pp 443-451 of "Fifty years of human genetics: A Festschrift and liber amicorum to celebrate the life and work of George Robert Fraser", edited by Oliver Mayo and Carolyn Leach, (ISBN 9781862547537, Wakefield Press, Adelaide, 2007)

A beautiful method of analysis

Oliver Penrose

Department of Mathematics, Heriot-Watt University,
Riccarton, Edinburgh EH14 4AS, Scotland, UK

Introduction

It is a pleasure to dedicate this article to my friend George Fraser. I have a curious symmetrical relationship with George: neither of us knew the other until recently, but each of us knew the other's father. I knew George's father through playing chess with him in various tournaments and club matches. He was a strong player and won the British Chess Championship in 1957; the only way I ever found of beating him was to tempt him occasionally with an opportunity for an over-optimistic piece sacrifice. George and my own father, the human geneticist L S Penrose, knew each other very well, since my father was George's thesis supervisor (and, as a keen chess player himself, he also knew George's father). Nevertheless it was only about five years ago that George Fraser and I first met, when he spoke at a conference held at University College London in 1998 to commemorate the 100th anniversary of my father's birth.

My own introduction to human genetics came at a very early age. When I was about four years old I would sometimes venture into my father's study, to find him doing what I described at the time as "red and blue busywork". I like to think that the work he was doing with his red and blue pencils was connected with the data collection and analysis for his 1933 paper *The relative effects of paternal and maternal age in mongolism* (Penrose 1933) which he was working on around that time. This paper was a milestone in the use of quantitative methods in human genetics. Until that time, according to Harris (1973) nothing was understood about the causation of the condition now known as Down's syndrome. It was known that the affected children were more often born to elderly parents and often came late in

the order of birth, but the relative importance of father's age, mother's age, and birth rank was a mystery. (At that time the possibility that the explanation was a chromosomal abnormality, suggested in a remarkably prescient passage by Waardenburg (1932)[1] was not taken seriously (Penrose 1966)). My father set out to disentangle these effects, and the 1933 paper showed convincingly that the mother's age, but not the father's, was an important causative factor.

When I got a little older, the object of greatest interest in my father's study was a handle-powered desk calculating machine called the Brunsviga. This was the latest thing in computing technology at the time and it made a very satisfying crunching noise when you turned the handle to do a big multiplication or division sum. A lot of the number crunching he did around that time must surely have been the analysis of the data for the 1933 paper. The collection of these data was a *tour de force* of dedicated field work by my father and his assistants Miss M Newlyn and Dr M Gunther, working from the Royal Eastern Counties Institution in Colchester. They took family histories of 150 families each of which included at least one child with Down's syndrome, a total of 727 children, recording, among other things, the age of each parent at the time of the child's birth. But it was my father on his own who invented the method of analysis that induced this confusing jumble of data to give up

[1] On pages 47-8 of [3] Waardenburg says 'In view of the persistent uncertainty of its genetic basis, I may have given too much space to this anomaly. On the other hand, the unfailing recurrence of a whole series of symptoms in mongoloid patients affords a fascinating problem. I would like to persuade the cytologists to investigate the possibility that we may be dealing with a particular chromosome aberration in man. If would surely not be surprising if such conditions could occasionally occur in man and that, if the effect is not lethal, they would be the cause of a far-reaching constitutional anomaly. Investigations should be carried out to see whether mongolism is associated with "chromosomal deficiency" caused by "nondisjunction", or on the contrary, we might be dealing with "chromosomal duplication". It is of course also possible that the cause could be due to an anomaly of only parts of chromosomes (chromomeres): a "sectional deficiency" caused by a "translocation", or a "sectional duplication": this would be more difficult demonstrate cytologically. [These terms were introduced by Morgan, Bridges and Sturtevant, The Genetics of Drosophila, *Bibliographia Genetica* 1925, II. 1-262]. My hypothesis has the advantage that it is testable, and it might also be able to explain the influence of maternal age, whereas in men, because of the very large number of meiotic divisions, the chance of chromosome aberrations would be expected to be increased even without any special influences due to age. If my hypothesis were shown to be correct, it would result in an important insight into problems of human constitution and the manifestations of syndromes' [English translation by Ursula Mittwoch]

its secret.

