MIXTURE DISTRIBUTION MODELS
FOR CROP LOSS ESTIMATION

6.1 Introduction

A plant affected by a disease is expected to give a reduced yield. Further, if a number of plants in a plot are affected, then we have a mixture of healthy and diseased plants. If we consider the healthy plants as one group (subpopulation) and the diseased ones as another (subpopulation), the plot as a whole is now an intimate mixture of two subpopulations. If the average yield and variance for healthy group of plants are different from the average yield and variance for the diseased group of plants, then the overall average and variance for this 'mixture' population of plants will depend upon these means and variances and also on the 'mixing proportion' of the two groups in the whole population of plants in the plot. The resulting crop loss in such plots was estimated in the previous chapters from the view point of sampling from a finite population. Since prior stratification of the population of plants into healthy and diseased groups was not feasible, the methods of crop loss estimation fell broadly into two types, viz., SRSWOR schemes with fixed sample size and
inverse sampling schemes with a random sample size. A significant point to note in respect of all the foregoing methods is that no assumption was made about the nature of (probability) distributions of yield observations, either in the healthy group or in the diseased group or in the 'mixture' of these two groups. From experience it is found that yield observations follow a normal distribution in a large number of cases.

When this assumption of normality is valid with one normal distribution for the healthy group and another for the diseased group, the yield in the population of plants as a whole follows a 'mixture' of two normal distributions, mixed in a certain proportion. It may also be noted that such a mixture need not necessarily be normal or even unimodal (Hasselblad, 1966). In general such a mixture distribution need not necessarily be a mixture of two normal distributions only. It could be a mixture of say, two exponential distributions or one a normal and another an exponential distribution. For example yield observations in respect of healthy plants may follow a normal distribution and that for diseased plants an exponential distribution. Sometimes experience may shed some light on the nature of this mixture population although the numerical values of characterising parameters of individual 'mixing' subpopulations may be unknown.
The shapes of two or more distributions mixed in different proportions are given in Everitt and Hand (1981) for typical cases.

The problem of estimation of parameters of a 'mixture distribution' in terms of the parameters of 'mixing' distributions and the mixing proportion largely depends upon the nature and amount of input information. For instance, the input information may be in the form of sample observations on only one character (yield) without any information on the group membership (healthy or diseased) of each observation. In this case the mixture distribution is univariate. If the input information also includes the knowledge about group membership (healthy or diseased) the resulting mixture distribution is bivariate. Accordingly the methods of estimation too depend on this.

A lot of literature is available for estimation of parameters of a mixture distribution in the former case.

the situations when mixture model is bivariate are not many. Two important references in this respect are Sclove (1977), Hosmer and Dick (1977).

In crop loss estimation the group membership may be always noted if plants are drawn one by one and harvested individually. Thus a bivariate mixture model is more appropriate. In this chapter two such mixtures will be considered:

(i) A mixture of two normal distributions;
(ii) A mixture of two exponential distributions. These are cases with fairly wide applicability. Methods for estimating the characterising parameters and crop loss are developed. The utility of such models is illustrated in chapters 10 and 11 for simulated data and also real life data from field experiments. These methods are pragmatically feasible whenever plants could be sampled and harvested individually (e.g., in vegetable crops like tomato, brinjal, cauliflower and knol kol). The experimental data used in chapter 11 are for cereal crops of ragi and paddy. This emphasises the point that the method can be applied whenever it is feasible to take the observations on plant basis.
6.2 Notations

\[ y_i \] : Yield of the \( i^{th} \) plant.

\[ x_i \] : Indicator variable of group membership.

\( N \) : Total number of plants in the plot.

\( N_1 \) : Total number of healthy plants in the plot.

\( N_2 \) : Total number of diseased plants in the plot.

\( p \) : The mixing proportion.

\( f_1(y) \) : The density function of \( y \) for the healthy group of plants.

\( f_2(y) \) : The density function of \( y \) for the diseased group of plants.

\( f(x, y) \) : Bivariate mixture distribution of the two mixing distributions \( f_1(y) \) and \( f_2(y) \).

\( n \) : Sample size (fixed).

\( n_1 \) : Number of healthy plants in the sample (a random number).

