Kawasaki Med. J. Vol. 1: 85-93, 1975

INCOMPLETE QUANTAL RESPONSE DATA ANALYSIS

—— Maximum Likelihood Estimation of Parameters Based on Mixtures of Interval Data and Ordinary Ones ——

by

TAICHI KARIYA

Division of Statistics, Kawasaki Medical School
Kurashiki 701-01, Japan

Accepted for Publication on Dec. 23, 1974

Abstract

In this paper, we attempt to classify the situations generating quantal response data into three main groups and develop a method of maximum likelihood estimation based on incomplete quantal response data. Incomplete quantal response data often arise in medical and biological examinations. In some cases the tolerance of each test subject can be known only to be above or below the value which is given or decided on the day of group examination. In such situations, the standard techniques of maximum likelihood estimation of parameters cannot be applied, because the values of observations are not specified. The following method of maximum likelihood estimation of parameters based on incomplete quantal response data will have a wide range of application in statistical estimation problems.

§ 1. The Classification of Quantal Response Data

Many problems of quantitative inference in medical and biological researches are concerned with the relation between a stimulus and a response. One exceedingly important type of response is known as all-or-nothing or *quantal*. For example, our interest will lie in the dependence of magnitude of the response on the dose of a drug. However, certain responses permit of no graduation and can be expressed only as 'occurring' or 'not-occurring'. If the characteristic response is quantal, occurrence or non-occurrence will depend upon the intensity of the stimulus (for example, a vitamin, a drug, a mental test, or an age). For any one subject, under controlled conditions, there will be a certain level of intensity below which the response does not occur and above which the response occurs; such a value has often been called *a tolerance value*. This tolerance value will vary from one subject to another in the population used.

The discussion of quantal response data therefore requires the recognition of frequency distribution of tolerances over the population studied.

The major situations of quantal response data can be identified as one of the following three models.

A. Dose-summing up model (ordinary data)
B. Dose-assigning model (cumulative frequency data)
C. Group examination model (incomplete data or interval data)

The term *dose* may be replaced by any other one which represents the intensity of the stimulus.

Model **A.** Sometimes the tolerance of each test subject in respect of a stimulus can be measured directly. For example, in the 'cat' method for the assay of digitalis, anaesthetized cats are given a continuous slow intravenous infusion of digitalis until death occurs. In such a case, we can obtain the measurements of tolerance like measurements of length or weight. But there are some weaknesses. If there is an appreciable time lag between the injection of the drug and its taking effect, the tolerance will be overestimated. Moreover, the dose required to cause death under conditions of slow infusion need not be the same as the tolerance for more rapid infusion.

An alternative direct measurement of tolerance can be practicable, if there is no cumulative effect of doses already given, either as lowering or as increasing the resistance of the subject. We give to each subject successive doses of different intensity, allowing after every dose a suitable time interval for a return to normal and making the differences sufficiently small for a satisfactory determination of the lowest dose which causes the characteristic response. With the direct measurement of tolerance, the appropriate methods of statistical analysis are the same as with other types of biological measurements. These methods of statistical analysis are detailed in many text-books.

Direct tolerance measurement is, however, often impracticable on account of the time the methods require. Even more commonly, it is ruled out entirely by the nature of the problem: a direct measurement technique for the poison tolerance of an insect, or of a fungus spore, is scarcely conceivable. Especially in the medical research whose subject is man, a direct tolerance measurement is almost impracticable.

Model **B.** An entirely different approach must be adopted if a direct tolerance measurement is impracticable on account of cumulative effect of doses already given, the time the methods require and the nature of the problem. The potency of the stimulus must be assessed from the proportions of subjects that respond, in random samples of the population tested at different doses.

If a batch of n test subjects (a random sample of size n from the population) is exposed to the same dose, and all react independently, the probabilities of n, (n−1), . . . , 1, 0 responding are the (n+1) terms in the expansion of the binomial $(P+Q)^n$. The probability P that the subject dies is the proportion of deaths that would occur if the whole population received the same dose, and Q=1−P. The probability of exactly r responding is therefore

$$P_r(r|n)=\frac{n!}{r!(n-r)!}P^rQ^{n-r}.$$

As the outcome of an experiment on different doses, applied under standardized conditions to random samples, we obtain the following data.

