A Model in Probit Analysis

Kudo, Akio
Kyushu University

Furukawa, Nagata
Kyushu University

https://doi.org/10.5109/12982

出版情報：統計数理研究. 8 (1/2), pp.1-7, 1958-03. Research Association of Statistical Sciences
バージョン：
権利関係：

A MODEL IN PROBIT ANALYSIS

By

Akio KUDÔ and Nagata FURUKAWA

(*Received* Feb. 15, 1958)

§ 1. **Introduction.** The statistical inference theory based on the assumption that the binomial probability P is a known function of the level, say x, with unknown parameters, say θ_1, θ_2, is called the probit analysis and has special importance in bioassays depending upon quantal responses. Usually the function is expressed as the probability integral of the form

$$P=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\alpha+\beta x-5} e^{-\frac{1}{2}u^2}du\,. \tag{1.1}$$

Other types of the function are adopted very rarely.

The practical meaning of the probit analysis in bioassay is as follows. Suppose some billogical subject, such as an insect, a plant etc, is applied a stimulus, such as a drug, X-ray etc, at a specified intensity, measured in a way. Then the subject makes response such as death, excitment etc, as its results. If the response is not quantative and is composed of only two, "response" or "non-response," then the minimum intensity of the stimulus to result in "response" is called the tolerance of the subject under application of the stimulus. Now the tolerance should vary from one individual to another, and if a subject is sampled from a population, the tolerance should be a random variable. If a subject is sampled from a population, and a stimulus is applied, then it should give "response," in case the intensity of stimulus is higher than the tolerance of this particular individual, and it should give "non-response" in the reverse case.

As a conclusion we can say that if a sample of some size is drawn from a population, and the stimulus at a specified intensity is given to each of the individuals then the number of individuals giving "response" should be distributed in binomial distribution whose probability is a function of the intensity.

The inference theory of this type has been developed by Finney [1].

The problem, which we shall discuss in this paper and does not appear to have been very fully discussed, is what happens when the number of subjects under application is not known.

In section 2, we shall discuss such a problem in case the average number of individuals is very large and an existing method due to K. Mather will be

valid under certain conditions, and in section 3 we shall discuss the problem when the number of observation subjects to Poisson distribution.

§2. **The case when the number of individuals is very large and the method of analysis due to Mather.** At first we shall state the following theorem. (c.f. [4])

THEOREM 1. Let $F(x\,;\,\alpha,\,\beta)$ be the distribution of the tolerance, and let $x_1, x_2, \cdots, x_N$ be the value of the tolerance of each of the individuals drawn from a population, then the probability that the tolerances of all the individuals are less than x_0 will tends to a function of the type $e^{-e^{-(Ax+B)}}$ as N tends to ∞, where A is positive, and both A and B are the constants depending on the parameter α and β and also the distribution function F itself.

PROOF. Proof is immediate as this probability is nothing but the probability of the maximum of independent random variables and we can make use of the theorem of the limit of the maximum of independent random variables due to R. A. Fisher [3]. Q.E.D.

This theorem enables us to formulate and analyse the problem of the following type.

Suppose we want to study the effect of the intensity of stimulus over the response, whatever it may be called, of a group of subjects whose number is unknown but supposed to be very large, and the group is assumed to give response when every individual gives response to the stimulus and to give non-response when even one of them gives non-response. The event that the group gives response means that the tolerance of each individuals is less than the intensity of stimulus applied, and hence the tolerance of the group is the maximum of the tolerances of all the individuals in our concern. As the number of individuals in the group is supposed to be very large, the distribution of the tolerance of the group should be of the type $exp[-exp[-(Ax+B)]]$.

The disinfection of bacteria will be the most standard example of it, and the arguments above are the mathematical justification of the model and the analysis of disinfection time data proposed by Mather [5], whose justifica tion such as above is not given by him.

It should be emphasized especially to the practical statisticians that the number of individuals should be enough large so that the limit distribution will be valid, and also that the experiments are so controled well that the variation of the number of individuals from one experiment to another will be negligibly small and the effects of its variation of both A and B may be neglected. Otherwise the experiments can not be analysed by the method due to Mather.

§3. **The case when the number of individuals is not large.** In case the number of the individual in an experiment is not very large, we can not

make use of the limit theorem as in § 2, and the natural assumption will be that the number of individuals in each experiment is distributed in Poisson distribution.