Despite the interest generated by these colourful early experiences, it was not until 65 years later that I took the trouble to find out exactly what my father's "red and blue busywork" might have consisted of, when I looked up the 1933 paper as part of my preparation for a lecture (Penrose 1998) about my father and his family, given at Wisbech in 1998. What I found when I did look at the paper was a seemingly simple statistical argument, but the more I thought about it the more fascinated I became by the subtle understanding that underlay the apparent simplicity. The object of this note is to reveal some of these subtleties. I hope that the historical importance of that paper, and the interesting statistical method used in it, make it worth while to give it this reconsideration, even after a lapse of over 70 years.

The 1933 paper begins with a preliminary investigation of the relative importance of the father's and the mother's ages, based on the method of partial correlations. The partial correlation of the father's age with the occurrence of DS (Down's syndrome) in the child, after elimination of the effect of the mother's age, turned out to be very small, only -0.01, whereas the partial correlation of the mother's age with DS in the child, after elimination of the effect of father's age, was much larger, 0.22, with an estimated statistical error of order 0.04 in both numbers. Thus the partial correlation method provided strong *prima facie* evidence that the mother's age was an important factor and that the father's age was not. However there was a difficulty with this method, namely that the standard tests for deciding whether or not the observed correlation is statistically significant were not applicable. These tests depend on the assumption that the random variables in the problem obey a multivariate Gaussian probability distribution. But in the present case this assumption cannot be used. Although the two age variables (father's age and mother's age) are capable of a continuous range of values and therefore might without too much

violence to the facts be assumed to obey a Gaussian distribution, the third variable is nothing like Gaussian because it can take only two values, depending on whether the child does or does not have Down's syndrome.

My father came up with a beautiful method of analysis which enabled him to obtain reliable deductions from the data he and his assistants had collected. The method avoided the manifestly false assumption that all the random variables were Gaussian, while at the same time making it unnecessary to work out *ab initio* the corresponding theory of statistical tests for partial correlations when one of the variables can take only two values. Like all the best ideas, the method looks very simple --- once you have been shown how to do it.

The idea is to test two competing hypotheses against the data. One of them, which I shall call **M**, is the hypothesis that, of the ages of the two parents at the time of birth of the baby, only the mother's age is relevant to whether the baby will have Down's syndrome. This hypothesis is compared with a "control", namely a hypothesis **F** that only the father's age is relevant. Because of the symmetry between the two hypotheses, they are easily compared. The test used is a prediction of the age of the parent whose age does not matter, based on the age of the parent whose age does matter and on whether or not the child has Down's syndrome. The results of the test are shown in Table 1 (a simplified version of Table II of Penrose 1933). The top half of the table summarizes the test of hypothesis **M**. On average, the prediction of the age of the "irrelevant" parent (in this case the father) based on the age of the "relevant" parent is very good: although the prediction in any individual case would of course be very inaccurate, the average of the predicted ages of the "irrelevant" parents is in error by only a few weeks. For the competing hypothesis **F**, on the other hand, the predictions are much worse, the error in the predicted average being more than ten times as large. Thus the evidence strongly favours hypothesis **M** over hypothesis **F**,

and the conclusion drawn in the paper was that "paternal age is not a significant factor, while maternal age is to be regarded as very important."

Test of hypothesis **M,** that only the mother's age matters

average age of fathers	***observed***	***predicted***	***error***
of affected babies	39.38	39.47	0.09
of unaffected babies	33.83	33.8	-0.03

Test of hypothesis **F,** that only the father's age matters

average age of mothers	***observed***	***predicted***	***error***
of affected babies	37.25	35.71	-1.54
of unaffected babies	31.25	31.68	0.43

Table 1: Summary of the method and the results.

Two things about this analysis are particularly intriguing. One is that it depends entirely on the one set of data. There was no separate survey of the population as a whole. The necessary information about the general population was gleaned from the group of parents studied - even though that group is far from typical, consisting entirely of parents of Down's syndrome children. The other intriguing aspect is the relation between the two hypotheses **F** and **M**. Most statistical tests use just one hypothesis, but this one uses two. In the basic theory they are treated completely symmetrically. However, the data reveal an asymmetry between them, and so in the end the same data serve two distinct purposes at the same time: analysed according to one hypothesis, they provide information in support of that hypothesis, analysed according to the other, they provide a foil against which the performance of the first hypothesis can be evaluated. In the rest of this paper the method and the results will be examined more closely, to see better how it has all been achieved.

Readers who are allergic to mathematics may skip to the beginning of the section headed "The sample".

