CHAPTER 1

INTRODUCTION

1.1 Introduction to the problem

Suppose that K different substances which have the same basic constitution and which are used for the same purpose, but originated at different sources are available. These substances may be fertilizers, insectisides, drugs, vitamins or any other.

It is natural to consider a substance of the above type to be good, if it produces the desired effect at a relatively lower dose, when applied to subjects of interest. The problem is to indentify the substance which is the efficient or 'the best' out of the K substances which are available.

In order to do so, one should know about the nature of responses of the individuals in the population, for each of K substances. Note that the population of the subjects, who may be given one or the other of K substances is fixed. The response of the same individual may be different for different substances or to the same substance at different levels of the dose.

With respect to a fixed substance, it is assumed that with every individual in the given population is associated a particular level of dose Z called as threshold or

tolerance of the individual, such that the individual responds at any dose higher than Z and does not respond at any dose below Z .

Thus tolerance is a random variable varying from individual to individual. Let it be denoted by Z. Let F*(z) denote the distribution of the tolerance, also called as tolerance distribution.

F*(z) = Probability that Z is less than or equal to z.

= Prob ($Z \leq z$)

The above probability can be interpreted in two ways :

i) It is the proportion of individuals responding at dose Z, that is the proportion of individuals whose tolerance is less than or equal to z; and

ii) It is the probability that a randomly selected individual will respond to the dose Z, which means that randomly selected individual will have its tolerance less than z .

The tolerance distribution is not usually completely known. Either

i) Form of distribution is known but for some unknown parameters; or

ii) The form itself is not known, wherein we have to consider non-parametric methods.

In this dissertation we consider only the first case. Generally it is observed that the tolerance distribution is positively skewed. So we need to make a transformation of the tolerance say $x = h(z)$. This transformed variable is known as dose metameter.

Usually the function h is taken to be a logarithmic function or square root function; that is $x = \log z$ or $x = z^{1/2}$. In the following discussion x denotes any dose metameter and the tolerance distribution refers to that of the transformed random variable.

It is found that log tolerance distribution has density given by

$$g(x) = \beta\, g(\alpha+\beta x) \qquad 1.1.1$$

where g is the standard density of one of the forms listed below. This is equivalent to assuming that the log tolerance distribution has location parameter $-\alpha/\beta$ and scale parameter β.

The standard densities refered above are

1] The uniform distribution over [0,1]

$$g(u) = \begin{cases} 1 & 0 \leq u \leq 1 \\ 0 & \text{otherwise} \end{cases} \qquad 1.1.2$$

$$1 < u < \infty$$

2] The standard normal distribution

$$g(u) = \frac{1}{\sqrt{\pi}} e^{-u^2/2} \qquad -\infty<u<\infty \qquad 1.1.3$$

3] The standard Cauchy distribution

$$g(u) = \frac{1}{\pi} \frac{1}{1+u^2} \qquad -\infty<u<\infty \qquad 1.1.4$$

4] The Exponential distribution

$$g(u) = \begin{cases} e^{-u} & o < u<\infty \\ 0 & \text{Otherwise} \end{cases} \qquad 1.1.5$$

5] The Laplace or double exponential distribution

$$g(u) = \frac{1}{2} e^{-|u|} \qquad -\infty<u<\infty \qquad 1.1.6$$

6] The Logistic distribution

$$g(u) = \frac{e^{-u}}{(1+e^{-u})^2} \qquad -\infty<u<\infty \qquad 1.1.7$$

7]

$$g(u) = \frac{1}{2} \operatorname{sech}^2 u \qquad -\infty<u<\infty \qquad 1.1.8$$

8] Sine distribution

$$g(u) = \begin{cases} \sin 2u & o < u < \pi/2 \\ 0 & \text{Otherwise} \end{cases} \qquad 1.1.9$$

The above discussion is restricted to a particular substance. We assume that form of the tolerance distribution is same for all the K substances, but they differ in one or more parameter values α and β. If the tolerance distribution for each substance is completely known, one

can compare the substances in terms of some real valued function of the parameters.

