Boston University - School of Medicine

THE " ACCURACY " OF ESTIMATES OF MNS GENE FREQUENCES

by
William C. Boyd

Wiener, has proposed calculating gene frequencies from MNS data obtained with three sera, anti-M, anti-N and anti-S, by a process which amounts to the use of the following formulas:

$$m_S = (1/2)\ (\sqrt{M + N + MN} + \sqrt{M} - \sqrt{N})$$

$$n_S = (1/2)\ (\sqrt{M + N + MN} - \sqrt{M} + \sqrt{N})$$

where m_S and n_S are the frequencies of the M and N genes lacking the factor$_S$, and MN are the frequencies of the MNS classes negative with anti-S serum. He has compared the results of his method with the results of the maximum likelihood method devised by me[2-3], and states that the maximum likelihood method does not increase the accuracy of the estimates, since the standard deviations indicate that the estimates "can at most be correct only to the second significant figure". It is seen from Wiener's Table 3 that the estimates by his method do in fact agree with the maximum likelihood estimates to two significant figures. But are the maxi mum likelihood estimates "correct only to tbe second significant figure"?

Significant Figures

The problem of how many decimal places to retain in a calculated statistic is not new, although Wiener is right in stating that some authors have more or less ignored it. The basic principle which applies is to give enough decimal places so that no substantial fraction of the information contained in the data will be wasted. It will be admitted that this is a most reasonable principle. The data which Wiener uses as an illustration of his method were obtained as a result of a journey just half way around the world and back, the work was financed by two Fulbright fellowships, involved a sabbatical leave from a university, and was only made possible by a high degree of kindness and cooperation on the part of many busy government officials in Pakistan. The results were therefore difficult and expensive to obtain, and the persone who did this work, at least, would not want any of their efforts to be wasted.

Exactly how much of the information in the data we shall allow ourselves to waste is a somewhat arbitrary matter, but Prof. Norton has suggested a rule which runs as follows:

Ricevuto il 6. II. 1956

https://doi.org/10.1017/S1120962300020618 Published online by Cambridge University Press

« On the general principle that data usually cost much compared to arithmetic, I usually follow a rule of wasting no more than one percent of the data. This means the variance of tho estimate should not exceed 101 per cent of the variance of the true likelihood estimate. If this is so, the variance of the difference between the two estimates is less than 1 per cent of the likelihood variance, and the standard deviation is less than 10 per cent of the last adjustment applied is as little as one-tenth of a standard error, it is reasonably certain that less than one per cent of the information is heing wasted ».

In order to apply this rule to the estimates discussed by Wiener it is necessary to calculate the standard deviations of the estimates of the gene frequences. The standard deviations given in the paper by Boyd[3] are unfortunately incorrect, being a little too small; their use would suggest even more stongly than the use of the correct values that the number of significant figures recommended by Wiener is not sufficient. The correct formulas for the standard deviations have since been published.[6]

In our symbols the gene frequences represented by Wiener as L^s, L, $l\sigma^s$, and $l\sigma$ are m_s, m_s, n_s and n_s. The standard deviations given by Boyd were $m_s = 0.018$, $m_s = 0.022$, $n_s = 0.014$, $n_s = 0.020$. (Wiener gives these and the gene frequences as percents, not frequences.) These values should have been 0.024 0.027, 0.021, and 0.025. If we follow the suggestion of Prof. Norton that the last adjustments to the estimated gene frequencies should not be more than 1/10 of a standard deviation, the last adjustment must be less, respectively, than 0.0024, 0.0027, 0.0021 and 0.0025. Now it would clearly not be sensible to apply adjustments of the order of two units or less in the third decimal place, only to round off to two decimals. Consequently three significant figures must be retained, and it is interesting to note that Wiener, in spite of his arguments, did retain just this number of places.

Another well known rule regarding the number of decimal places, the « one-third sigma rule »[7], would allow eleven times as much of the information to be wasted, but would still point to the retention of three decimal places.

" Accuracy " of Wiener's Estimate

Now let us consider how much of the information contained in the data is wasted by using Wiener's method of estimation.

