NOTES ON PROBIT ANALYSIS

S. R. Searle*

BU-202-M July, 1965

Abstract

These notes consist of a brief outline of the statistical and mathematical procedures involved in probit analysis. They are based on "Probit Analysis", by D. J. Finney, Cambridge University Press.

* Cornell Computing Center and Biometrics Unit, Cornell University.

NOTES ON PROBIT ANALYSIS

S. R. Searle*

BU-202-M July, 1965

INTRODUCTION

These notes provide a brief outline of the statistical and mathematical procedures involved in probit analysis. They are based on "Probit Analysis", D. J. Finney, Cambridge University Press, 1952, which should be consulted for details and for illustrative examples. Numbers enclosed in square brackets are references to sections of Finney's book. Notation used here is similar but not identical to that of Finney.

DEFINITIONS

The basic objective of probit analysis is the study of threshold tolerances to stimuli. The problem is usually thought of in terms of insecticide dosing. The following definitions establish the situation usually considered.

Threshold tolerances

If λ represents a dose level, we suppose that $f(\lambda)$ is the distribution of threshold tolerances throughout the population. Then, [5],

$f(\lambda)d\lambda$ = proportion of the population whose tolerance limit (threshold) lies between dose levels λ and $\lambda + d\lambda$.

Proportion of population responding

P_i = proportion of the population that respond to a dose of level λ_i.

This proportion includes, of course, all those members of the population whose

---

* Cornell Computing Center and Biometrics Unit, Cornell University.

threshold tolerance is at some dose level lies than λ_i. (Insects that would die from a dose level of $\lambda < \lambda_i$ will also die from a dose level λ_i.) Therefore,

$$P_i = \int_0^{\lambda_i} f(\lambda)d\lambda \ .$$

Normality transformation [5]

It so happens that the distribution $f(\lambda)$ is usually skew (a few insects can tolerate very high doses). But the distribution of $x = \log_{10}\lambda$ is usually sufficiently near normal that there is no serious error in assuming that it is normal. We therefore make the transformation

$$x_i = \log_{10}\lambda_i$$

for which the distribution of x_i is

$$f(x_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}} \ .$$

The proportion P_i is accordingly redefined as

$$P_i = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x_i} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx \qquad \text{- - - - - - - (1)}$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{(x_i-\mu)/\sigma} e^{-t^2/2} dt \ . \qquad \text{- - - - - - - (2)}$$

Estimation problem

The problem is to estimate μ and σ^2, the parameters of the (assumed) normal distribution of $x = \log_{10}\lambda$ where λ is the dose level. Knowing μ and σ^2 we could then, from (1), find P_i for any given x_i.

Dose and dosage [5]

λ_i is referred to by Finney as dose, and the corresponding $x_i (= \log_{10}\lambda_i)$ as dosage.

Probit [9]

A probit value Y is defined as follows:

$$Y \text{ is the probit of } \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(Y-5)} e^{-t^2/2}\, dt. \quad \text{-----(3)}$$

This is a definition. It has nothing to do, per se, with the study of insecticides, but because of the importance of equation (2) in describing the insecticide problem, the concept of a probit is useful. By comparing (2) and (3) we find that

$$Y_i = 5 + (x_i - \mu)/\sigma \quad \text{-----(4)}$$

is the probit of P_i defined by (1).

Equation (4) is equivalent to

$$Y_i = \alpha + \beta x_i \quad \text{-----(5)}$$

where $$\alpha = 5 - \mu/\sigma \text{ and } \beta = 1/\sigma\,, \quad \text{-----(6)}$$

or $$\mu = (5-\alpha)/\beta \text{ and } \sigma = 1/\beta\,. \quad \text{-----(7)}$$

Furthermore, by using (4), equation (2) can be written as

$$P_i = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Y_i - 5} e^{-u^2/2}\, du\,. \quad \text{-----(8)}$$

In equations (8) and (5) we see the basis for the probit analysis. If, for a dose λ_i we have an estimate from data of P_i, then tables of the normal distribution will yield from equation (8) a corresponding estimate of Y_i. But for any λ_i the dosage is $x_i = \log_{10}\lambda_i$, and equation (5) can then be used to estimate α and β from a regression of Y_i on x_i. Estimation by maximum likelihood leads to procedures not as simple as this, although iterative approximations thereto are almost as simple.