Dose level	Sample size	No. affected	%
x_1	n_1	r_1	p_1
x_2	n_2	r_2	p_2
⋮	⋮	⋮	⋮
x_s	n_s	r_s	p_s

Among the statistical analysis based on these data, probit analysis is the most popular and is developed in detail. [1] In this model the estimate $\hat{p}_i$ is given not by r_i/n but by r_i/n_i and so it is desirable that n_i should be sufficiently large from the point of precision, but this is often impracticable on account of the time, cost and so on. Furthermore, optimal allocation of sample n $(n=n_1+n_2+\cdots+n_s)$ to dose levels and optimal spacing between dose levels are very difficult because the tolerance distribution is unknown. This model is often inapplicable in the medical research whose subject is man.

Model **C.** In the medical research we are often obliged to deal with incomplete quantal response data obtained by survey or group examination. Then the quantal response techniques as probit analysis have been used in the study of such phenomenon in some population as cannot be exactly dated but can readily be recorded as having occurred or not occurred in any one individual. It is organizationally and possibly psychologically difficult to obtain reliable records of the age of menarche in adolescent girls. On the other hand, a sample of girls distributed over the appropriate age range can relatively easily be classified according to whether or not each has yet menstruated. The standard dose-response techniques are not applicable, even though age is regarded as analogous to dose, and 'having passed menarche' as response. With the standard dose-response techniques, ages on the day of survey must be grouped according to their magnitude, into some classes. We select a suitable representative age of each class and the age is regarded as analogous to dose. Then these representative ages of classes are not exact values and so the method of regression analysis should not be applied, for the use of regression methods may give unreliable results if there is any appreciable error in the measurement of independent variate.

We should develop the estimation method of menarche age distribution directly from incomplete quantal response data where the incomplete data are sets of the ages of menarche each of which is known only to be above or below her age on the day of survey. There are many problems of the same type, for example, the estimation of the age distribution of the first milk-teething, the estimation of the distribution of tolerance concentration of some toxin accumulated in the liver and so on.

§ 2. Maximum Likelihood Estimation Based on Incomplete Quantal Response Data

§ 2.1 Introduction

Here we are going to be concerned with incomplete quantal response data in Model C. In Model C the tolerance value x^* of each subject is known only to be above or below the value x which is given or decided on the day of survey. That is to say, we can get either $x^* > x$ or $x^* < x$. (If $x^* = x$, x^* is an ordinary datum.) If the type of distribution of tolerance x^* is given, the relation $x^* > x$ or $x^* < x$ may be replaced by interval data. For example, with the normal distribution $x^* > x$ may be replaced by (x, ∞) and $x^* < x$ by $(-\infty, x)$, and with the log-normal distribution $x^* > x$ may be replaced by (x, ∞) and $x^* < x$ by $(0, x)$. From this point of view interval data may be designated general censored data. In the following discussion we will use the interval data (general censored data) (y, z) instead of x^*, where $y < x^* < z$ and y, z are determined respectively as the examples mentioned above.

At the obvious risk of excluding certain situations from our discussion we shall confine our concept of 'data analysis' to the problem of 'estimation of parameters from sample data'. Generally our sample data will be composed of mixtures of interval data and ordinary ones.

Therefore the following method of maximum likelihood estimation of parameters can be applicable to various statistical estimation problems.

§ 2.2 Formulation and Maximum Likelihood Solution

Suppose $(x_1, x_2, \cdots, x_n, x_1^*, x_2^*, \cdots, x_m^*)$ is a random sample from a population having continuous distribution function $F(x; \theta)$;

(i) The x_is are ordinary measurement values of sample units.

(ii) The x_i^*s are unspecified values of sample units which cannot be measured but where the relation $y_j < x_j^* < z_j$ can be known to hold, and y_j, z_j are independent of x_j^* and are the values which are given or decided on the day of survey or group examination.

(iii) Moreover, we assume that x_j^* distributes according to the probability density function $f(x; \theta)/(F(z_j; \theta)-F(y_j; \theta))$ $(y_j < x < z_j)$, where $f(x; \theta)=\frac{dF(x; \theta)}{dx}$. It is clear that the assumption for x_j^* is always satisfied in Model C. The mathematical form of the distribution function $F(x; \theta)$ is known, but the value of parameter θ is unknown.

