# CHAPTER TWO
# BACKGROUND

## 2.1 Mixture Models For Zero Heavy Data

As mentioned in Chapter One, a subpopulation may solely be responsible for the extra zeroes. In addition, the nonzero portion of the data often fits very well to some existing distributions. Hence it would be appropriate to use a mixture distribution, which combines the two components, to model zero heavy data. Several authors have proposed different but similar models for the analysis of zero heavy data based on the use of a mixture distribution. These models are able to incorporate the excess zeroes in the data.

Aitchison (1955) proposes a mixture approach to model a positive random variable with a discrete probability of zero. He also develops the estimates of the mean and the variance of the mixture distribution. Aitchison's model is motivated by economic studies in which the data often consists of a large number of zeroes, and at the same time, the positive portion of the data often fits very well to some existing distributions. Aitchison and Brown (1957) consider a special case of Aitchison's model which they call the delta-lognormal distribution (a combination of a zero mass and a lognormal density). Crow and Shimizu (1988) extend the delta-lognormal to a multivariate case and Shimizu (1993) applies the bivariate case to model rainfall data. Pennington (1983) develops an estimator of the variance of the mean for the delta-lognormal distribution and applied it to estimate the abundance for fish and plankton in marine survey.

Lachenbruch (1976) proposes an Aitchison-like model for testing hypotheses in a cell transplant experiment. His model is a mixture of a degenerate distribution at zero and a positive distribution. This is a simple model that includes the excess zeroes and tests different model based hypotheses. This model is motivated by the data analysis of a cell transplantation experiment where cell rejection is very common. As a result, there are a large number of zeroes recorded due to the rejection.

Heilbron (1989) proposes the "zero altered" and "added zero" model for modeling high-risk behavior in homosexuals. The zero altered model is the same as Lachenbruch's model except the models are discussed in the notation of generalized linear model with covariates. The model he uses is a degenerated distribution at zero with either a positive Poisson or negative binomial distribution for the non-zero numbers. On the other hand, the zero added model uses a regular Poisson or negative binomial distribution which included its own zero values. Lambert (1992) also proposes a "zero-inflated" Poisson (ZIP) regression to model the number of defects in a manufacturing process. This model is actually the same as Heibron's "zero added" model.

The basic idea of a mixture distribution is to model the data Y by two variables X and Z where Y = X*Z. In general, X can follow any distribution of a non-negative random variable, for example, the Poisson, lognormal, or exponential distribution, or any distribution of non-negative random variables. Z is a binary variable which takes on the value 0 or 1 and contributes the extra zero to Y. With this concept and Heilbron's terminology, the models mentioned above can be divided into two categories. One is the zero altered model for which *Prob*(Y=0) = *Prob*(Z=0), and the other is the added zero model in which *Prob*(Y=0) = *Prob*(X=0) + *Prob*(Z=0). They are shown in Table 2.1. It can be seen clearly that the problem of mixing zeroes occurs only in the added zero model, and it does not exist in the zero altered model.

**Table 2.1.** Mixture Models

| Model (Y= X*Z) | Probability of zero observations |
|---|---|
| Zero Altered model | *Prob*(Y=0) = *Prob*(Z=0) |
| Added Zero Model | *Prob*(Y=0) = *Prob*(Z=0) + *Prob*(X=0) |

## 2.2 Zero Altered Model

The zero altered model is a mixture of a zero mass and an index distribution with the zero values truncated. The concept behind the zero altered model is that zeroes come from a source (modeled by the probability p) other than the underlying distribution, however, the zeroes and the underlying distribution may be related due to some unknown function. Since the zeroes come from only one source, the problem of mixing zeroes is not a concern. In the following sections, Lachenbruch's model is the first zero altered model being presented, in which the probability p is not related to the underlying distribution of X.

### 2.2.1 Analysis of Variance (ANOVA) Model

Lachenbruch (1976) proposes a simple model to include the excess zeroes and tests different hypotheses based on this model. The development of the model is motivated by the analysis of a data set obtained from a cell transplantation experiment conducted by J. Hulka. Cells are transplanted from one strain of mice to another, and the number of transplanted cells grown is counted after fifteen days. In some animals some growth occurs, however, in others, the transplant is completely rejected and no cells remain. As a result, there are a large number of zeroes recorded due to the rejection. Since there are

several types of donor and recipients, the researchers are interested in testing hypotheses about the distribution of the cells between different donors and recipients.