In the present problem it is natural to compare the efficiencies in terms of certain percentiles of the distribution. Usually three percentiles are considered for the tolerance distribution.

i) ED 10 : It is the value of dose at which 10% of the subjects respond.

ii) ED 50 : or Median effective dose which is the level of dose such that 50% of the subjects respond; and

iii) ED 90 : which corresponds to the level of the dose at which 90% of the subjects respond.

The effective doeses ED 10, ED 50 and ED 90 will be called as Lethal doses LD10, LD50, and LD90 respectively whenever the desired response is death of the subject.

We will consider the most commonly used measure, the parametric function ED50.

The expression for ED50 for some tolerance distributions are given further in 1.1.20 to 1.1.22.

Suppose that f(x) is a log tolerance distribution which is a location scale family of distributions. Hence there

exists a distribution function G such that

$$F(x) = G\left(\frac{x-\mu}{\sigma}\right)$$
$$= G(\alpha+\beta x) \qquad 1.1.10$$

Thus proportion of respondants at dose metameter x, as discussed earlier will be

$$P = F(X \leq x)$$
$$= \int_{-\infty}^{x} dF(x)$$
$$= \int_{-\infty}^{x} dG\left(\frac{x-\mu}{\sigma}\right)$$
$$= \int_{-\infty}^{\alpha+\beta x} g(u)\, du \qquad 1.1.11$$

Suppose that at a given dose, the true proportion that responds be P. Since $0 < P < 1$ there exists a real number Y such that

$$\int_{-\infty}^{Y} g(u)\, du = P \qquad 1.1.12$$

This Y is called as equivalent deviate. Note that it depends on the dose level. The equivalent deviate transforms the response at a dose measured as P to Y. This Y is called as respond metameter.

Remark : If we have $P = \int_{-\infty}^{Y-5} g(u)\, du$ 1.1.13

then Y is called as probit.

From equation 1.1.10 we have

From equation 1.1.10 we have

$$Y = \frac{x - \mu}{\sigma}$$

$$= -\frac{\mu}{\sigma} + \frac{1}{\sigma} \qquad 1.1.14$$

By putting $\alpha = -\mu/\sigma$ 1.1.15

and $\beta = 1/\sigma$ 1.1.16

the equation 1.1.13 can be written as

$$Y = \alpha + \beta x \qquad 1.1.17$$

This is called the equivalent deviate regression line.

Thus, eventhough the relationship between dose and response is non-linear, the relationship between dose-metameter and response metameter is linear.

Now ED50 is the dose at which 50% of the subjects respond. Let ED50 be denoted by M. Then M is such that

$$\int_{-\infty}^{\alpha+\beta \log M} g(u)\,du = 0.5 \qquad 1.1.18$$

therefore,

$$G^{-}(0.5) = \alpha + \beta \log M$$

so that

$$\log M = \frac{G^{-1}(0.5) - \alpha}{\beta} \qquad 1.1.19$$

G is distribution function corresponding to density g. The expressions for log ED50 for different tolerance distributions are obtained below :

1) If the log tolerance distribution G is symmetric about zero, then

$$G^{-1}(0.5) = 0$$

and from 1.1.18

$$\log M = \frac{G^{-1}(0.5)-\alpha}{\beta}$$

$$= -\alpha/\beta \qquad 1.1.20$$

In this case, if the mean exists, log M is the mean of the distribution and its variance also exists and is equal to $1/\beta$. We shall let $\mu = -\alpha/\beta$ and $\sigma = 1/\beta$

2) Uniform distribution

$P = \gamma$ is the equivalent deviate in this case,

for ED50 ,

$$\log M = \frac{G^{-1}(0.5)-\alpha}{\beta}$$

$$= \frac{0.5-\alpha}{\beta} \qquad 1.1.21$$

3) Sine distribution

$$P = \int_0^Y f(u)du$$

$$= \int_0^Y \sin 2u du$$

$$= \left[-\frac{\cos 2u}{2} \right]_0^Y$$

$$= 1 - \frac{\cos 2Y}{2}$$

$$= \text{Sin}^2 Y$$

$$\therefore \quad \text{Sin } Y = \sqrt{P}$$

or equivalently

$$Y = \text{Sin}^{-1}(\sqrt{P})$$

at ED50 we have

$$Y = \text{Sin}^{-1}(0.5)$$

$$\therefore \quad \log M = \frac{\text{Sin}^{-1}(0.5)-\alpha}{\beta} \qquad 1.1.22$$

It can be easily seen that for a substance to be the best its ED50 should be minimum of the K values of ED50, since the substances for which subject responds at a lower dose is more efficient.