Here we are obliged to use the interval data (y_j, z_j) instead of x_j^*. Then we define the likelihood function of the random sample $(x_1, \cdots, x_n, x_1^*, \cdots, x_m^*)$ by

$$L(\theta) \propto \prod_{i=1}^{n} f(x_i)dx_i \cdot \prod_{j=1}^{m} (F(z_j)-F(y_j)), \qquad (1)$$

where $F(x)$ is the distribution function and $f(x)$ is the probability density function of X. [2] And then we obtain

$$L^*(\theta)=\log L(\theta)=\log c+\sum_{i=1}^{n}\log f(x_i)+\sum_{j=1}^{m}\log (F(z_j)-F(y_j)), \qquad (2)$$

where c is constant.

The log-likelihood function $L^*(\theta)$ and $L(\theta)$ have their maximum at the same value of θ, and it is sometimes easier to find the maximum of $L^*(\theta)$. If certain regularity conditions are satisfied, the point where the $L^*(\theta)$ is a maximum is a solution of the equation

$$\frac{\partial L^*}{\partial \theta}=0 \ . \qquad (3)$$

We assume that the tolerance distribution is normal or log-normal, and then the tolerance distribution involves two parameters, say μ and σ^2. For convenience of notation, let the variance σ^2 be denoted by ϕ.

Then the maximum likelihood equations for estimating the parameters μ and ϕ are given by

$$\frac{\partial L^*}{\partial \mu}=0, \qquad \frac{\partial L^*}{\partial \phi}=0. \tag{4}$$

Solving these equations we obtain the maximum likelihood estimates μ, ϕ which do maximize L^*. The explicit solution of such equations is impossible, but iterative methods can give successive approximations converging to the solutions. In most cases the following iterative procedure is generally used. Suppose that μ_0, ϕ_0 are any approximations to the solutions of the equations (4). By the Taylor-Maclaurin expansion, to the first order of small quantities, second approximations will be $\mu_0+\delta\mu$, $\phi_0+\delta\phi$, where $\delta\mu$, $\delta\phi$ are obtained from

$$\left.\begin{aligned} \frac{\partial L^*}{\partial \mu_0}+\delta\theta\frac{\partial^2 L^*}{\partial {\mu_0}^2}+\delta\phi\frac{\partial^2 L^*}{\partial \mu_0\partial\phi_0}=0, \\ \frac{\partial L^*}{\partial \phi_0}+\delta\theta\frac{\partial^2 L^*}{\partial \mu_0\partial\phi_0}+\delta\phi\frac{\partial^2 L^*}{\partial {\phi_0}^2}=0\,; \end{aligned}\right\} \tag{5}$$

the addition of the suffix to μ, ϕ indicates that the first approximations are to be substitued after differentiation. In general these linear equations are easily solved for $\delta\mu$, $\delta\phi$. The process may now be repeated with

$$\left.\begin{aligned} \mu_1=\mu_0+\delta\mu, \\ \phi_1=\phi_0+\delta\phi, \end{aligned}\right\} \tag{6}$$

in place of μ_0, ϕ_0, and further cycles computed until the latest set of adjustments is negligible. The maximum likelihood estimates $(\hat{\mu}, \hat{\phi})$ of the parameters (μ, ϕ) are the values satisfying simultaneously the two estimating equations (4).

The asymptotic variance-covariance matrix of (μ, ϕ) can be approximated by

$$\begin{pmatrix} \hat{V}(\hat{\mu}) & \hat{\mathrm{C}}\mathrm{ov}\,(\hat{\mu}, \hat{\phi}) \\ \hat{\mathrm{C}}\mathrm{ov}\,(\hat{\phi}, \hat{\mu}) & \hat{V}(\hat{\phi}) \end{pmatrix} = -\begin{pmatrix} \partial^2 L^*/\partial\mu^2 & \partial^2 L^*/\partial\mu\partial\phi \\ \partial^2 L^*/\partial\phi\partial\mu & \partial^2 L^*/\partial\phi^2 \end{pmatrix}^{-1}_{(\mu,\,\phi)=(\hat{\mu},\,\hat{\phi})}. \tag{7}$$

Inconveniently the solution of the equations (5) is impossible for the $L^*(\theta)$ under certain conditions and so we propose the following relaxation method which is a modified method of Gauss-Seidel's.

