114. *On Weinberg's Statistical Method in Human Heredity.*

By Junjiro OGAWA.

(Comm. by K. KUNUGI, M.J.A., Nov. 12, 1951.)

In the study of human heredity, where, of course, any artificial crossing experiments are impossible, when it is required to estimate the probability of a patient of a certain desease (= true rate of appearance of patients of the desease) among brothers, provided that the desease is known to be inherited according to a certain type of Mendelian law of inheritance, the difficulties occur, for instance, even brothers which were borne by parents having genes of that desease are not counted in the statistical data, unless they actually have at least one patient. To overcome such difficulty, W. Weinberg[1] had developed methods for estimating the true rate of appearance of patients among brothers early in 1913. Recently, Dr. M. Masuyama[2] criticized the Weinberg's method on the basis of the theory of modern mathematical statistics and derived the formula of estimate, which, however, turned out to be the same as that of Weinberg. Further, Masuyama derived the confidence limits for the parameter representing the probability from the asymptotical properties of the maximum likelihood estimate.

It seems to the author, however, that Masuyama's theory is not completely reasonable in the following two respects:

(1) First the probability of a patient of some desease among brothers is not constant, but varies in value according to genotype of the parent by which the brothers were borne. And further the existence probability or probability á priori of each genotype in the population of parents is unknown in almost all cases.

(2) Second, Masuyama's assumption that all patients in the data may be considered as randomly drawn from the population is implausible in practice. It will be rather more plausible to consider that one patient of brothers happened to come to the doctor, and then the doctor knows that there are more patients among his brothers. So we should not consider each patient but each brother as a sumpling unit.

The purpose of this note is to modify Weinberg-Masuyama's theory in two respects above mentioned to conform the situations which seems to be of more frequent occurrence in practice.

First, for the sake of simplicity of explanations, we shall assume that genotypes of parents by which the brothers were borne, of which the data on hand consists of, are such that the probability of appearance of a patient among brothers is constant value p. Since the probability that a n_α-brother[3] drawn at random has x patients is

$$\frac{n_\alpha!}{x!\,(n_\alpha-x)!}p^x(1-p)^{n_\alpha-x}, \quad x=0,\,1,\,\dots,\,n_\alpha,$$

the probability that a n_α-brother has at least one patient is

$$\sum_{x=1}^{n_\alpha}\frac{n_\alpha!}{x!\,(n_\alpha-x)!}p^x(1-p)^{n_\alpha-x}=1-(1-p)^{n_\alpha}\equiv\varphi_\alpha, \quad \text{say.}$$

Therefore the conditional probability that a n_α-brother has x patients under the condition that it is already known that it has at least one patient is

$$\varphi_\alpha^{-1}\cdot\frac{n_\alpha!}{x!\,(n_\alpha-x)!}p^x(1-p)^{n_\alpha-x}.$$

Consequently the probability that a patient drawn at random from the population of n_α-brothers having at least one patient will belong to such a class of n_α-brothers as having x patients becomes

$$\begin{aligned}F_{\alpha x}&=\varphi_\alpha^{-1}\cdot x\frac{n_\alpha!}{x!\,(n_\alpha-x)!}p^x(1-p)^{n_\alpha-x}\Big/\varphi_\alpha^{-1}\cdot\sum_{x=1}^{n_\alpha}x\frac{n_\alpha!}{x!\,(n_\alpha-x)!}p^x(1-p)^{n_\alpha-x}\\&=\frac{(n_\alpha-1)!}{(x-1)!\,(n_\alpha-x)!}p^{x-1}(1-p)^{n_\alpha-x}.\end{aligned} \tag{1}$$

Let the number of n_α-brothers having x patients be $y_{\alpha x}$, and the total number of n_α-brothers be N_α, then we have

$$\sum_{x=1}^{n_\alpha}y_{\alpha x}=N_\alpha, \quad \alpha=1,\,2,\,\dots.$$

The probability of getting $y_{\alpha x}$, $x=1,2,\dots,n_\alpha$, $\alpha=1,2,\dots$ simultaneously is

$$L=\prod_\alpha\frac{N_\alpha!}{y_{\alpha1}!\dots y_{\alpha n_\alpha}!}F_{\alpha1}^{y_{\alpha1}}\dots F_{\alpha n_\alpha}^{y_{\alpha n_\alpha}},$$

whence the maximum likelihood estimate[4] $\hat{p}$ of p should be obtained from the equation

$$\left.\frac{\partial M}{\partial p}\right|_{p=\hat{p}}=0,$$

where $M=\log L=\log p\cdot\sum_\alpha\sum_x(x-1)y_{\alpha x}+\log(1-p)\sum_\alpha\sum_x(n_\alpha-x)y_{\alpha x}$
$\quad +\text{terms independent of } p$

Differentiating M with respect to p, we have

$$\frac{\partial M}{\partial p}=\frac{1}{p}\sum_{\alpha}\sum_{x}(x-1)y_{\alpha x}-\frac{1}{1-p}\sum_{\alpha}\sum_{x}(n_{\alpha}-x)y_{\alpha x}\,, \tag{2}$$

thence, we get

$$\hat{p}=\sum_{\alpha}\sum_{x}(x-1)y_{\alpha x}/\sum_{\alpha}\sum_{x}(n_{\alpha}-1)y_{\alpha x}\,.$$

$$=\sum_{\alpha}\sum_{x}(x-1)y_{\alpha x}/\sum_{\alpha}N_{\alpha}(n_{\alpha}-1)\,. \tag{3}$$