Wiener states not only that the estimates under discussion can at most be correct only to the second significant figure, but that the additional effort required by the maximum likelihood method is not justified « as it does not increase the accuracy of the estimates ». Now estimates are accurate only if they summarize all the information contained in a body of data; otherwise they are inaccurate. It is known that some methods of estimating parameters are inefficient and vaste more of the information than do efficient methods, and Fisher's suggestion[8], that the efficiency of a method be obtained by calculating the ratio of the variance of the maximum like-

https://doi.org/10.1017/S1120962300020618 Published online by Cambridge University Press

lihood estimate (which is always efficient) to the variance of the inefficient estimate, has been generally eccepted. This enables us to determine the efficiency of Wiener's estimates of the MNS gene frequencies. The variances of the maximum likelihood estimates are obtained as described by Boyd[6], and the variance of Wiener's estimdte by a recently derived formula[9], which is

$$V(m_S) = (1/16G)\ [1 - 4m^2_S + 4m_S/s]$$

were G is the number of persons tested and s is the combined frequency of the S-negative genes, m_S, + n_S.

If this method of estimating the efficiency is applied to Wiener's estimates based on Boyd's Bengali data, the efficiency of Wiener's method turns out not to be bad. But it is a characteristic of inefficient methods that their efficiency may vary from one case to another, and for another typical set of data Wiener's estimates are not so satisfactory. Let us consider also the data on natives of the Cook Islands[10]: MS = 3, M = 92, MNS = 28, MN = 97, NS = 17, N = 30; total 267. Wiener's method gives for these data m_S = 0.5787, and the maximum likeliood method gives 0.5804. The difference in the estimated frequencies is not great, but there is a considerable difference in the efficiencies of the two estimates. The variance of the Wiener estimate is 0.000519, and that of the maximum likeliood method is 0.000470. From the ratio 470/519 we find that the efficiency of the Wiener estimate is 90.7 per cent. This means that relying on this estimate is equivalent to throwing away the results of the tests on 25 of the 267 persons tested. It may be suspected that Dr Fry, after spending a year in the Cook Islands collecting the bloods, and Mr. Simmons, after spending a considerable amount of time in his laboratory testing them, would be unwilling to do this.

Wiener has frequently stated that instead of doing maximum likelihood calculations he finds it simpler to test a few more persons. Let us see how much extra testing is involved. In the case of the Cook Islanders he would have to test 267/0.907 = 294 persons, or 27 more persons. This might not be easy, and under certain condition might require weeks of additional field work.

The problem can also be put into economic terms. It may be surmised that the cost of testing each individual in such a survey is not less than $3.00, and it is probably a good deal more. The maximum likelihood calculation can be run through, several times if need be, by anybody who can use a desk calculator, in an eight hour working day. At present rates of pay for technicians this would cost only about $12.00. Therefore the amount of information about m_S which cost Dr. Fry 3 × 267 + 12 = $813.00 would cost Dr. Wiener 3 × 294 = $882.00. Here as in other connections, the application of modern statistical methods can result in considerable savings.

https://doi.org/10.1017/S1120962300020618 Published online by Cambridge University Press

Summary

Wiener's methods for estimating gene frequencies from MNS data are compared with the maximum likelihood methods, and found not to be fully efficient. The extra effort and expense which would result from the use of such methods is discussed. It is concluded also that the number of significant figures advised by Wiener is not sufficient.

References

1. WIENER, A. S. (1954). Serology, Genetics and Nomenclature of the M-N-S Types. *Acta Genet. Med. Gemell. 3*, 314.
2. BOYD, W. C. (1953). Estimation of Gene Frequencies from MNS Data. *Science* 118, 756.
3. BOYD, W. C. (1954). Maximum Likelihood Method for Estimation of Gene Frequencies from MNS Data. *Am. J. Hum. Gen.* 6, 1.
4. NORTON, H. W. (1954). *Personal communication.*
5. BOYD, W. C. (1954). Shortened Maximum Likelihood Estimation of Rh Gene Frequencies Am. J *Hum. Gen.* 6, 303.
6. BOYD, W. C. (1955). (Letter to the editor.) Am. J. *Hum. Gen.* 1, 444.
7. KELLEY, T. L. (1947). *Fundamentals of Statistics*, Harvard University Press, Cambridge, Mass.
8. FISHER. R. A. (1950). *Statistical Methods for Research Workers.* Llth edition. Oliver and Boyd, Edinburg.
9. BOYD, W. C. (1956). Am. J. *Hum. Gen.* 8, 24.
10. FRY, E. I. (1955). An Analysis of Cook Island Blood Groups. Paper presented at the 1955 meeting of the Am. *Assoc. Phys. Anthrop.*

https://doi.org/10.1017/S1120962300020618 Published online by Cambridge University Press