\( n_2 \) : Number of diseased plants in the sample (a random number).

\( \phi(y|x_1, x_2) \) : The characteristic function of \( f(x, y) \).

6.3 The Probability Model

Consider a plot which is affected by a disease. Further assume that there are only two classes of plants: healthy and diseased. When a plant is chosen, two measurements are made on it:
(i) the yield \( y \) from the plant and (ii) the group membership \( x \). Here \( x \) is a Bernoulli variate taking values 0 (healthy) and 1 (diseased). Although one could conceive of different models for representing a mixture distribution, one which is most apt for estimation may be taken as

\[
f(x, y) = P^x(1-P)^{(1-x)} [x_1f_1(y_1) + (1-x_1)f_2(y_1)], \quad (6.1)
\]

where \( P \) is the mixing proportion, and \( f_1(y) \), \( f_2(y) \) are the probability functions of \( y \) for the healthy and diseased groups of plants respectively. It is easy to verify that (6.1) represents a probability function.

The model (6.1) for the mixture distribution depends on the assumptions we make about the distributions of \( y \) in the healthy and diseased groups of plants and also on the proportion '\( P \)' of mixture of healthy and diseased plants in the population. While section 6.4 deals with the mixture of two normal densities, section 6.5 treats the case of two exponential distributions. Although the methods developed here are general, specialisation to the above two cases is motivated by their scope in field situations (Chap. 11).
Assume that yield \((y_1)\) of healthy plants has a normal distribution with unknown mean \(\mu_1\) and variance \(\sigma_1^2\) and that of the diseased plants another normal distribution with mean \(\mu_2\) and variance \(\sigma_2^2\). Let these subpopulations be mixed in proportion \(P\). Thus Healthy : Diseased :: \(P : (1-P)\). Then in the model (6.1)

\[
f_1(y) = \frac{1}{\sigma_1 \sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(y_1 - \mu_1)^2}{\sigma_1^2}\right), \quad -\infty < \mu_1 < +\infty, \quad \sigma_1 > 0
\]

and

\[
f_2(y) = \frac{1}{\sigma_2 \sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(y_1 - \mu_2)^2}{\sigma_2^2}\right), \quad -\infty < \mu_2 < +\infty, \quad \sigma_2 > 0
\]

The following may be easily verified.

1) The marginal distribution of \(x\) is a Bernoulli distribution with probability mass function

\[
g(x) = P^x(1-P)^{(1-x)}, \quad x = 0, 1; \quad 0 < P < 1. \quad \text{(6.2)}
\]

ii) The marginal distribution of \(y\) is a mixture with density

\[
h(y) = P f_1(y) + (1-P) f_2(y) \quad \text{(6.3)}
\]

where the two densities \(f_1(y)\) and \(f_2(y)\) are mixed in the proportion \(P : (1-P)\).
Here \( h(y) \) is a univariate mixture density in \( y \).

While the bivariate mixture density \( f(x, y) \) requires input information on \( y \) and \( x \), \( h(y) \) requires information only on \( y \) for estimation of parameters of \( f_1(y) \), \( f_2(y) \) and \( P \). It is this marginal distribution which has received wide attention from many researchers as already mentioned. In fact Macdonald (1975) has shown the proliferation of the papers on finite mixtures by plotting the cumulative number of papers on a logarithmic scale against time.

iii) Conditional distributions: The conditional density of \( y \) for a given \( x \) may be easily shown to be

\[
h(y|x) = \frac{f(y, x)}{g(x)} = x f_1(y) + (1-x) f_2(y) \quad (6.4)
\]

The expected value of \( y \) can be shown to be

\[
E(y) = P\mu_1 + (1-P)\mu_2 = \mu \text{ (say)} \quad (6.5)
\]

and the variance of \( y \) to be

\[
V(y) = P\sigma_1^2 + (1-P)\sigma_2^2 + P(1-P)(\mu_1 - \mu_2)^2 \quad (6.6)
\]

6.4.1 Estimation of Parameters

The method of estimation of parameters of a mixture distribution depends upon the input information. The case when information available from the sample refers only to \( y \) with no information on \( x \) has been studied by several authors, notably by Hill (1963), Policello II
(1981), Tan et al (1972), Quandt and Ramsay (1978) besides those cited in section 6.1. But in the context of crop loss estimation, additional information on $x$ will also be available when selected plants are harvested individually for yield information. This should make the problem of estimation relatively simple. Let $(x_i, y_i)$, $i=1, 2, \ldots, n$ be a random sample of $n$ vector observations from the mixture model (6.1).