The probability that the group under an experiment will contain exactly n individuals and all the tolerances of the individuals will not exceed the intensity x of the stimulus applied on the group will be

$$e^{-\lambda}\frac{\lambda^n}{n!}(F(x,\theta))^n,$$

where λ is the average number of individuals and $F(x|\theta)$ is the tolerance distribution of the individuals in our concern. Summing up these probabilities for all n, the tolerance distribution will be given by $e^{-\lambda(1-F(x|\theta))}$.

Before we proceed to the analysis of the data of the experiments according to this model, we shall discuss the comparison of these models.

In the first model given in § 2, the parameters A and B represents only the relation between the group of a given number of individuals and the stimulus, but not the one between the individual and the stimulus, where θ in the model in this section represents the relation between each one of the individual and the stimulus. This is the main point in the difference of meanings of two models.

For instance, if we want to study the effect of a medicine on a sort of bacteria, our main interest will not be confined to the effect of a medicine on the solution of bacteria at a given density, but also the effect of the medicine on individual bacteria, namely the tolerance distribution of the individual bacteria, from which the tolerance distribution of the solution should be resulted. In this case the model due to Mather will not be valid, and the model in this section will be useful, as we can separate the parameters expressing the effect and the one expressing the characteristics of the group, the practical namely the average numbers in the solution.

For the calculation of the maximum likelihood estimates of these parameters, we need a provisional values of them, as the likelihood equations are very much complicated. After obtaining a provisional value in some way, we shall show, we can calculate the approximations of maximum likelihood estimates recursively, and the general theory of numerical calculus states that if the series of approximations converges we can say the limit should be the maximum likelihood estimates.

For the calculation of the approximations from the provisional values, or from the approximations of previous stage we may employ the following method.

Let $P(x;\alpha,\beta,\mu)=e^{-\mu(1-F(x;\alpha,\beta))}$ be the binomial probability, with controlable variable x and unknown parameters α, β and μ, and suppose experiments were carried out at the level of controlable variables $x_1, x_2, \cdots, x_k$, with the sample sizes $n_1, n_2, \cdots, n_k$. In our case x represents the intensity

of stimulus, and μ is the average number of individuals in the group, and $F(x;\alpha,\beta)$ is the tolerance distribution of individuals.

Let $r_1, r_2, \cdots, r_k$ be the number of success in each experiment, then the likelihood function is given by

$$(3.1)\qquad \prod_{i=1}^{k}\left(\frac{n_i!}{r_i!(n_i-r_i)!}\right)(P(x_i;\alpha,\beta,\mu))^{r_i}(1-P(x_i;\alpha,\beta,\mu))^{n_i-r_i},$$

which we shall write as e^L, and the maximum likelihood estimates will be given by solving the equations

$$(3.2)\qquad \frac{\partial L}{\partial\alpha}=\frac{\partial L}{\partial\beta}=\frac{\partial L}{\partial\mu}=0.$$

Let (α_1,β_1,μ_1) be the provisional value of the maximum likelihood estimates, and let (α_0,β_0,μ_0) be their true values, and further let us write

$$(3.3)\qquad \begin{aligned}\delta\alpha&=\alpha_0-\alpha_1,\\ \delta\beta&=\beta_0-\beta_1,\\ \delta\mu&=\mu_0-\mu_1,\end{aligned}$$

expanding $\frac{\partial L}{\partial\alpha}$, $\frac{\partial L}{\partial\beta}$ and $\frac{\partial L}{\partial\mu}$ around (α_1,β_1,μ_1) and taking the terms up till the second order, we have the following equations

$$(3.4)\qquad \begin{aligned}&L_\mu(\alpha_1,\beta_1,\mu_1)+\delta\mu L_{\mu\mu}(\alpha_1,\beta_1,\mu_1)+\delta\alpha L_{\mu\alpha}(\alpha_1,\beta_1,\mu_1)\\&\qquad+\delta\beta L_{\mu\beta}(\alpha_1,\beta_1,\mu_1)=0,\\ &L_\alpha(\alpha_1,\beta_1,\mu_1)+\delta\mu L_{\mu\alpha}(\alpha_1,\beta_1,\mu_1)+\delta\alpha L_{\alpha\alpha}(\alpha_1,\beta_1,\mu_1)\\&\qquad+\delta\beta L_{\alpha\beta}(\alpha_1,\beta_1,\mu_1)=0,\\ &L_\beta(\alpha_1,\beta_1,\mu_1)+\delta\mu L_{\mu\beta}(\alpha_1,\beta_1,\mu_1)+\delta\alpha L_{\alpha\beta}(\alpha_1,\beta_1,\mu_1)\\&\qquad+\delta\beta L_{\beta\beta}(\alpha_1,\beta_1,\mu_1)=0.\end{aligned}$$

The solution of the equations above may be taken as its approximates to $(\delta\alpha,\delta\beta,\delta\mu)$.

