 0.1083$ and $\mathrm{Var}(\hat{d}) = 0.1836$ and the standard errors are $\mathrm{SE}(\hat{m}) = 0.329$ and $\mathrm{SE}(\hat{d}) = 0.428$. The correlation between these estimators is -0.658. The 95% confidence region is $16.2993\,(\hat{m} - m)^2 + 16.4808\,(\hat{m} - m)\left(\hat{d} - d\right) + 9.6138\left(\hat{d} - d\right)^2 \leq 5.99$. The confidence region is given in Figure 1.

Temporal Variations in the MZ and DZ Twinning Rates

The MZ twinning rate estimated by WDR has increased in many western European countries simultaneously with a decline in DZ rates. It has been suggested that the increase in MZ twinning rates in developed countries could be due to the introduction of oral contraception in the 1960s.

https://doi.org/10.1017/thg.2013.65 Published online by Cambridge University Press

Fellman and Eriksson (2006) noted that the increasing MZ rate observed in Åland–Åboland between 1653 and 1949 is statistically insignificant ($\chi^2 = 5.92$ with 5 degrees of freedom; $p > .05$) (see Figure 2). The temporal trend of the DZ twinning is given in Figure 3, and the temporal variation is statistically significant ($\chi^2 = 26.14$ with 5 degrees of freedom; $p < .001$), caused by a decreasing trend (Fellman & Eriksson, 2006).

Discussion

One also has to consider the assumed independence between the sexes in a DZ twin pair. James (1971) observed an excess of DZ twin pairs of the same sex, primarily due to males, claiming that WDR underestimates the number of DZ twins. Fellman and Eriksson (2006) have critically scrutinized WDR for the estimation of the MZ and DZ twinning rates. The basic assumptions for this rule are that the probability of a male birth is .5 and that the sexes within a DZ twin set are independent. These assumptions may be considered too simple, but their study indicates that WDR seems to be quite satisfactory, when large national birth register data are considered. Finally, the new standard errors for the estimated MZ and DZ rates should be used (cf. Figures 2 and 3). Especially, the standard errors for the MZ rate based on the correct formula are so large that variations other than random ones may be difficult to identify. If the WDR holds, the linearity indicates that the impact of external factors, such as maternal age and parity, seasonality, temporal effects, and ethnicity, has no disturbing effects on the analyses. Recently, Hardin et al. (2008) analyzed twin registers from several countries and their analyses of WDR yielded satisfactory results.

References

Allen, G. (1981). Errors of Weinberg's difference method. In L. Gedda, P. Parisi & W. E. Nance (Eds.), *Twin research 3: Part A, Twin biology and multiple pregnancy* (pp. 71–74). New York: Alan R. Liss.

Bertillon, M. (1874). Des combinaisons de sexe dans les grossesses gemellaires (doubles ou triples) de leur cause et de leur caractère ethnique. *Bulletins de la Societe d'Anthropologie de Paris*, *9*, 267–290.

Boklage, C. E. (1985). Interactions between opposite-sex dizygotic fetuses and the assumptions of Weinberg difference method epidemiology. *American Journal of Human Genetics*, *37*, 591–605.

Bulmer, M. G. (1970). *The biology of the twinning in man.* Oxford: Oxford University Press.

Eriksson, A. W. (1973). Human twinning in and around the Åland islands. *Commentationes Biologicae*, *64*, 1–159.

Fellman, J., & Eriksson, A. W. (2006). Weinberg's differential rule reconsidered. *Human Biology*, *78*, 253–275.

Hardin, J., Selvin, S., Carmichael, S. L., & Shaw, G. M. (2008). The estimated probability of dizygotic twins: A comparison of two methods. *Twin Research and Human Genetics*, *12*, 79–85.

James, W. H. (1971). Excess of like sexed pairs of dizygotic twins. *Nature*, *232*, 277–278.

Little, J., & Thompson, B. (1988). Descriptive epidemiology. In J. MacGillivray, D. M. Campbell, & B. Thompson (Eds.), *Twins and twinning* (pp. 37–66). Chichester, UK: John Wiley & Sons.

Weinberg, W. (1902). Beiträge zur Physiologie und Pathologie der Mehrlingsbeburten beim Menschen. *Archiv für die gesammte Physiologie des Menschen und der Thiere*, *88*, 346–430.

https://doi.org/10.1017/thg.2013.65 Published online by Cambridge University Press