The log likelihood function is given by

$$
\ln L = \ln \prod_{i=1}^{n} \left[ P_i (1-P_i) \right]^{x_i} \left[ x_i f_1(y_i) + (1-x_i) f_2(y_i) \right],
$$

(6.7)

$$
= \ln P(\Sigma x_i) + \ln (1-P) \Sigma (1-M_i) + \Sigma \ln [x_i f_1(y_i) + (1-x_i)f_2(y_i)].
$$

(6.8)

Differentiating this w.r.t $\mu_1$, $\mu_2$, $\sigma_1^2$, $\sigma_2^2$ and $P$ and equating the derivatives to zero we get

$$
\frac{\partial \ln L}{\partial \mu_1} = \sum_{i=1}^{n} \left[ x_i (y_i - \mu_1) f_1(y_i) / [x_i f_1(y_i) + (1-x_i)f_2(y_i)] \right] = 0,
$$

(6.9)

$$
\frac{\partial \ln L}{\partial \mu_2} = \sum_{i=1}^{n} \left[ (1-x_i) (y_i - \mu_2) f_2(y_i) / [x_i f_1(y_i) + (1-x_i)f_2(y_i)] \right] = 0.
$$

(6.10)
\[
\frac{\delta \ln L}{\delta \overline{z}_1} = \frac{n}{1} \left[ x_1 \frac{\left( y_1 - \mu_1 \right)^2 - \sigma_1^2}{f_1(y_1)} - \frac{(1-x_1)f_2(y_1)}{f_1(y_1)} \right] = 0,
\]
(6.11)

\[
\frac{\delta \ln L}{\delta \overline{z}_2} = \frac{n}{1} \left[ (1-x_1) \frac{\left( y_1 - \mu_2 \right)^2 - \sigma_2^2}{f_2(y_1)} \right] - \frac{x_1 f_1(y_1) + (1-x_1)f_2(y_1)}{f_1(y_1)} = 0,
\]
(6.12)

and
\[
\frac{\delta \ln L}{\delta P} = \frac{\sum x_1}{P} - \frac{n(1-x_1)}{1-P} = 0.
\]
(6.13)

Though these equations seem formidable at first, they get simplified by noting that \( x_1 \) takes value either 0 or 1. The sample of size \( n \) splits at random into (say) \( n_1 \) observations in class 1 (healthy) and \( n_2 \) observations in class 2 (diseased), where \( n_1 + n_2 = n \). The equations (6.9) to (6.13) may be simplified as given below where \( \Sigma_j \) extends over class \( j \), \( (j = 1, 2) \) with \( n_j \) observations.

\[
\Sigma_1 \left[ (y_1 - \mu_1) \right] = 0
\]
\[
\Sigma_2 \left[ (y_1 - \mu_2) \right] = 0
\]
\[
\Sigma_1 \left[ (y_1 - \mu_1)^2 - \sigma_1^2 \right] = 0
\]
(6.14)

\[
\Sigma_2 \left[ (y_1 - \mu_2)^2 - \sigma_2^2 \right] = 0.
\]

These equations lead to the estimators
\[
\hat{\mu}_1 = \frac{\Sigma_1 y_1}{n_1} \quad \text{and} \quad \hat{\mu}_2 = \frac{\Sigma_2 y_1}{n_2}
\]
\[ \hat{\sigma}_1^2 = \frac{\sum(y_1 - \hat{\mu}_1)^2}{n_1}; \quad \hat{\sigma}_2^2 = \frac{\sum(y_1 - \hat{\mu}_2)^2}{n_2}. \]

and \[ \bar{\mu} = \frac{1}{n} \sum x_i \] (6.15)

These results demonstrate that \( \hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1^2, \hat{\sigma}_2^2 \) are the customary maximum likelihood estimators from the corresponding classes conditional on \( n_1 \) and \( n_2 \). In fact \( \hat{\mu}_1 \) is the sample mean from the first class and \( \hat{\mu}_2 \) is the sample mean from the second class.