And also we have

$$(3.5)\qquad \begin{aligned}L_\alpha&=\sum\frac{n(p-P)}{PQ}P_\alpha,\\ L_{\alpha\alpha}&=\sum\frac{PQ(-nP_\alpha)-n(p-P)P_\alpha(1-2P)}{(PQ)^2}P_\alpha+\sum\frac{n(p-P)}{PQ}P_{\alpha\alpha},\\ L_{\alpha\beta}&=\sum\frac{PQ(-nP_\beta)-n(p-P)P_\beta(1-2P)}{(PQ)^2}P_\alpha+\sum\frac{n(p-P)}{PQ}P_{\alpha\beta},\\ &\text{etc.}.\end{aligned}$$

For the purpose of approximation we may take $p=P$ in the second order derivatives and thus we have

$$
\begin{aligned}
L_{\alpha\alpha} &\simeq -\sum \frac{n}{PQ} P_\alpha^2 , \\
L_{\alpha\beta} &\simeq -\sum \frac{n}{PQ} P_\alpha P_\beta , \\
&\text{etc.}
\end{aligned} \tag{3.6}
$$

Therefore the solutions of the following equations

$$
\begin{aligned}
\delta\mu \sum \frac{n}{PQ} P_\mu^2 + \delta\alpha \sum \frac{n}{PQ} P_\mu P_\alpha + \delta\beta \sum \frac{n}{PQ} P_\mu P_\beta &= \sum \frac{n(p-P)}{PQ} P_\mu , \\
\delta\mu \sum \frac{n}{PQ} P_\mu P_\alpha + \delta\alpha \sum \frac{n}{PQ} P_\alpha^2 + \delta\beta \sum \frac{n}{PQ} P_\alpha P_\beta &= \sum \frac{n(p-P)}{PQ} P_\alpha , \\
\delta\mu \sum \frac{n}{PQ} P_\mu P_\beta + \delta\alpha \sum \frac{n}{PQ} P_\alpha P_\beta + \delta\beta \sum \frac{n}{PQ} P_\beta^2 &= \sum \frac{n(p-P)}{PQ} P_\beta ,
\end{aligned} \tag{3.7}
$$

may be taken as the approximations of the solutions of (3.4) and also $(\alpha_1 + \delta\alpha, \beta_1 + \delta\beta, \mu_1 + \delta\mu)$ may be taken as the approximations of the maximum likelihood estimates.

Then we have the following

THEOREM 2. In case $F(x; \alpha, \beta)$ is assumed to be of the form

$$
F(x; \alpha, \beta) = \int_{-\infty}^{\alpha+\beta x} Z(x)\, dx , \tag{3.8}
$$

where $Z(x)$ is some known function, the solution of (3.7) minimizes

$$
\begin{aligned}
M = \sum_i n_i \frac{P(x_i; \alpha_1, \beta_1, \mu_1)}{Q(x_i; \alpha_1, \beta_1, \mu_1)} \{ y_i + \mu G(x_i; \alpha_1, \beta_1) - \alpha\mu_1 Z(\alpha_1 + \beta_1 x_i) \\
- \beta\mu_1 x_i Z(\alpha_1 + \beta_1 x_i) \}^2 ,
\end{aligned} \tag{3.9}
$$

where y_i is defined by

$$
\begin{aligned}
y_i = \frac{p_i - P(x_i; \alpha_1, \beta_1, \mu_1)}{P(x_i; \alpha_1, \beta_1, \mu_1)} + \mu_1(\alpha_1 + \beta_1 x_i) Z(\alpha_1 + \beta_1 x_i) \\
- \mu_1 G(x_i; \alpha_1, \beta_1) ,
\end{aligned} \tag{3.10}
$$

and $G(x; \alpha, \beta)$ is defined by

$$
G(x; \alpha, \beta) = 1 - F(x; \alpha, \beta) . \tag{3.11}
$$

PROOF. Inserting each derivative in (3.7) the value calculated from the relation $P(x; \alpha, \beta, \mu) = e^{-\mu(1-P(x; \alpha, \beta))}$ and (3.8), we have the following equations

$$
\begin{aligned}
&-\delta\mu\sum\frac{nP}{Q}G^2+\delta\alpha\sum\frac{nP}{Q}G\mu_1Z+\delta\beta\sum\frac{nP}{Q}G\mu_1Zx=\sum\frac{n(p-P)}{Q}G,\\
&-\delta\mu\sum\frac{nP}{Q}GZ+\delta\alpha\sum\frac{nP}{Q}\mu_1Z^2+\delta\beta\sum\frac{nP}{Q}\mu_1Z^2x=\sum\frac{n(p-P)}{Q}Z,\\
&-\delta\mu\sum\frac{nP}{Q}GZx+\delta\alpha\sum\frac{nP}{Q}\mu_1Z^2x+\delta\beta\sum\frac{nP}{Q}\mu_1Z^2x^2\\
&\qquad\qquad=\sum\frac{n(p-P)}{Q}Zx.
\end{aligned}
\tag{3.12}
$$