ESTIMATION

Data

We suppose that the insecticide has been tested on groups of insects at

k different dose levels, $d_1, d_2, \ldots, d_k$. For dose level d_i

n_i = number of insects tested

r_i = number of insects that respond (dead)

$p_i = r_i/n_i$

$x_i = \log_{10} d_i$

P_i = true, unknown, proportion of population that respond at dose level λ_i

$Q_i = 1 - P_i$

Likelihood [7 and App. II]

The situation under discussion is a binomial one and the likelihood of the sample is

$$\prod_{i=1}^{k} {}^{n_i}C_{r_i} P_i^{r_i} (1-P_i)^{n_i-r_i} .$$

Apart from a constant, the logarithm of this is

$$\sum_{i=1}^{k} r_i \log P_i + \sum_{i=1}^{k} (n_i-r_i) \log (1-P_i) . \qquad ------(9)$$

Estimation by maximum likelihood

Estimation of α and β of equation (5) is achieved by maximizing L with respect to α and β after substituting (5) into (8) and then (8) into (9). This is done by solving the equations

$$\frac{dL}{d\alpha} = 0 \text{ and } \frac{dL}{d\beta} = 0 . \qquad ------(10)$$

Note: the symbol d is used for partial differentiation.

To develop equations (10) after substituting from (5) and (8) we note that they are equivalent to

$$\sum_i \frac{dL}{dP_i} \frac{dP_i}{dY_i} \frac{dY_i}{d\alpha} = 0 \text{ and } \sum_i \frac{dL}{dP_i} \frac{dP_i}{dY_i} \frac{dY_i}{d\beta} = 0 . \quad ------(11)$$

The four terms involved in these equations are as follows. From (9)

$$\frac{dL}{dP_i} = \frac{n_i(p_i - P_i)}{P_i Q_i} ;$$

from (8) $$\frac{dP_i}{dY_i} = Z_i$$

where $$Z_i = \frac{1}{\sqrt{2\pi}} e^{-(Y_i-5)^2/2} ; \qquad \text{- - - - - - - - - -(12)}$$

and from (5)

$$\frac{dY_i}{d\alpha} = 1 \text{ and } \frac{dY_i}{d\beta} = x_i .$$

Substitution in (11) gives

$$\sum \frac{n_i(p_i - P_i)Z_i}{P_i Q_i} = 0 \text{ and } \sum \frac{n_i(p_i - P_i)Z_i x_i}{P_i Q_i} = 0 . \quad \text{- -(13)}$$

Summations here, and in all that follows, are with respect to the parameter i, for i = 1, 2, ..., k, the number of dose levels.

These are the equations that have to be solved for α and β after substituting

for P_i from (8)

for Q_i as $1 - P_i$

for Z_i as in (12)

for Y_i as $\alpha + \beta x_i$

where x_i is an observation, $x_i = \log_{10} d_i$. It is clear that after making these substitutions no explicit solution for $\alpha + \beta$ can be found. However, a first approximation can be made.

First approximation

Using p_i as the value of P_i, equation (8) yields a first approximation to Y_i. This is often called the empirical probit. With Y_i so obtained, and

$x_i = \log_{10} d_i$, a regression analysis based on $Y_i = \alpha + \beta x_i$ as in (5) yields first approximations to α and β. They shall be denoted by a_1 and b_1 respectively.

Iterative approximations

An application of the Taylor-MacLaurin expansion leads to a method for improving the first estimates a_1 and b_1 of α and β suggested in the preceding paragraph. Better estimates are

$$a_2 = a_1 + a' \qquad \text{- - - - - - - - - - - -(14)}$$

and $$b_2 = b_1 + b'$$

where a' and b' are obtained by solving the equations

$$a' \sum \frac{n_i Z_i^2}{P_i Q_i} + b' \sum \frac{n_i Z_i^2 x_i}{P_i Q_i} = \sum \frac{n_i Z_i^2}{P_i Q_i} \frac{(p_i - P_i)}{Z_i} \qquad \text{- - - - -(15)}$$

and $$a' \sum \frac{n_i Z_i^2 x_i}{P_i Q_i} + b' \sum \frac{n_i Z_i^2 x_i^2}{P_i Q_i} = \sum \frac{n_i Z_i^2 x_i}{P_i Q_i} \frac{(p_i - P_i)}{Z_i} ,$$

where, for what is called the provisional probit,

$$Y_i = a_1 + b_1 x_i ,$$

we have $$Z_i = \left(1/\sqrt{2\pi}\right) e^{-\frac{1}{2}(Y_i - 5)^2} , \qquad \text{- - - - -(16)}$$

$$P_i = \left(1/\sqrt{2\pi}\right) \int_{-\infty}^{Y_i - 5} e^{-t^2/2}\, dt$$

and $$Q_i = 1 - P_i .$$

Note: the expression for Y_i: it involves a_1 and b_1.