**THEOREM 6.1**

\[ \hat{\mu} = \hat{P}\hat{\mu}_1 + (1-\hat{P})\hat{\mu}_2 \] is an unbiased estimator (6.16) of the population mean \( \mu = P\mu_1 + (1-P)\mu_2 \).

**Proof:**

\[ E(\hat{\mu}) = E[\hat{P}\hat{\mu}_1 + (1-\hat{P})\hat{\mu}_2], \]

\[ = E(\hat{P}\hat{\mu}_1) + E((1-\hat{P})\hat{\mu}_2). \] (6.17)

Consider \( E(\hat{P}\hat{\mu}_1) = \sum E(\hat{P}\hat{\mu}_1|\hat{P}) \),

\[ E \left[ \frac{n_1}{n} \frac{\sum y_i}{n_1} \right] \]

since taking expectation w.r.t. \( \hat{P} \) is the same as taking expectation w.r.t. \( n_1 \).

\[ E \left[ \frac{n_1}{n} \mu_1 \right] = \frac{\mu_1}{n} E(n_1) = P\mu_1. \] (6.18)
Similarly \( E(1-P)\hat{\mu}_2 = (1-P)\mu_2 \). \hspace{1cm} (6.19)

Thus from (6.17), \( E(\hat{\mu}) = P\mu_1 + (1-P)\mu_2 \). \hspace{1cm} (6.20)

Hence \( \hat{\mu} \) is an unbiased estimator of \( \mu \). Note that \( \frac{\sum y_i}{n} \) reduces to \( \frac{\sum y_i}{n} \) which is the sample mean \( \bar{y} \).

6.4.2 Characteristic Function

For knowing the precision of the estimate \( \hat{\mu} \) we need the expression for \( V(\hat{\mu}) \) which can be easily obtained through the characteristic function of (6.1) where \( f_1(y_1) \) and \( f_2(y_1) \) are the normal densities.

The characteristic function of \( (x, y) \) is given by

\[
\phi_{(x,y)}(t_1, t_2) = E[\exp(it_1x + it_2y)] , \text{ with } x=0,1 \text{ and } i=\sqrt{-1}
\]

\[
= \int_{-\infty}^{\infty} \left[ \int \left[ \exp(it_2y)(1-P)f_2(y) + \exp(it_1+it_2y)Pf_1(y) \right] dy \right]
\]

\[
(1-P)\left[ \exp(it_2\mu_2 - \frac{t_2^2\sigma_2^2}{2}) \right] + P\left[ \exp(it_1+it_2\mu_1 - \frac{t_2^2\sigma_1^2}{2}) \right]. \hspace{1cm} (6.21)
\]

From equation (6.21) we can get the characteristic function of \( (\bar{x}, \hat{\mu}) \) as

\[
\phi_{(\bar{x},\hat{\mu})}(t_1, t_2) = E[\exp(it_1\bar{x} + it_2\hat{\mu})],
\]
\[ \phi_\hat{\mu}(t_1, t_2) = \sum_{r=0}^{n} nC_r (1-P)^r p^{n-r} a^r b^{n-r}, \]

\[ = \sum_{r=0}^{n} nC_r (1-P)^r p^{n-r} \left[ \exp[\text{i} t_1 (n-r)/n] + it_2 M(r) - t_2^2 S^2(r)/2n^2 \right], \]

(6.22)

where \( M(r) = r \mu_2 + (n-r) \mu_1 \) and \( S^2(r) = r \sigma_2^2 + (n-r) \sigma_1^2 \). \hspace{1cm} (6.23)

The characteristic function of \( \hat{\mu} \) is obtained by putting \( t_1 = 0 \) in (6.22). This gives

\[ \phi_{\hat{\mu}}(t_2) = \sum_{r} nC_r (1-P)^r p^{n-r} \left[ \exp[\text{i} t_2 M(r)/n] - t_2^2 S^2(r)/2n^2 \right]. \]

(6.24)