The probability model and the two competing hypotheses

The underlying probability model of the 1933 paper can be set out in the following way. A child is described in the model by just three variables: the mother's age *m* (measured in years) at the time of birth, the father's age *f* at the time of birth, and a non-numerical variable *c* describing the clinical situation. The variable *c* is capable of just two values: *A* if the baby is *A*ffected with Down's syndrome and *U* if the baby is *U*naffected. (Thus, $c = U$ means 'the child is unaffected'). We can call the triple *(f,m,c)* the 'state' of the child. Other characteristics of the child, such as its sex, its date of birth, and the number of brothers and sisters, are ignored.

We consider some relevant large population, which might be (but in fact is not) all the infants born in Great Britain during a particular year. For each possible state *(f,m,c)* we denote the number of children in the chosen population having that state by *N(f,m,c)*. The probability *w(f,m,c)* that an infant randomly chosen from that population would have had the state *(f,m,c)* is then equal to *N(f,m,c)/N,* where *N* is the total number of children in the population.

A couple expecting a baby will naturally be interested in knowing whether their child is likely to be affected with DS. The statistician cannot predict the future, but if he knows their ages *f, m* and the probability distribution function *w* he can tell them the fraction of couples of their age whose babies were affected with DS in the past. This fraction, which I will denote *p(A|fm),* is the conditional probability of the child's being in the state *A*, given the ages *f,m* of the two parents at the time of its birth. As a formula, it is given by setting $c = A$ in the formula

$$p(c|fm) = N(f,m,c)/N(f,m) \qquad c = \text{either } A \text{ or } U \qquad (1)$$

where

$$N(f,m) = \Sigma_{c'=A,U} N(f,m,c') \quad (2)$$

denotes the total number of children in the population whose fathers and mothers at the time of the child's birth were aged *f,m.*

The analysis that follows will use other conditional probabilities besides *p(c|fm);* for example, the conditional probability that the child's state is *c* and that in addition the father's age (at the time of birth) is *f*, given that the mother's age at that time is *m,* is defined by

$$p(cf|m) = p(fc|m) = N(f,m,c)/N(m) \quad (3)$$

where

$$N(m) = \Sigma_{c'}\Sigma_{f'}N(f',m,c') \quad (4)$$

is the total number of children in the population who were born to mothers aged *m* ; and the conditional probability that the father's age is *f ,* given that the mother's age is *m*, and independent of the state of the child, is

$$p(f|m) = N(f,m)/N(m) \quad (5)$$

The following identity is a direct consequence of the definitions (1), (3), (5):

$$p(cf|m) = p(c|fm)p(f|m) \quad \text{for all} \quad f,m,c \quad (6)$$

In principle, the probability that the baby born to a particular couple with ages *f,m* will turn out to have DS can depend on both parents' ages, *i.e.* the conditional probability *p(A|fm)* as defined in (1) may depend on both the variables *f,m*. However, each of the two hypotheses to be tested can be phrased as a statement that *p(c|fm)* depends on only one of these variables.

Hypothesis **M** : *only the mother's age affects the newborn baby's chance of having DS; the father's age is irrelevant (*i.e. *p(c|fm) is independent of f)*.

This statement of the hypothesis is equivalent[2] to the formula

$$p(c|fm) = p(c|m) \quad \text{for all} \quad f,m,c \tag{7}$$

Two alternative mathematical statements of this hypothesis are[3]

$$p(fc|m) = p(f|m)p(c|m) \tag{8}$$

$$p(f|cm) = p(f|m) \qquad \text{provided that} \qquad p(c|m)>0 \tag{9}$$

Hypothesis **F** : *only the father's age affects the newborn baby's chance of having DS; the mother's age is irrelevant.* As a formula, this is

$$p(c|fm) = p(c|f) \qquad \text{for all} \quad f,m,c \tag{10}$$

Two alternative formulations of hypothesis **F**, analogous to (8) and (9), can be obtained by interchanging the symbols *f* and *m* in (8) and (9).

A prediction

The statistical test summarized in Table 1 is based on the idea of using the hypothesis **M** or **F** to predict the average ages of the "irrelevant" parents at the birth of affected children and also of unaffected children. Thus, one of the two tests of hypothesis **M** is to use it to predict the average age of the fathers of affected children, given the ages of the mothers of those children. This average age, corresponding to the first entry in Table 1, can be written in terms of conditional probabilities as $\boldsymbol{E}(f|A)$, where

$$\boldsymbol{E}(f|c) = \Sigma_f \, f\, p(f|c) \qquad c = A \text{ or } U \tag{11}$$

denotes the conditional expectation of f in the sample, conditional on the given value of c. To use hypothesis **M** we express the conditional probability in the above formula as a sum over maternal ages:

$$p(f|c) = \Sigma_m \, p(fm|c) \qquad \text{by equations analogous to (5) and (3)}$$

---

[2] To prove (7), sum the identity (6) over f and use **M** which says that $p(c|fm)$ has a common value independent of f ; then use the sum rules $\Sigma_f \, p(cf|m) = p(c|m)$ and $\Sigma_f \, p(f|m) = 1$ to obtain $p(c|m) =$ (common value of $p(c|fm)$).