By making use of the definition of y, we have the following equations,

$$
\begin{aligned}
&-\mu\sum\frac{nP}{Q}G^2+\alpha\sum\frac{nP}{Q}G\mu_1Z+\beta\sum\frac{nP}{Q}G\mu_1Zx=\sum\frac{nP}{Q}Gy,\\
&-\mu\sum\frac{nP}{Q}GZ+\alpha\sum\frac{nP}{Q}\mu_1Z^2+\beta\sum\frac{nP}{Q}\mu_1Z^2x=\sum\frac{nP}{Q}Zy,\\
&-\mu\sum\frac{nP}{Q}GZx+\alpha\sum\frac{nP}{Q}\mu_1Z^2x+\beta\sum\frac{nP}{Q}\mu_1Z^2x^2=\sum\frac{nP}{Q}Zxy,
\end{aligned}
\tag{3.13}
$$

where $\mu=\mu_1+\delta\mu$, $\alpha=\alpha_1+\delta\alpha$, $\beta=\beta_1+\delta\beta$ and value of G, Z, y are the ones when $\mu=\mu_1$, $\alpha=\alpha_1$, $\beta=\beta_1$. Q.E.D.

§ 4. Further discussion.

The authors understand that data of this type can be analysed in the same method as was proposed by Finney [2], and there is some similarity between our problem and Wadley's problem. The method proposed by Finney makes approximation on both the first and the second derivatives in (3.4), whereas ours makes the same on the second ones only, and our method appears to be more natural way of analysis along the usual method of probit analysis. Moreover there are essential differences between our problem and Wadley's one, as in the later one the number of individuals giving non-response is assumed to be possible to be counted, and also the analysis of Wadley's model enables us to study only the effects of the stimulus on each individuals but not the effect on, so to say, the group itself. Therefore in the example giving in the paper [2] repeated experimentations at the same intensity of stimulus were not required. Therefore the authors dare to present their result here.

There remains an important problem unsolved as to finding out the first provisional value of (α, β, μ). There do exist situations when we cannot count the number of individuals under aplication of stimulus. For instance the data presented by Mather is produced through the experiments of such nature.

The authors regret that they cannot give any example of the analysis. The data so far available to them is only the one given in the paper of Mather [5]. As our model assumes $P=e^{-\mu(-\lambda)}$, we have $F=1-1/\mu\cdot\log P$,

we first calculated $1-1/\mu \cdot \log r/n$, for various value of μ, and then, assuming that $F(x; \alpha, \beta)$ be log-normal distribution function, we calculated the probit transformation of $1-1/\mu \cdot \log r/n$. The graphical representation shows $\mu = 150$ gives the closest fit to the straight line with $\alpha = 1.3$ and $\beta = 5.2$. Taking them as the first approximates, we calculated the approximations to the maximum likelihood estimates which, however, failed to converge. We tried to analize the same according to the method due to Finney [3], which resulted in the same. Therefore further investigations concerning the tolerance distribution of individual bacterium cell would be necessary.

The method above would be suitable for finding out the provisional value.

§ 5. **Acknowledgements.** The authors are gratefull to professor T. Kitagawa for his encouragements and criticism while this paper was beeing prepared.

References

[1] D. J. FINNEY; *Probit analysis*, Cambridge, (1952).
[2] D. J. FINNEY; *The estimation of the parameters of tolerance distributions*, Biometrika, **36**, (1949), 239-256.
[3] R. A. FISHER; *Limitting form of the frequency distribution of the largest or smallest member of a sample*, Proceedings of the Cambridge Philosophical Society, **24**, (1928), 180-190.
[4] E. J. GUMBEL; *Statistical theory of extreme values and some practical applications*, National Bureau of Standards Applied Mathematical Series, **33**, (1954).
[5] K. MATHER; *The analysis of extinction time data in bioassay*, Biometrics, **5**, (1949), 127-143.