§ 2.3 Development of Maximum Likelihood Equations

The most important form is that of the estimation of the parameters of the tolerance distribution given by equation

$$F(x_0;\mu,\sigma^2)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x_0}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx,$$

where x measures dose on a logarithmic or other suitable matametric scale (x_0 being a particular value of x). Assume that $F(x,\theta)$ is a normal distribution function having unknown parameter θ, and then for the data described above the log-likelihood function $L^*(\mu,\phi)$ is given by

$$L^*(\mu,\phi)=\log c+\sum_{i=1}^{n}\log f(x_i)+\sum_{j=1}^{m}\log(F(z_j)-F(y_j)), \tag{8}$$

where c is constant,

$$f(x)=\frac{1}{\sqrt{2\pi\phi}}\exp\left(-\frac{(x-\mu)^2}{2\phi}\right) \text{ and } F(t)=\int_{-\infty}^{t}f(x)dx$$

On differentiating $L^*(\mu,\phi)$ with respect to μ and ϕ and putting these derivates equal to 0, we obtain the estimating equations for the normal parameters μ, $\phi(=\sigma^2)$

$$0=\sigma\frac{\partial L^*}{\partial\mu}=\sum_{i=1}^{n}x'_i+\sum_{j=1}^{m}\frac{-\varphi(z'_j)+\varphi(y'_j)}{\Phi(z'_j)-\Phi(y'_j)}, \tag{9}$$

$$0=2\phi\frac{\partial L^*}{\partial\phi}=-n+\sum_{i=1}^{n}x'^2_i+\sum_{j=1}^{m}\frac{-z'_j\varphi(z'_j)+y'_j\varphi(y'_j)}{\Phi(z'_j)-\Phi(y'_j)}, \tag{10}$$

where $x'_i=(x_i-\mu)/\sigma,\quad y'_j=(y_j-\mu)/\sigma,\quad z'_j=(z_j-\mu)/\sigma,$

$$\varphi(y'_j)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y'^2_j}{2}\right),\quad \varphi(z'_j)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z'^2_j}{2}\right),$$

$$\Phi(y'_j)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{y'_j}\exp\left(-\frac{t^2}{2}\right)dt,\quad \Phi(z'_j)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z'_j}\exp\left(-\frac{t^2}{2}\right)dt.$$

Substituting $\mu=\mu_0+\delta\mu$, $\phi=\phi_0$ $(\sigma=\sigma_0)$ for (9) gives

$$\sum_{i=1}^{n}\left(x'_{i_0}-\frac{\delta\mu}{\sigma_0}\right)+\sum_{j=1}^{m}\frac{-\varphi(z'_{j_0})+\varphi(y'_{j_0})}{\Phi(z'_{j_0})-\Phi(y'_{j_0})}$$

$$+\frac{\delta\mu}{\sigma_0}\sum_{j=1}^{m}\left\{\frac{-z'_{j_0}\varphi(z'_{j_0})+y'_{j_0}\varphi(y'_{j_0})}{\Phi(z'_{j_0})-\Phi(y'_{j_0})}-\left(\frac{-\varphi(z'_{j_0})+\varphi(y'_{j_0})}{\Phi(z'_{j_0})-\Phi(y'_{j_0})}\right)^2\right\}\doteqdot 0 \quad .$$