[3] To prove (8), put (7) into the identity (6) and use the symmetry exhibited on the left side of (3). To prove (9), interchange f and c in (6) to get $p(f|cm) = p(fc|m)/p(c|m)$; then use (8)in the right side of this last formula.

$$= \Sigma_m\, p(f|mc)p(m|c) \qquad \text{by an analogue of (6)}$$

$$= \Sigma_m\, p(f|m)p(m|c) \qquad \text{by hypothesis } \mathbf{M} \text{ (equation (9)} \qquad (12)$$

so that equation (11) becomes (after interchanging the two summations)

$$\boldsymbol{E}(f|c) = \Sigma_m\, \boldsymbol{E}(f|m)p(m|c) \qquad \text{under hypothesis } \mathbf{M} \qquad (13)$$

where we have defined $\boldsymbol{E}(f|m) := \Sigma_f\, f\, p(f|m)$, which is the expectation (average) of the father's age, given the age of the mother.

To evaluate the right side of (13) we make a standard simplifying assumption used in statistics, the assumption of *linear regression.* This assumption is that $\boldsymbol{E}(f|m)$ depends linearly on m, i.e. that a relation of the form

$$\boldsymbol{E}(f|m) = Jm + K \qquad \text{for all} \quad m \qquad (14)$$

holds, where J, K are constants which can be estimated by the least-squares method. Putting (14) into (13), we obtain a formula for the predicted value of $\boldsymbol{E}(f|c)$ under hypothesis **M,** in terms of $\Sigma_m\, mp(m|c)$ which is the same as $\boldsymbol{E}(m|c)$, the expectation of the mother's age conditional upon the state c of the baby. Since the baby may be either affected or unaffected, there are in fact two predictions, one for each of the two possible values of c, which can be tested against the actual data. Written out explicitly these predictions are

$$\boldsymbol{E}(f|A) = J\,\boldsymbol{E}(m|A) + K$$

$$\boldsymbol{E}(f|U) = J\,\boldsymbol{E}(m|U) + K \qquad \text{both under hypothesis } \mathbf{M} \qquad (15)$$

where $\boldsymbol{E}(m|c)$ means $\Sigma_m\, mp(m|c)$, the conditional expectation of the mother's age for given condition of the child, in analogy with equation (11).

The sample

The question that now arises is how to estimate the numbers on the left and right sides of equations such as (15). The 1933 paper treated this as a perfectly straightforward matter. The regression coefficients J, K were estimated by the least squares method from the data from the 150 families, and the expectations $\boldsymbol{E}(f|A)$, etc. were estimated

by taking appropriate averages of these data. But is it really quite so straightforward?

A standard method of estimating such expectations would be to take a random sample of children from the general population, and to treat the sample as being typical of the general population. In the 1933 paper, however, the sample was not taken at random from the general population, but consisted of a very particular class of children, namely those families of children with Down's syndrome to whom the researchers had access. Because of this method of selection, the sample is far from being typical of the general population; for example, since the parents of children with DS tend to be older than parents in the general population, we might expect the parents of the children in the sample to be older, on average, than parents in the general population. Even worse, the families in the sample all contain at least one child affected with DS, so one would expect the frequency of DS in the sample to be much higher than in the general population - and indeed, the sample contained 153 cases of DS among 727 children, a frequency of about 1 in 5, whereas the frequency of DS in the general population is about 1 in 600).

Because of this bias, we cannot be sure of getting reliable results from the usual assumption of sampling theory that the members of the sample were drawn at random from the general population. For example, we have no reason to believe that the average age of the fathers of the unaffected children in the sample is even approximately the same as $\boldsymbol{E}(f|U)$, the average age of the fathers of such children in the general population. A more reliable assumption would be that the sample was drawn at random from what might be called the special population, consisting of those families in the general population containing at least one child with DS. But analysing this assumption properly would be a complicated task, and moreover it would require statistical information not supplied in the 1933 paper, about things like the sizes and age structures of families. The following analysis is instead based on a

plausible simplifying assumption which eliminates any need for additional information of this kind.