Traditionally, there are several methods used to carry out the test. The first method is to compare two populations using the t-test on the nonzero data. This ignores information from the excess zeroes as well as any other distributional features of the data that may make the test more sensitive. The second method is to use a nonparametric test, such as the Mann-Whitney test, on the complete data. It is sometimes recommended that the zeroes are eliminated in this situation, but again this ignores an important feature of the data. If zeroes are included, then there will be many ties and the assumption of a continuous underlying distribution will be violated. There are two major features in the data for this experiment: the zeroes due to total rejection and the positive values due to the successful transplant and growth of the cells. The assumption of a mixture density for the underlying distribution makes sense. Lachenbruch proposes the following model based on the mixture distribution idea that incorporates the excess zeroes.

The Model

Consider a random variable X with a density function as follows:

$$
\begin{aligned}
f(x) &= 0 && \text{for } x < 0, \\
f(x) &= (1 - p)\, h(x) && \text{for } x > 0, \text{ and} \\
Prob(X = 0) &= p && \text{with } 0 \leq p < 1,
\end{aligned}
$$

where h(x) is such that $\int_0^\infty h(x)dx = 1$. Then X is a random variable with possibly excess zeroes. In this case, h(x) can be a positive or truncated Poisson, exponential, or even a lognormal distribution that takes on only positive values. In other words, the zeroes of the

distribution f(x) depend only on p. This is similar to Heilbron's zero altered model which will be discussed in the next section.

The analysis is based on the likelihood function and the likelihood of a sample of size n is (assuming the first $n - n_0$ are nonzero)

$$L(x_1, \ldots, x_n) = p^{n_0} (1 - p)^{n-n_0} \prod_{i=1}^{n-n_0} h(x_i).$$

The maximum likelihood estimate of p is easily seen to be $\hat{p} = n_0/n$. Similarly we can use the $n - n_0$ nonzero x values to estimate parameters of h(x). This estimate can be obtained exactly as if no excess zeroes are present and only the positive observations are used. The estimate of p is independent of the estimate of the parameters of $h(x)$.

A variety of hypotheses may be tested. In the one sample case, we may be concerned with:

$$H_0: f(x) = f_0(x)$$

which is equivalent to

$$H_0: p = p_o, h(x) = h_0(x).$$

The test can be carried out using the likelihood ratio test with an approximate $\chi^2$ distribution.

Another possible hypothesis is:

$$H_0: h(x) = h_0(x) \quad \text{given} \quad x > 0.$$

In this case, a nonparametric test can be used, however, there are some difficulties if $h(x)$ is not symmetric. A likelihood ratio test can again be used. If the form of the distribution allows, an exact test can be carried out.

In the two-sample case, the hypotheses may be:

$$H_0: f_1(x) = f_2(x),$$

or assuming a common distributional form,

$$H_0' : h_1(x) = h_2(x), \text{ given } x > 0,$$
$$H_0'' : p_1 = p_2.$$

Again, the likelihood ratio test can be used to test $H_0$ or the subhypotheses, but this is affected by the choice of the functional form of $h(x)$. Lachenbruch further suggests a chi-square test which combines the Mann-Whitney test for nonzero data and the test of zero proportions in order to offer protection against a poor choice of $h(x)$. The analysis of the Lachenbruch model is quite simple and straightforward, however, this model does not involve the mixing of zeroes coming from two sources and the use of covariates. Another zero altered model which involves covariates is Heilbron's (1989) regression model for the analysis of count data.

### 2.2.2 Regression Model

Heilbron's (1989) regression model is based on the mixture of two distributions including a zero mass. This model is similar to Lachenbruch's model for hypotheses testing. However, the model is fitted using the generalized linear model, and the effect of overdispersion is also considered. The two-part model was first introduced which separately models the zero counts and the positive counts with the assumption that the two parts are independent of each other. Furthermore, Heilbron extends it to a zero altered model where the probability of zero depends on the covariates through the mean of

the distribution that gives rise to the positive counts. Before discussing the model in detail, some notation should be introduced. The notation is as follows:

$\mu$ and $\sigma^2$ are the mean and the variance of the mixture distribution; and

$i_0$ is the overdispersion index ($i_0 = \sigma^2/\mu - 1$).