Similarly the characteristic function of \( \bar{x} \) is

\[ \phi_{\bar{x}}(t_1) = \sum_{r} nC_r (1-P)^r p^{n-r} \left[ \exp[\text{i} t_1 (n-r)/n] \right], \]

\[ = [Q + P(\exp(\text{i} t_1))]^n. \] \hspace{1cm} (6.25)
THEOREM 6.2

Variance of \( \hat{\mu} \) is given by

\[
V(\hat{\mu}) = \frac{1}{n} \left[ \sigma_1^2 + Q\sigma_2^2 + PQ(\mu_1 - \mu_2)^2 \right] \frac{V(y)}{n}
\]  
(6.26)

Proof:

Consider the characteristic function of \( \hat{\mu} \) given by (6.24). This leads to

\[
(\hat{\mu})^{(t_2)} = \sum \frac{nC_r(1-P)^r}{t} \left( \frac{P}{n} \right)^{n-r} M(r)/n ,
\]

\[
\sum r(n-r) \frac{nC_r(1-P)^r}{t} \left( \frac{P}{n} \right)^{n-r} (n-r)\mu_1/n + \sum \frac{nC_r(1-P)^r}{t} \left( \frac{P}{n} \right)^{n-r} r\mu_2/n ,
\]

\[
\frac{\mu_1}{n} \sum r(n-r) nC_r(1-P)^r \left( \frac{P}{n} \right)^{n-r} + \frac{\mu_2}{n} \sum r nC_r(1-P)^r \left( \frac{P}{n} \right)^{n-r} ,
\]

\[
\frac{\mu_1}{n}(nP) + \frac{\mu_2}{n}(n(1-P)),
\]

= \( P\mu_1 + (1-P)\mu_2 \).

This shows that \( E(\hat{\mu}) = P\mu_1 + (1-P)\mu_2 \) as in (6.20).

For obtaining \( E(\hat{\mu}^2) \) we consider next

\[
\frac{\delta^2\phi}{\delta t^2} (\hat{\mu})^{(t_2)} = \sum \frac{n}{r} nC_r(1-P)^r \left( \frac{P}{n} \right)^{n-r} \left[ \frac{M^2(r)}{n^2} + \frac{S^2(r)}{n^2} \right] ,
\]
\[
\frac{1}{n^2}\left[ \frac{\Sigma C_x (1-P)^x P^{n-x} [(n-x)^2 \mu_1^2 + x^2 \mu_2^2 + 2(n-x) \mu_1 \mu_2 + (n-x) \sigma_1^2 + x \sigma_2^2]}{\Sigma x^2(1-P)^x P^{n-x}} = \frac{nPQ + n^2Q^2}{\Sigma x^3(1-P)^x P^{n-x}}.
\]

Noting \( \Sigma x^n(1-P)^x P^{n-x} = nPQ + n^2Q^2 \),

and \( \Sigma (n-x)^n(1-P)^x P^{n-x} = nPQ + n^2P^2 \),

(6.27) reduces to

\[
\frac{1}{n^2}\left[ (nPQ + n^2P^2) \mu_1^2 + (nPQ + n^2Q^2) \mu_2^2 + 2n(1-P)PQ \mu_1 \mu_2 + nP \sigma_1^2 + nQ \sigma_2^2 \right],
\]

which is the expression for \( E(\hat{\mu}^2) \). Therefore

\[
V(\hat{\mu}) = E(\hat{\mu}^2) - (E(\hat{\mu}))^2 = \frac{1}{n^2}\left[ (nPQ + n^2P^2) \mu_1^2 + (nPQ + n^2Q^2) \mu_2^2 + 2n(1-P)PQ \mu_1 \mu_2 + nP \sigma_1^2 + nQ \sigma_2^2 - n^2P^2 \mu_1^2 - n^2(1-P)^2 \mu_2^2 - 2nPQ \mu_1 \mu_2 \right],
\]

\[
= \frac{1}{n}\left[ PQ \sigma_1^2 + Q \sigma_2^2 + PQ(\mu_1 - \mu_2)^2 \right].
\]

which is the same as \( V(y)/n \).