Thus is the iterative procedure set up: when equations (15) are solved for a' and b', (14) gives a_2 and b_2: using these in (16) enables another solution for a' and b' to be found from (15), and hence a_3 and b_3, and so on.

Simplification of iterative procedure

Define $$w_i = Z_i^2/P_iQ_i \qquad \text{- - - - - - - - - - - - - - - (17)}$$

then equations (15) become

$$\begin{aligned} a'\Sigma n_i w_i &+ b'\Sigma n_i w_i x_i = \Sigma n_i w_i (p_i - P_i)/Z_i \\ a'\Sigma n_i w_i x_i &+ b'\Sigma n_i w_i x_i^2 = \Sigma n_i w_i x_i (p_i - P_i)/Z_i \end{aligned} \qquad \text{- - - - - - (18)}$$

Now define what is often called the working probit:

$$y_i = Y_i + (p_i - P_i)/Z_i \,. \qquad \text{- - - - - - - - - - - - (19)}$$

Then equations (18) become

$$\begin{aligned} \Sigma n_i w_i Y_i &+ a'\Sigma n_i w_i + b'\Sigma n_i w_i x_i = \Sigma n_i w_i y_i \\ \Sigma n_i w_i x_i Y_i &+ a'\Sigma n_i w_i x_i + b'\Sigma n_i w_i x_i^2 = \Sigma n_i w_i x_i y_i \,. \end{aligned} \qquad \text{- - - (20)}$$

But in these equations Y_i has the form $Y_i = a_1 + b_1 x_i$. Substitution in (20) leads to

$$\begin{aligned} \Sigma n_i w_i (a_1 + b_1 x_i) &+ a'\Sigma n_i w_i + b'\Sigma n_i w_i x_i = \Sigma n_i w_i y_i \\ \Sigma n_i w_i x_i (a_1 + b_1 x_i) &+ a'\Sigma n_i w_i x_i + b'\Sigma n_i w_i x_i^2 = \Sigma n_i w_i x_i y_i \,. \end{aligned} \qquad \text{- - (21)}$$

Now, of course, $a_1 + a' = a_2$ and $b_1 + b' = b_2$, and so equations (21) can be written as

$$\begin{aligned} a_2\Sigma n_i w_i &+ b_2\Sigma n_i w_i x_i = \Sigma n_i w_i y_i \\ a_2\Sigma n_i w_i x_i &+ b_2\Sigma n_i w_i x_i^2 = \Sigma n_i w_i x_i y_i \,. \end{aligned} \qquad \text{- - - - - - - - -(22)}$$

But these are exactly the equations for a weighted regression of y_i on x_i using $n_i w_i$ as weights. Hence, for

$$\tilde{x} = \Sigma n_i w_i x_i / \Sigma n_i w_i$$

and

$$\tilde{y} = \Sigma n_i w_i y_i / \Sigma n_i w_i \,,$$

the solutions to (22) for a_2 and b_2 are

$$a_2 = \tilde{y} - b_2\tilde{x}$$

$$b_2 = \frac{\Sigma n_i w_i x_i y_i - \tilde{x}\tilde{y}(\Sigma n_i w_i)}{\Sigma n_i w_i x_i^2 - \tilde{x}^2(\Sigma n_i w_i)} \,.$$

The procedure for obtaining a_2 and b_2 therefore resolves itself into being a weighted regression of y_i on x_i using weights $n_i w_i$ where:

d_i = dose

$x_i = \log_{10} d_i$

p_i = proportion of n_i insects affected by d_i

y_i is as given in (19) and (16)

w_i is as given in (17) and (16) .

Having obtained a_2 and b_2 from a_1 and b_1, the process can be repeated to derive a_3 and b_3 from a_2 and b_2: and so on.