To formulate this simplifying assumption, let us extend the model described in section 2 by including one further variable into the description of the "state" of a child. In addition to the mother's age, the father's age and the clinical state of the child, we include a fourth variable *s* capable of two non-numerical values which will be represented by symbols as follows: $s = *$ if the child is included in the sample and $s = \wedge$ if it is not[4]. In the standard random sampling procedure each child has the same probability of being included in the sample, regardless of its clinical state and the ages of its parents; that is to say, the conditional probability $p(*|fmc)$ is independent of *f,m* and *c*. By a mathematical argument very similar to the one leading from (5) to (7), this statement about independence implies

$$p(fmc|*) = p(fmc) \qquad \text{for standard random sampling} \qquad (16)$$

if we exclude the possibility that $p(*|fmc)$, the probability of going into the sample, is zero. Equation (16) says that the expected relative frequency of state (f,m,c) in the sample is the same as in the general population.

For the special sampling method used in (Penrose 1933), however, not all children in the general population have the same probability of going into the sample; that is to say, $p(*|fmc)$ depends on the values of some or all of *f, m* and *c*. For example $p(*|fmc)$ will be larger for $c = A$ than for $c = U$ since the intention of the researchers was to include as many affected children into their sample as possible. In general it will depend on both *f* and *m* as well; for it is quite likely that the other children in the family are born within a few years before or after the child being

[4] A curious feature of the variable *s* is that its value may not be known at the time of birth: unless the infant already has an affected older brother or sister the value of *s* is not known until either an affected sibling is born or it becomes clear that no more children from that family will be included in the sample. Moreover, its definition depends on when the sample is chosen: an unaffected first child born in 1932 with an affected sister born in 1934 would have $s=*$ for a survey done in 1933, but $s=\wedge$ for one done in 1935.

considered, and so one would expect that an unaffected child of older parents is more likely to have a sibling with DS, and therefore more likely to be in the sample, than a child of younger parents. Working out the actual dependence would require some fairly complicated theory and would also require additional factual information about things like the distribution of children's ages in families with more than one child. However there is a simple way of evading all this, if one of the following plausible extensions of the hypotheses **M** and **F** is accepted:

Hypothesis **M'** (to be used with hypothesis **M**)**:** *only the mother's age affects whether or not the baby has (or will have) a sibling with DS*, so that $p(*|cmf)$ depends only on c and m, but not on f.

Hypothesis **F'** (to be used with hypothesis **F**)**:** *only the father's age affects whether or not the baby has (or will have) a sibling with DS*, so that $p(*|cmf)$ depends only on c and f, but not on m.

By manipulations analogous to those used in deriving (7), (9) we can state the new hypothesis **M'** in either of the following ways:

$$p(*|cmf) = p(*|cm) \quad \text{for all } c,m,f \tag{17}$$

$$p(f|cm*) = p(f|cm) \quad \text{for all } c,m,f \tag{18}$$

provided, in equation (18), that $p(*|cm)>0$ Equation (18) tells us that, under this new hypothesis **M',** the conditional probability distribution of f at given c,m is the same in the sample as it is in the general population.

Hypothesis **M'** is certainly not a truism, nor is it a logical consequence of its close relative **M**. Moreover, it is not strictly biological, but contains a sociological component as well. If it were the case, for example, that old fathers in our society tended to have smaller families, then an unaffected child born to an old father would be less likely than one with a young father to enter the sample later on as a result of the subsequent birth of an affected child (the mother's age being the same in both

cases). In that case, unaffected children with old fathers would be under-represented in the sample. Nevertheless **M'** is a useful working hypothesis. It can be combined with **M** to give the following composite hypothesis:

Hypothesis **M***: *both* **M** *and* **M'** *are true; i.e. only the mother's age matters, both as to whether the child will be affected and as to whether any of its siblings are, or will be, affected.*

For a mathematical statement of hypothesis **M***, we combine **M** in the form (9) with **M'** in the form (18), to obtain

$$p(f|cm^*) = p(f|m) \qquad \text{under hypothesis } \mathbf{M^*} \qquad (19)$$

Hypothesis **M*** also implies

$$\begin{aligned} p(f|m^*) &= \Sigma_c\, p(fc|m^*) && \text{by analogues of (5) and (3)} \\ &= \Sigma_c\, p(f|cm^*)\, p(c|m^*) && \text{by an analogue of (6)} \\ &= \Sigma_c\, p(f|m)\, p(c|m^*) && \text{by (19)} \\ &= p(f|m) && \text{since } \Sigma_c\, p(c|m^*) = 1. \qquad (20) \end{aligned}$$

Equations (19) and (20) can be combined to show that hypothesis **M*** implies the following analogue of (9):

$$p(f|cm^*) = p(f|m^*) \qquad (21)$$

This is just like (9), but the asterisks show that it is a statement about the probabilities in the special rather than the general population.