6.4.3 Measurement of Crop Loss

Suppose there are \( N \) plants in a plot and \( N_1 \) of them are healthy while \( N - N_1 = N_2 \) are diseased. The proportion \( P \) of healthy plants is \( N_1/N \) and that of the diseased plants is \( N_2/N \). The average yield from the healthy plants is
denoted by $\mu_1$ while that from the diseased plants is denoted by $\mu_2$. The overall average yield realised is

$$\mu = P\mu_1 + (1-P)\mu_2.$$ 

Hence the total loss in the plot as defined in chapter 1 is

$$N_2(\mu_1-\mu_2) = N(\mu_1-\mu). \quad (6.30)$$

In order to construct estimators for the total loss we need to distinguish between the two cases:

(i) When $N_1$ and $N_2$ are known.

(ii) When these are not known (more realistic case)

**Theorem 6.3**

When $N_1$ and $N_2$ are not known $N(Q(\hat{\mu}_1-\hat{\mu}_2)) \quad (6.31)$

where $Q$, $\hat{\mu}_1$ and $\hat{\mu}_2$ are as defined earlier, is an unbiased estimator of the total crop loss with its variance given by

$$V[NQ(\hat{\mu}_1-\hat{\mu}_2)] = \frac{N^2}{n} \left[ \frac{\sigma_1^2}{n} \left[ nE\left(\frac{1}{n_1}\right) - 2n + E(n_1) \right] 
+ \frac{\sigma_2^2}{n} E(n_2) + PQ(\mu_1-\mu_2) \right]. \quad (6.32)$$

**Proof:**

$$E[NQ(\hat{\mu}_1-\hat{\mu}_2)] = NE(Q\hat{\mu}_1 - Q\hat{\mu}_2),$$
\[ E[NQ(\hat{\mu}_1 - \hat{\mu}_2)] = N\left[ E\left( \hat{Q}\mu_1 \mid n_2 \right) \right] - E\left( \hat{Q}\mu_2 \mid n_2 \right) = N\left[ E\left( \hat{Q}\mu_1 \right) - E\left( \hat{Q}\mu_2 \right) \right], \]

\[ NQ(\mu_1 - \mu_2), \]

\[ N_2(\mu_1 - \mu_2), \]

\[ N(\mu_1 - \mu). \]

Hence \( N[\hat{Q}(\hat{\mu}_1 - \hat{\mu}_2)] \) is an unbiased estimator of the total crop loss as in (6.30).

To evaluate the variance, consider

\[ V[\hat{Q}(\hat{\mu}_1 - \hat{\mu}_2)] = V(\hat{Q}\mu_1) + V(\hat{Q}\mu_2) - 2\text{Cov}(\hat{Q}\mu_1, \hat{Q}\mu_2). \quad (6.33) \]

Let us evaluate the right hand side term by term. First consider

\[ V(\hat{Q}\mu_1) = E\left[ V(\hat{Q}\mu_1 \mid n_2) \right] + V\left[ E(\hat{Q}\mu_1 \mid n_2) \right], \]

\[ = E\left[ \hat{Q}^2 \frac{\sigma^2}{n_1} \right] + V[\hat{Q}\mu_1], \]

\[ \frac{\sigma^2}{n_2} E\left( \frac{n_2}{n_1} \right) + \mu_n^2 V(\hat{Q}). \]

Since \( n_2 = n-n_1 \), the above expectation \( E \) may be be evaluated more easily by operating the expectation \( \hat{\mu}_1 \) via \( n_1 \).
Finally consider \( \text{Cov}[\hat{Q}_1, \hat{Q}_2] = \)

\[
= E \left[ \text{Cov}(\hat{Q}_1|n_2, \hat{Q}_2|n_2) \right] + \text{Cov} \left[ E(\hat{Q}_1|n_2), E(\hat{Q}_2|n_2) \right],
\]

\[
= E \left[ \hat{Q}^2 \text{ Cov}(\mu_1, \mu_2) \right] + \mu_1 \mu_2 \text{ Cov}(\hat{Q}, \hat{Q}) = \mu_1 \mu_2 \ V(\hat{Q}),
\]

\[
= \mu_1 \mu_2 \ \frac{PQ}{n}. \tag{6.36}
\]

Substituting (6.34), (6.35) and (6.36) in (6.33) we get

\[
V[\hat{Q}(\hat{\mu}_1 - \hat{\mu}_2)] = \frac{\sigma_1^2}{n_1} \left[ n^2 E(\frac{1}{n_1}) - 2n + E(n_1) \right] + \\
+ \frac{\mu_1^2}{n} + \frac{\sigma_2^2}{n_2} E(n_2) + \frac{\mu_2}{n} - 2\mu_1 \mu_2 \ \frac{PQ}{n},
\]
This can be evaluated by substituting the expression 
for \( E(1/n_1) \) from (4.9).