Then we get

$$\delta\mu \doteqdot \sigma_0\Big(\sum_{i=1}^{n} x'_{i_0}+\sum_{j=1}^{m} A_{j_0}\Big) \Big/ \Big\{n+\sum_{j=1}^{m}(A^2_{j_0}-B_{j_0})\Big\} \tag{11}$$

and

$$\mu_1=\mu_0+\delta\mu \doteqdot \mu_0+\sigma_0\Big(\sum_{i=1}^{n} x'_{i_0}+\sum_{j=1}^{m} A_{j_0}\Big) \Big/ \Big\{n+\sum_{j=1}^{m}(A^2_{j_0}-B_{j_0})\Big\}, \tag{12}$$

where $x'_{i_0}=(x_i-\mu_0)/\sigma_0,\quad y'_{j_0}=(y_i-\mu_0)/\sigma_0,\quad z'_{j_0}=(z_j-\mu_0)/\sigma_0,$

$$A_{j_0}=\frac{-\varphi(z'_{j_0})+\varphi(y'_{j_0})}{\Phi(z'_{j_0})-\Phi(y'_{j_0})}, \text{ and } B_{j_0}=\frac{-z'_{j_0}\varphi(z'_{j_0})+y'_{j_0}\varphi(y'_{j_0})}{\Phi(z'_{j_0})-\Phi(y'_{j_0})}.$$

Now on substituting $\mu=\mu_1$, $\phi=\phi_0+\delta\phi$ ($\phi_0=\sigma_0^2$) for (10), we obtain

$$-n+\sum_{i=1}^{n} x'^2_{i1}\Big(1-\frac{\delta\phi}{\phi_0}\Big)+\sum_{j=1}^{m}\frac{-z'_{j_1}\varphi(z'_{j_1})+y'_{j_1}\varphi(y'_{j_1})}{\Phi(z'_{j_1})-\Phi(y'_{j_1})}$$

$$+\frac{1}{2}\frac{\delta\phi}{\phi_0}\sum_{j=1}^{m}\Big\{\frac{-(z'_{j_1}+z'^2_{j_1})\varphi(z'_{j_1})+(y'_{j_1}+y'^2_{j_1})\varphi(y'_{j1})}{\Phi(z'_{j_1})-\Phi(y'_{j_1})}$$

$$-\Big(\frac{-z'_{j_1}\varphi(z'_{j_1})+y'_{j_1}\varphi(y'_{j1})}{\Phi(z'_{j_1})-\Phi(y'_{j_1})}\Big)^2\Big\} \doteqdot 0 .$$

Then we get

$$\delta\phi \doteqdot \phi_0\Big(-n+\sum_{i=1}^{n} x'^2_{i_1}+\sum_{j=1}^{m} B_{j_1}\Big) \Big/ \Big\{\sum_{i=1}^{n} x'^2_{i_1}+\frac{1}{2}\sum_{j=1}^{m}(B_{j_1}^2-B_{j_1}-C_{j_1})\Big\} \tag{13}$$

and

$$\phi_1=\phi_0+\delta\phi \doteqdot \phi_0+\phi_0\Big(-n+\sum_{i=1}^{n} x'^2_{i_1}+\sum_{j=1}^{m} B_{j_1}\Big) \Big/ \Big\{\sum_{i=1}^{n} x'^2_{i_1}+\frac{1}{2}\sum_{j=1}^{m}(B_{j_1}^2-B_{j_1}-C_{j_1})\Big\} \tag{14}$$

where $x'_{i_1}=(x_i-\mu_1)/\sigma_0,\quad y'_{j_1}=(y_j-\mu_1)/\sigma_0,\quad z'_{j_1}=(z_j-\mu_1)/\sigma_0,$

$$B_{j_1}=\frac{-z'_{j_1}\varphi(z'_{j_1})+y'_{j_1}\varphi(y'_{j_1})}{\Phi(z'_{j_1})-\Phi(y'_{j_1})} \text{ and } C_{j_1}=\frac{-z'^2_{j_1}\varphi(z'_{j_1})+y'^2_{j_1}\varphi(y'_{j_1})}{\Phi(z'_{j_1})-\Phi(y'_{j_1})}.$$

§ 2.4 Iterative Maximum Likelihood Estimating Procedure

The iterative maximum likelihood estimating procedure is, then, as follows.

(1°) Set initial values μ_0 and ϕ_0 ($=\sigma_0^2$).