It is easily seen that

$$E(\hat{p})=p,$$

and

$$D^2(\hat{p})=\frac{p(1-p)}{\sum_{\alpha}N_{\alpha}(n_{\alpha}-1)}\,. \tag{4}$$

When N_{α} becomes indefinitely large, then the limiting distribution of $\hat{p}$ approaches the normal distribution with the mean value p and variance given by (4) above. Further, the estimate $\hat{p}$ given by (3) is an efficient one. For, the amount of information concerning p is

$$E\left(\frac{\partial M}{\partial p}\right)^2=-E\left(\frac{\partial^2 M}{\partial p^2}\right)=\frac{\sum_{\alpha}N_{\alpha}(n_{\alpha}-1)}{p(1-p)}\,. \tag{5}$$

In consequence the confidence interval for p of confidence coefficient 100 $(1-\varepsilon)$ percent is asymptotically

$$P(\underline{p}\leqq p\leqq\bar{p})=1-\varepsilon\,,$$

where

$$\underline{p}=\frac{2K\hat{p}+t_{\varepsilon}^2-\sqrt{D}}{2(K+t_{\varepsilon}^2)}\,,\quad \bar{p}=\frac{2K\hat{p}+t_{\varepsilon}^2+\sqrt{D}}{2(K+t_{\varepsilon}^2)}\,, \tag{6}$$

and

$$D=(2K\hat{p}+t_{\varepsilon}^2)^2-4K\hat{p}^2(K+t_{\varepsilon}^2)\,,$$

$$K=\sum_{\alpha}N_{\alpha}(n_{\alpha}-1)\,,$$

$$1-\varepsilon=\frac{1}{\sqrt{2\pi}}\int_{-t_{\varepsilon}}^{t_{\varepsilon}}e^{-\frac{u^2}{2}}du\,,$$

i e., t_{ε} is the 100 ε percent point of the standard normal distribution.

If it is not so unreasonable to assume that the existence probabilities of various genotypes of parents having n_α children are independent of the number of children, we may put, for instance, the existence probabilities $\pi_1, \ldots, \pi_s$ of genotypes $G_1, G_2, \ldots, G_s$ are constant for all populations of parents having n_α children for $\alpha = 1, 2. \ldots .$

Let the probability that a patient will appear among brothers which was borne by parent of genotype G_α be p_α, then a parent drawn at random from the population of parents having n_α children will bear a patient is

$$P = \sum_{\alpha=1}^{s} \pi_\alpha p_\alpha . \qquad (7)$$

When genotypes of parents are unknown, the above method of estimation is only applicable for estimating the parameter P of (7), i.e., the probability that a parent drawn at rondom from the population will have a patient among its children.

The following table shows the frequencies of brothers having at least one harelip, classified by their numbers of members and patients, which were collected by Mr. H. Tsutsui of the Department of Oral Surgery, Dental School of Osaka University[5].

Frequencies of brothers having harelips.

n_α: Number of brothers
x: Number of patients

x \ n_α	1	2	3	4	5	6	7	8	9	10	$\sum_\alpha y_{\alpha x}$	$\sum_\alpha x y_{\alpha x}$
1	111	121	106	68	61	24	19	11	3	3	527	527
2		4	3	7	4	6	4	1	1	0	30	60
3			0	0	1	1	0	0	0	0	2	6
4				0	0	0	0	0	0	0	0	0
5					0	1	0	0	0	0	1	5
N_α	111	125	109	75	66	32	23	12	4	3	560	598

$$K = \sum_\alpha N_\alpha (n_\alpha - 1) = 1273$$

$$\sum_\alpha \sum_x (x-1) y_{\alpha x} = \sum_\alpha \sum_x x y_{\alpha x} - \sum_\alpha \sum_x y_{\alpha x} = 38$$

$$\hat{p} = \frac{38}{1273} = 0.0298.$$

By the method explained above, it will be seen that the confidence interval for P of confidence coefficient 90% is

$$0.018984 \leqq P \leqq 0.038690$$

On the other hand, Tsutsui has found by examining 13061 briths at the infirmary hospital of Osaka University that the appearance rate of harelips in all is $25/1306 = 0.002$[6], so the appearance rate estimated from brothers having at least one patient is significantly higher than the general appearance rate. From this fact we have concluded that the hereditary incline of harelip is not denied. But about the type inheritance nothing can be specified.

Notes and References.

1) Weinberg, W., Auslesewirkungen bei biologisch-statistischen Problemen, Arkiv für Rassen- und Gesellschaftsbiologie (1913) pp. 417–451.

2) Masuyama, Motosaburo, On a statistical method of heredity, Medecine and Biology, Vol. 5, No. 7 (1944), pp. 383–3 4 (in Japanese).

3) n_a-brother means here n_a children who were born by one pair of parents.

4) Dugué, D., Applications des propriétés de la limité au sens de calcul des probabilités à l'étude de diverses questions d'estimation, Journ. de l'École Polytechnique, (1937).

Doob, J., Probability and statistics, Trans. of Amer. Math. Soc., Vol. 36 (1934).

5) Tsutsui, Hideo, Study on the Etiology of Clefts of the Lips and the Palate. (in Japanese, unpublished.)

6) Tsutsui, H., Study on the Etiology of Clefts of the Lips and the Palate, (I) Clinico-Statistical Observation (in Japanese). Shigaku-Zasshi, Vol. 8, No. 1 (1951).

The author expresses here his deep thanks to Mr. H. Tsutsui for his kindness of letting his data at the author's disposal.