The estimated variance can be obtained by 
substituting the unbiased estimators of \( \sigma_1^2, \sigma_2^2, \mu_1, \mu_2 \) and 
P in the above expression for variance.

**Theorem 6.4**

When \( N_1 \) and \( N_2 \) are known \( N_2(\hat{\mu}_1-\hat{\mu}_2) \) (6.37) 
is an unbiased estimator of the total crop loss with its variance given by

\[
V[N_2(\hat{\mu}_1-\hat{\mu}_2)] = N_2^2 \left[ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \right].
\] (6.38)

**Proof:**

\[
E[N_2(\hat{\mu}_1-\hat{\mu}_2)] = N_2 E(\hat{\mu}_1-\hat{\mu}_2),
\]

\[
= N_2(\mu_1-\mu_2),
\]

\[
= N_2 \mu_1 - \mu_2, \mu_2,
\]

\[
= N(\mu_1-\mu).
\]
\[ \cdot \cdot \cdot N_2(\hat{\mu}_1 - \hat{\mu}_2) \] is an unbiased estimator of the total crop loss as in \((6.30)\).

Consider \( V(\hat{\mu}_1 - \hat{\mu}_2) = V(\hat{\mu}_1) + V(\hat{\mu}_2) - 2\text{Cov}(\hat{\mu}_1, \hat{\mu}_2) \),

\[ = \sigma_1^2 E\left(\frac{1}{n_1}\right) + \sigma_2^2 E\left(\frac{1}{n_2}\right), \]

since \( \text{Cov}(\hat{\mu}_1, \hat{\mu}_2) = 0 \).

Hence \( V[N_2(\hat{\mu}_1 - \hat{\mu}_2)] = N_2^{2} \left[ \sigma_1^2 E\left(\frac{1}{n_1}\right) + \sigma_2^2 E\left(\frac{1}{n_2}\right) \right] \).

This can be evaluated after substituting the expressions for \( E(1/n_1) \) and \( E(1/n_2) \), which are available in \((4.9)\).

6.5 Mixture of Exponential Densities

When plants are affected by a disease, the extent of damage caused is not generally uniform over the entire field. Usually a large percentage of plants will be affected to a lesser extent and a small percentage will have more severity. This kind of situation might result in a change of distributional form with a large number being concentrated at a relatively higher yield than at other levels. One could therefore expect an exponential type of distribution for plant yield in such a situation.
In fact in a field experiment carried out at the University of Agricultural Sciences, Bangalore and presented in chapter 11, it is found that the yield of crop in each category follows an exponential distribution. Crop loss estimators are developed below for this situation in the case of mixture of two exponential distributions.

**The Model**

A mixture model for two exponential distributions with parameters $\gamma_1$ and $\gamma_2$ respectively may be written as

$$f(x, \gamma_1) = \frac{x^i(1-P)}{(1-x_1)} [x_1\gamma_1 \exp(-\gamma_1x_1) + (1-x_1)\gamma_2 \exp(-\gamma_2x_1)] ,$$

$$x = 0,1 ;$$

$$0 \leq y < \infty$$

The characteristic function of (6.39) may be shown to be

$$\phi(x, t_1, t_2) = (1-P)(1 - \frac{it_2}{\gamma_2})^{-1} + Pe^{it_1}(1 - \frac{it_2}{\gamma_2})^{-1} .$$

Using this, the first two moments of $y$ are obtained as

$$E(y) = \frac{P}{\gamma_1} + \frac{(1-P)}{\gamma_2} ,$$

$$E(y^2) = 2(1-P)/\gamma_2^2 + 2P/\gamma_1^2 ,$$
and variance as

\[ V(\gamma) = \frac{P}{\gamma_1^2} + \frac{(1-P)}{\gamma_2^2} + PQ(1/\gamma_1 - 1/\gamma_2)^2. \]  

(6.43)

\( V(\gamma) \) has a structure similar to (6.6) with \( \sigma_1^2, \sigma_2^2, \mu_1 \) and \( \mu_2 \) replaced by their respective analogs for an exponential distribution.