The Models

Heilbron's first model, which made use of the Poisson-zero mixture distribution, is shown in the following:

**1. Zero-Altered Poisson (ZAP)**

Consider the random variable X with

$$Prob(X < 0) = 0,$$

$$Prob(X = 0) = p,$$

$$Prob(X = k) = (1-p)\frac{e^{-\lambda}\lambda^{k}}{(1-e^{-\lambda})k!},$$

where $0 \le p \le 1$, and $\lambda > 0$ for $k > 0$ with $\mu = (1 - p)\,\lambda/(1 - e^{-\lambda})$, $\sigma^2 = \mu(\lambda - \mu + 1)$ and the overdispersion index $i_0 = \sigma^2/\mu - 1 = (\mu(\lambda - \mu + 1)/\mu) - 1 = \lambda - \mu$.

This model is denoted as ZAP[$\lambda$, p] and apart from (1 – p), the expression for *Prob*(X = k) is just the probability for a Poisson distribution with zero truncated. It is called the positive Poisson or the truncated Poisson. The parameters of the Poisson and the truncated Poisson distribution are shown in Table 2.2. This model is actually Lachenbruch's model with h(x) as the Poisson distribution discussed previously, however, in this ZAP model, the probability of zero, p, may be related to the parameter of the Poisson distribution $\lambda$ by a function f such that $p = f(\lambda)$. The function $p = e^{-\gamma\lambda}$ is a sensible choice for $f(\lambda)$ since the probability of X = 0 approaches zero as $\lambda \to \infty$.

**Table 2.2.** The parameters of the Poisson and the truncated Poisson distributions.

| Distribution | Poisson | Truncated/Positive Poisson |
|---|---|---|
| Density | $P(X = x) = \frac{e^{-l} l^{x}}{x!}$ $x = 0, 1, 2, \ldots$ | $P(X = x) = \frac{e^{-l} l^{x}}{(1-e^{-l})x!}$ $x = 1, 2, 3, \ldots$ |
| Mean | $l$ | $l / (1 - e^{-l})$ |
| Variance | $l$ | $[l / (1 - e^{-l})][1 - l\ e^{-l} - e^{-l}/(1 - e^{-l})]$ |

The above model is denoted as ZAP[$\lambda$, $e^{-\gamma\lambda}$] with $\mu = \lambda(1 - e^{-\gamma\lambda})/(1 - e^{-\lambda})$ and $i_0 = \lambda\ (\ e^{-\gamma\lambda} - e^{-\lambda})/(1 - e^{-\lambda})$. $i_0$ approaches zero as $\lambda \rightarrow \infty$ for all $\gamma > 0$. The parameter $\gamma$ acts like a shape parameter for the distribution. If $\gamma$ approaches zero, the distribution approaches zero mass. If $0 < \gamma < 1$, the distribution is overdispersed with $\sigma^2 > \mu$ and $\lambda > \mu$. If $\gamma > 1$, the distribution is underdispersed with $\sigma^2 < \mu$ and $\lambda < \mu$. As $\gamma$ approaches infinity, the distribution approaches the positive Poisson ($\lambda$).

### 2. Zero-Altered Negative Binomial (ZANB)

Heilbron also uses the negative binomial distribution in his zero altered regression model to handle the overdispersion effect. Two different parameterization of the negative binomial distribution are discussed in Heilbron's work as shown in Table 2.2, but only one is presented here. The negative binomial used in the following zero altered model is analogous to that used by Lawless (1987) and is denoted as NB1($\lambda$, $\alpha$) with $\alpha > 0$. It is as follows:

$$\text{Prob}(X = k) = \frac{\Gamma(k + a^{-1})}{k!\Gamma(a^{-1})}\left(\frac{al}{1+al}\right)^{k}(1+al)^{-1/a}$$

where $\mu = \lambda$ and $\sigma^2 = \lambda + \alpha\lambda^2$. This distribution can be treated as an overdispersed Poisson with $i_0$ equal to $\alpha\lambda$. Here, NB1($\lambda$, $\alpha$) approaches Poisson($\lambda$) as $\alpha$ approaches zero.