Although the values of a_1 and b_1 could, as described above, be derived from the simple regression of Y_i on x_i, a weighted regression is customarily used. The weights are $n_i w_i$ where w_i is calculated using p_i for P_i, in consequence of which $y_i = Y_i$, namely the first working probit is taken as the empirical probit. The whole procedure therefore resolves itself into being an iterative weighted regression of y_i on x_i using weights $n_i w_i$.

Summary of procedure

Preliminary calculations are:

dosage: $x_i = \log_{10} d_i$

response: $p_i = r_i/n_i$

empirical probit: Y_{1i} from $p_i = \left(1/\sqrt{2\pi}\right) \int_{-\infty}^{Y_{1i}-5} e^{-u^2/2} du$.

At the t'th round of the iterations, for $t > 1$, the estimates a_{t-1} and b_{t-1} obtained at the preceding round are used, and the steps involved are:

$Y_{ti} = a_{t-1} + b_{t-1} x_i$ (provisional probit)

$Z_{ti} = \left(1/\sqrt{2\pi}\right) e^{-(Y_{ti}-5)^2}$

$P_{ti} = \left(1/\sqrt{2\pi}\right) \int_{-\infty}^{Y_{ti}-5} e^{-u^2/2} du$

$$Q_{ti} = 1 - P_{ti}$$

$$w_{ti} = Z_{ti}^2 / P_{ti} Q_{ti}$$

$$y_{ti} = Y_{ti} + (p_i - P_{ti})/Z_{ti} \qquad \text{(working probit)}$$

The weighted regression of y_{ti} on x_i using weights $n_i w_{ti}$ is then as follows:

$$\tilde{y}_t = \Sigma n_i w_{ti} y_{ti} / \Sigma n_i w_{ti}$$

$$\tilde{x}_t = \Sigma n_i w_{ti} x_{ti} / \Sigma n_i w_{ti}$$

$$SXX_t = \Sigma n_i w_{ti} x_i^2 - (\Sigma n_i w_{ti}) \tilde{x}_t^2$$

$$SYY_t = \Sigma n_i w_{ti} y_{ti}^2 - (\Sigma n_i w_{ti}) \tilde{y}_t^2$$

$$SXY_t = \Sigma n_i w_{ti} x_i y_{ti} - (\Sigma n_i w_{ti}) \tilde{x}_t \tilde{y}_t$$

$$b_t = SXY_t / SXX_t$$

$$a_t = \tilde{y}_t - b_t \tilde{x}_t \, .$$

For the first iteration, t = 1, the working probit y_{1i} is taken as the empirical probit Y_{1i} obtained in the preliminary calculations, and P_{1i} is taken as p_i in deriving w_{1i}.

Goodness of fit

When a satisfactory fit has been obtained, the measure of its goodness of fit is:

$$X^2_{k-2} = \sum \frac{(r_i - n_i P_i)^2}{n_i P_i (1 - P_i)}$$

$$= \sum \frac{n_i (p_i - P_i)^2}{P_i Q_i}$$

$$= \Sigma n_i w_i \left(\frac{p_i - P_i}{Z_i}\right)^2 , \qquad \text{from (17)}$$

$$= \Sigma n_i w_i (y_i - Y_i)^2 , \qquad \text{from (19)}$$

$$= \Sigma n_i w_i (y_i - \hat{\alpha} - \hat{\beta} x_i)^2$$

$$= SYY - (SXY)^2/SXX .$$

Therefore the χ^2 value from the weighted regression represents, directly, the goodness of fit of predicted P_i's to the observed p_i's. If the computed χ^2 is significant, the data are said to be heterogeneous and a heterogeneity factor is computed as $\sigma^2 = \chi^2_{k-2}/(k-2)$. Otherwise it is taken as $\sigma^2 = 1$. In either case $\text{var}(b) = \sigma^2/SXY$.

Median dose

The median dose m, is defined as being the dose which affects 50% of the population; i.e. the dose for which $P_i = 50\%$. From equation (2) x_i then equals μ, which from (7) gives $x_i = m$ for $P_i = 50\%$ as

$$m = (5-\alpha)/\beta .$$

For estimated values of α and β this is

$$m = (5-\bar{y}+\beta\bar{x})/\beta$$

$$= \bar{x} + (5-\bar{y})/\beta .$$

The corresponding dose level is 10^m.