### 6.5.1 Maximum Likelihood Estimators

Suppose the pairs \((x_i, y_i) ; i = 1, 2, 3, \ldots, n\) are a random sample from mixture population (6.39) then the likelihood of the sample is

\[ L = \prod_{i=1}^{n} P^{x_i}(1-P)^{(1-x_i)} \left[ x_i \gamma_1 e^{-\gamma_1 y_i} + (1-x_i) \gamma_2 e^{-\gamma_2 y_i} \right]. \]  

(6.44)

The estimator for \( P \) turns out to be the usual binomial estimator given by \[ \hat{P} = \frac{\sum x_i}{n}. \]  

(6.45)

As in section 6.4.1 the likelihood equations lead to the solution

\[ \hat{\gamma}_1 = \frac{n_1}{\sum_1 \gamma_1} \quad \text{and} \quad \hat{\gamma}_2 = \frac{n_2}{\sum_2 \gamma_1}. \]  

(6.46)

\( 1/\hat{\gamma}_1 \) and \( 1/\hat{\gamma}_2 \) are the sample means of the healthy and diseased groups respectively and

\[ \hat{\mu} = \frac{\hat{P}}{\hat{\gamma}_1} + \frac{(1-\hat{P})}{\hat{\gamma}_2}. \]  

(6.47)
is an unbiased estimator of $\mu$, the mean for the mixture of two distributions as defined in (6.41).

6.5.2 The Variance of $\hat{\mu}$

The characteristic function of $\hat{\mu}$ is seen to be
\[ \phi_{\hat{\mu}}(t) = \text{E}C_{x}(1-P)^{2}P(n-x)(1 - \frac{it_{2}}{n\gamma_{1}})^{r-n}(1 - \frac{it_{2}}{n\gamma_{2}})^{-r}. \] (6.48)

From this it follows that
\[ E(\hat{\mu}) = \frac{P}{\gamma_{1}} + \frac{(1-P)}{\gamma_{2}}, \] (6.49)

and
\[ V(\hat{\mu}) = \frac{1}{n}[ \frac{P}{\gamma_{1}^{2}} + \frac{(1-P)}{\gamma_{2}^{2}} + \frac{P(1-P)(1/\gamma_{1} - 1/\gamma_{2})^{2}}{\gamma_{1}^{2}} ] , \] (6.50)

which takes the usual form $V(\gamma)/n$. (6.51)

6.5.3 Measurement of Crop Loss

As in section (6.4.3), the estimators to be constructed depend upon whether $N_{1}$ and $N_{2}$ are known or not. Although usually $N_{1}$ and $N_{2}$ are not known in practice, the case when they are known is also presented here. (1) If $N_{1}$ and $N_{2}$ are not known $N[\hat{\gamma}(1/\hat{\gamma}_{1} - 1/\hat{\gamma}_{2})] \) (6.52) is an unbiased estimator of the total crop loss and its variance is
\[
\frac{N^2}{n} \left[ \frac{1}{n\gamma_1^2} \left( n^2E(1/n_1) - 2nE(n_2) \right) + \frac{1}{n\gamma_2^2} E(n_2) + PQ(1/\gamma_1 - 1/\gamma_2)^2 \right].
\] (6.53)

(ii) When \( N_1 \) and \( N_2 \) are known

\[
N_2(1/\gamma_1 - 1/\gamma_2)
\] (6.54)

is an unbiased estimator of the total crop loss and its variance can be shown to be (from theorem 6.4)

\[
N_2^2 \left[ \frac{1}{\gamma_1^2} E(1/n_1) + \frac{1}{\gamma_2^2} E(1/n_2) \right].
\] (6.55)

The approximate expression for \( E(1/n_1) \) and \( E(1/n_2) \) are available in (4.9).

As pointed out in section 6.5, the yield data for paddy crop was found to be exponential rather than normal. The above formulations are applied to these field data and results are reported in chapter 11.