For ED50 we would concentrate on the case where the tolerance Z has log normal distribution, that is X=logZ has normal distribution. In this case as we have seen log ED50 turns out to be the mean $\mu = -\alpha/\beta$. So if the values of means are known then one can identify the best substance easily. We order the substances according to means and select the one as the best which has smallest mean. But these values are not known in practice. In that case one has to find estimators of μ from each population and based on the estimator one should suggest

a method identifying the best substance. Such problem is called as selection problem.

There are two ways of obtaining the data in the present case. One is to observe the exact dose required by a sample of individuals to show a desired response. This method of obtaining the data is called as direct assay. Based on such data the selection procedures from normal populations are described in chapter 2.

The second method is called as indirect assay. In this method we give different doses of the particular substance to certain number of subjects. This gives number of respondants and proportion of respondants at each dose.

The data will be in the following form :

Table 1

log dose x_i	number of subjects n_i	number of respondants r_i	Proportion of respondants P_i
x_1	n_1	r_1	P_1
x_2	n_2	r_2	P_2
'	'	'	'
'	'	'	'
'	'	'	'
x_l	n_l	r_l	P_l

We have such K sets of data for K substances.

Here data is not directly available, since we do not observe directly the tolerances of the subjects. We have to estimate μ in terms of α and β which are to be estimated. The method of obtaining these estimates is given in chapter 3.

The problem of selecting the best of K normal populations when the data from K populations is as given in the above table is the subject matter of this dissertation and is dealt with in chapter 3. In the meanwhile, we describe the general selection problem in the next section.

1.2 Nature of selection procedure :

Suppose that $\pi_1, \pi_2, \ldots, \pi_k$ are K (Statistical) populations. Suppose that for i=1,2,..,k individuals in the i^{th} population are characterised by a random variable X_i which has distribution F_i. We shall assume that form of F_i remains the same for all the K populations, however they differ in one or more parameter values.

Let $F(x,\theta)$ be a distribution function which depends on a parameter θ and let Θ be the parameter space, which

is a subset of same finite dimensional Euclidian space. We suppose that for $i = 1,2,\ldots,k$ the distribution F_i is equal to F with the value of Q equal to θ_i.

Now, the problem is to identify the population (out of K), which is 'best' in some sense, based on observations from each population. Usually the criteria on of comparing the K populations is expressed in terms of 'largeness' or 'smallness' of a real valued function θ defined on parameter space .

This selection problems involves comparison of the components of K dimensional vector $(Q_1,Q_2,\ldots,Q_k)$ which takes values in a subset of K dimentional function space, where $Q_i = Q(\theta_i)$ $i = 1,2,\ldots,k$. Any vector $(Q_1, Q_2, \ldots, Q_k)$ is called a parameter configuration.

To fix ideas, suppose that i^{th} population is considered to be best iff

$$Q_i = \max (Q_1,Q_2,\ldots,Q_k) \qquad 1.2.1$$

If $Q_{[i]}$ denote i^{th} ordered value of $Q_1,Q_2,\ldots,Q_k$ then the ith population is best iff

$$Q_i = Q_{[k]} \qquad 1.2.2$$

If there is more than one value of i for which 1.2.2 holds, then we may select any one of them.

If the best one differs by at least some minimum threshold value from all the others, then we have a strong preferance to it. For this we define an appropriate distance measure to serve as a measure of differances between population we want to identify and remaining populations. Since we have ordered populations according as

$$Q_{[1]} \leq Q_{[2]} \leq \cdots \leq Q_{[k]} \qquad 1.2.3$$

It is natural to define the distance measure on $Q_{[k]}$ and $Q_{[k-1]}$.

In the parametric case some distance measures usually considered are

1) $$\partial = Q_{[k]} - Q_{[k-1]} \qquad 1.2.4$$

This is the definition of ∂ when parameter is location parameter.