By a calculation just like the one that led to (15), but with stars inserted everywhere to make all the probabilities refer to the special population, we can now derive the following prediction, which, unlike its analogue (15), involves only things that can be estimated from the sample:

$$\begin{aligned} \boldsymbol{E}(f|A^*) &= J^*\, \boldsymbol{E}(m|A^*) + K^* \\ \boldsymbol{E}(f|U^*) &= J^*\, \boldsymbol{E}(m|U^*) + K^* \qquad \text{both under hypothesis } \mathbf{M^*} \qquad (22) \end{aligned}$$

where J^*, K^* are the coefficients in the linear regression formula

$$\boldsymbol{E}(f|m^*) = J^*m + K^* \qquad \text{for all} \quad m \qquad (23)$$

The quantities in (22) are the ones that were estimated from the data in (Penrose 1933): the regression coefficients were estimated by the least squares method to be $J^* = 0.944$, $K^* = 4.304$, the conditional expectations were estimated to be the observed conditional averages given in column 2 of Table I, and column 3 shows the right side of (22). As noted already, the agreement was good: the data are consistent with hypothesis **M***.

The hypothesis to be compared with **M*** is obtained by combining **F** and **F'**:

Hypothesis **F***: *only the father's age matters*, *both as to whether the child will be affected and as to whether any of its siblings are, or will be, affected*.

For this hypothesis a calculation just like the one that gave (22), but with *m* and *f* interchanged throughout, leads to the following predictions:

$$\boldsymbol{E}(m|A^*) = J^{*\prime}\,\boldsymbol{E}(f|A^*) + K^{*\prime}$$

$$\boldsymbol{E}(m|U^*) = J^{*\prime}\,\boldsymbol{E}(f|U^*) + K^{*\prime} \qquad \text{both under hypothesis } \mathbf{F^*} \qquad (24)$$

where $J^{*\prime}, K^{*\prime}$ are the coefficients in the linear regression formula

$$\boldsymbol{E}(m|f^*) = J^{*\prime}f + K^{*\prime} \qquad \text{for all} \quad m \qquad (25)$$

The least-squares estimates of the coefficients $J^{*\prime}, K^{*\prime}$ given in (Penrose 1933) are $J^{*\prime} = 0.726$, $K^{*\prime} = 7.120$. The results in Table II show that (24) does not agree with the data, so that at either the biological component **F** or the sociological component **F'** of the composite hypothesis **F*** (or both) should be rejected.

Conclusion

In Penrose 1933 it was concluded from the data that paternal age is not a significant factor in determining whether the child would have DS, and that maternal age is very important. The statistical argument used was highly ingenious and original, but could be criticized on the grounds that it does not make any explicit allowance for the bias in the sample of children studied. The present paper suggests a way of allowing for

that bias, and the conclusion about the biological importance of maternal age is the same as before, although the strength of the conclusion in respect of paternal age is weaker because it may only be the sociological hypothesis **F'** that has been falsified by the data rather than the biological hypothesis **F**.

Moral: Originality, ingenuity, and tireless observation are more important than flawless statistics when it comes to doing good science.

Acknowledgements

I am grateful to George Fraser for providing references Harris 1973 and Waardenburg 1932 and for obtaining the translation of footnote 1, and to Ursula Mittwoch for doing the translation. I am also grateful to Mathew Penrose and Shirley Hodgson for informative discussions.

References

Harris H 1973 Lionel Sharples Penrose *Biographical memoirs of Fellows of the Royal Society* **19** 521-561.

Penrose LS 1966 Human chromosomes, normal and aberrant. *Proceedings of the Royal Society of London Series B* **164** 311-319.

Penrose LS 1933 The relative effects of paternal and maternal age in mongolism. *Journal of Genetics* **27** 219-224.

Penrose O 1998 *Lionel S Penrose, human geneticist and human being*, talk given to the Wisbech Society on 11 May 1998 (unpublished booklet).

Waardenburg PJ 1932 *Das Menschliche Auge and seine Erblangen*, pp.47-48 Martinus Nijhoff, the Hague.