(2°) From μ_0 and ϕ_0 ($=\sigma_0^2$), compute the standardized variables

$$x'_{i_0}=(x_i-\mu_0)/\sigma_0,\quad y'_{j_0}=(y_j-\mu_0)/\sigma_0,\quad z'_{j_0}=(z_j-\mu_0)/\sigma_0$$

and then calculate

$$A_{j_0}=\frac{-\varphi(z'_{j_0})+\varphi(y'_{j_0})}{\Phi(z'_{j_0})-\Phi(y'_{j_0})} \text{ and } B_{j0}=\frac{-z'_{j_0}\varphi(z'_{j_0})+y'_{j_0}\varphi(y'_{j_0})}{\Phi(z'_{j_0})-\Phi(y'_{j_0})}.$$

(3°) With these calculated values, estimate $\delta\mu$ and μ_1 by the following equations.

$$\delta\mu=\sigma_0\left(\sum_{i=1}^{n}x'_{i_0}+\sum_{j=1}^{m}A_{j_0}\right)\Big/\left\{n+\sum_{j=1}^{m}(A^2_{j0}-B_{j_0})\right\},$$

$$\mu_1=\mu_0+\delta\mu.$$

(4°) Now from μ_1 and ϕ_0 $(=\sigma_0^2)$, compute the standardized variables

$$x'_{i_1}=(x_i-\mu_1)/\sigma_0,\quad y'_{j_1}=(y_j-\mu_1)/\sigma_0,\quad z'_{j_1}=(z_j-\mu_1)/\sigma_0$$

and then calculate

$$B_{j_1}=\frac{-z'_{j_1}\varphi(z'_{j_1})+y'_{j_1}\varphi(y'_{j_1})}{\Phi(z'_{j_1})-\Phi(y'_{j_1})}\quad\text{and}\quad C_{j_1}=\frac{-z'^{2}_{j_1}\varphi(z'_{j_1})+y'^{2}_{j_1}\varphi(y'_{j_1})}{\Phi(z'_{j_1})-\Phi(y'_{j_1})}.$$

(5°) With these calculated values, estimate $\delta\phi$ and ϕ_1 by the following equations.

$$\delta\phi=\phi_0\left(-n+\sum_{i=1}^{n}x'^{2}_{i_1}+\sum_{j=1}^{m}B_{j_1}\right)\Big/\left\{\sum_{i=1}^{n}x'^{2}_{i_1}+\frac{1}{2}\sum_{j=1}^{m}(B^2_{j1}-B_{j_1}-C_{j_1})\right\},$$

$$\phi_1=\phi_0+\delta\phi.$$

(6°) Setting μ_1, ϕ_1 as initial values in place of μ_0, ϕ_0 respectively, continue steps (1°)~(5°) until the latest set of adjustments, say $\delta\mu$, $\delta\phi$ is negligible.

We will denote the latest set of estimates by $\hat{\mu}$, $\hat{\phi}$. These estimates $\hat{\mu}$ and $\hat{\phi}$ are the solution of the maximum likelihood equations (9), (10), and the joint maximum likelihood estimates.

The asymptotic variance-covariance matrix of (μ, ϕ) can be approximated by

$$\begin{pmatrix}\hat{V}(\hat{\mu}) & \hat{\mathrm{Cov}}(\hat{\mu},\hat{\phi})\\ \hat{\mathrm{Cov}}(\hat{\phi},\hat{\mu}) & \hat{V}(\hat{\phi})\end{pmatrix}=-\begin{pmatrix}\dfrac{\partial^2L^*}{\partial\mu^2} & \dfrac{\partial^2L^*}{\partial\mu\partial\phi}\\ \dfrac{\partial^2L^*}{\partial\phi\partial\mu} & \dfrac{\partial^2L^2}{\partial\phi^2}\end{pmatrix}^{-1}_{(\mu,\phi)=(\hat{\mu},\hat{\phi})}. \tag{15}$$

[Remark] For the computation of $\Phi(y')$ it is convenient to use the Hastings' approximation formulus. [3]

REFERENCES

[1] Finney, D. J.: Probit Analysis, (Third edition), Cambridge Univ. Press, 1971.
[2] Kendall. and A. Stuart: The Advanced Theory of Statistics, Vol. 2. (Third edition), Griffin, London, 1973.
[3] Hastings, C.: Approximations for Digital Computers, Princeton Univ. Press, 1955.
[4] Hartley, H. O. and R. R. Hocking: The Analysis of Incomplete Data, Biometrics 27, 783-823, 1971.