2) When the parameter is scale parameter, we define ∂ as

$$\partial = \frac{Q_{[k]}}{Q_{[k-1]}} \qquad 1.2.5$$

3) When the parameter is neither the scale parameter nor the location parameter, like in binomial model we have

$$\partial = \frac{Q_{[k]}(1-Q_{[k-1]})}{Q_{[k-1]}(1-Q_{[k]})} \qquad 1.2.6$$

which is an odd's ratio.

With reference to a chosen distance measure, the K dimentional subset of possible parameter configuration is divided into

i) Preference Zone ii) Indifferance Zone.

The configurations with a distance larger than ∂^*, a constant known as indifferance constant, are known to form a preferance zone.

The configurations with their distance smaller than ∂^* are said to form an indifferance zone.

Now the selection procedure is as follows :

For the populations $\pi_1, \pi_2, \ldots, \pi_k$ we are interested in their parameter function $Q_1, Q_2, \ldots, Q_k$.

Remark: If the parameter under consideration is real valued then we consider $Q_1, Q_2, \ldots, Q_k$ directly, as it happens in case of selection of means of normal populations. For $i = 1,2,\ldots,k$, we take observations on n_i individuals from the i^{th} population and obtain an appropriate estimator of Q_i, say $\hat{Q}_i$. Then order the estimators as

$$\hat{Q}_{[1]} \leq \hat{Q}_{[2]} \cdots\cdots \leq \hat{Q}_{[k]} \qquad 1.2.7$$

We find the best population as one which gives largest sample estimate of parameter. But it is quite likely that

the population giving largest sample value need not always be the one with largest parameter value. This is called as an error of incorrect selection.

Also the selected populations differ if a different estimator is used.

In this regard we define the probability of correct selection for configuration $\underline{Q}$. It is the probability of correctly choosing a population as best which is actually the population with largest parameter value. Notationally if $\hat{Q}_{(1)},\hat{Q}_{(2)},\ldots,\hat{Q}_{(k)}$ denote the ordered estimator corresponding to $Q_{[1]},Q_{[2]},\ldots,Q_{[k]}$, then probability of correct selection is

$$\text{Prob.}\left[\hat{Q}_{(k)} = \hat{Q}_{[k]}\right]$$

$$= \text{Prob}\left[\hat{Q}_{(k)} = \max\left(\hat{Q}_{(1)},\ldots,\hat{Q}_{(k)}\right)\right]$$

$$= \text{Prob}\left[\hat{Q}_{(k)} = \hat{Q}_{(i)} \text{ for all } i = 1,2,\ldots,k\right]$$

Obviously, this should be large whenever Q is in preferance zone.

Suppose there are 2 selection procedures or rules say R_1 and R_2 for selection of best population. In order to choose and use an appropriate one, we should compare their relative performance in terms of probability of correct selection $P_{(CS/\underline{Q})}$. The procedure which gurantees the larger

P_{CS} is to be preferred. But since $P_{(CS/\underline{Q})}$ is different for different $\underline{Q}$ and we have infinitely many configurations $\underline{Q}$ in the preferance zone, a good idea is to compare in terms of $\inf_{Q} P_{CS/Q}$

The configuration which gives infimum value of $P_{(CS/Q)}$ in the preferance zone is called as least favourable configuration denoted as LFC.

This approach to select a 'best' population which achieves the given P_{CS} is called the 'Indifferance zone approach'. There are two other methods which are generalization of this approach; one known as 'Ranking populations' and the other as 'Subset selection procedure'. these two methods are not discussed in this dissertation.

1.3 A selection problem in bioassay :

Our aim is to apply a selection problem in bioassay In bioassays we have to compare the populations in terms of ED50's, denoted by M.

We know that $\log M = \dfrac{y_{0.5} - \alpha}{\beta}$

$$= \frac{G^{-1}(0.5) - \alpha}{\beta}$$

from 1.1.19.

Now, since log M is the strictly monotonic function of M we can take log M as our function for comparison of populations. Since log (M) turns out to be μ the estimated mean of normal population we first make a survey

of selection problems for means of normal population in various cases in the second chapter.

In the third chapter we try to find out the exact distribution for μ . But it is found that exact solutions cannot be obtained because of the assymprotic property of α and β, μ can only be estimated through α and β. Some examples are worked out in section 3.3.

Sections 3.2 and 3.3 contain original results.