Consider a random variable X with

Prob(X = 0) = $p$,

$$\text{Prob}(X = k) = \frac{(1-p)}{(1-(1+\alpha\lambda))^{-1/\alpha}} \frac{\Gamma(k+\alpha^{-1})}{k!\Gamma(\alpha^{-1})} \left(\frac{\alpha\lambda}{1+\alpha\lambda}\right)^k (1+\alpha\lambda)^{-1/\alpha}, \quad \text{for } k > 0$$

and $m = \frac{(1-p)}{1-(1+\alpha\lambda)^{-1/\alpha}}\lambda$, $s^2 = m\lambda[\lambda(1+\alpha) - m + 1]$.

The function $f$ is chosen to be $p = (1 + \gamma\alpha\lambda)^{-1/\alpha}$ with the property that the probability of X = 0 approaches zero as $\lambda \to \infty$. This model is denoted as ZANB1[$\lambda$, $\alpha$, $(1+\gamma\alpha\lambda)^{-1/\alpha}$] which approaches ZAP[$\lambda$, $e^{-\gamma\lambda}$] as $\alpha \to 0$.

These models are fitted using generalized linear models. Heilbron proposes two different links that equate the linear predictors to the means of the overall distribution and the mean of the index distribution respectively. He then uses the quasi-Newton method to obtain the maximum likelihood estimators of the coefficients.

**2.3 Added Zero Model**

The added zero model is a mixture of zero mass and an index distribution. It is very similar to the zero altered model except the zero of the index distribution is not truncated. The conceptual difference between the zero altered model and the added zero model is that the zeroes of the added zero model not only come from one source, but they also come directly from the underlying distribution. Hence, by just looking at the observations, one would not be able to distinguish whether a zero observation actually comes from the extra source or comes from the underlying distribution. From a

theoretical standpoint, if the index distribution is a discrete distribution with a probability of (Y=0) >0, we will have the problem of mixing zeroes. For continuous distributions, since the probability of (Y=0) = 0, the problem of mixed zeroes can be avoided. This may make the added zero model more interesting when the index distribution is a discrete distribution with zeroes then when the index distribution is a continuous one. However, from a practical standpoint, the problem of mixing zeroes is always possible.

Since there is not much work done in the ANOVA model, I will go directly into the regression models. Heilbron's (1989) Poisson with zero (PWZ) model is actually the same as Lambert's (1992) zero-inflated Poisson (ZIP) regression and both are considered to be added zero models. Both models use a mixture of Poisson and a zero mass, but only ZIP is presented in the following section.

### 2.3.1 Zero-inflated Poisson regression

Lambert (1992) proposes the zero-inflated Poisson regression model for the analysis of defective counts on soldering boards. The formulation of the model consists of a Poisson regression part and a logistic regression part. The link notation in the generalized linear model is also used to relate the parameters of the distribution to the covariates of the model.

The Model

In ZIP regression, the responses $\underline{Y} = (Y_1,\ldots,Y_n)'$ are independent with

$$Prob\ (Y_i = 0) = p_i + (1 - p_i)e^{-\lambda_i}$$

$$Prob\ (Y_i = k) = (1 - p_i)e^{-\lambda_i}\lambda_i^k / k!, \text{ where } k > 0.$$

The parameters $\underline{\lambda} = (\lambda_1, \ldots, \lambda_n)'$ and $\underline{p} = (p_1, \ldots, p_n)'$ satisfy

$$\log(\underline{\lambda}) = \mathbf{B}\underline{\beta} \quad \text{and}$$

$$\text{logit}(\underline{p}) = \log(\underline{p}/(1 - \underline{p})) = \mathbf{G}\underline{\gamma}$$

for the covariate matrices **B** and **G**. The log and logit function act exactly like the link function in the generalized linear model which "link" the parameters $\lambda$ and p to the covariates **B** and **G**.

The covariates that affect the Poisson mean may or may not be the same as the covariates that affect the probability p. When the covariates are the same and $\underline{\lambda}$ and $\underline{p}$ are not functionally related, **B** = **G** and the ZIP regression requires twice as many parameters as Poisson regression. At the other extreme, when the probability p does not depend on the covariates, **G** is a column of ones, and ZIP regression requires only one more parameter than Poisson regression. The number of parameters that can be estimated depends on the richness of the data. If there are only a few positive counts and $\underline{\lambda}$ and $\underline{p}$ are not functionally related, then only simple models should be considered for $\underline{\lambda}$. Since the zeroes of this model may either come from the Poisson distribution or from the subpopulation of zeroes, the model needs to distinguish between the two different kind of zeroes. By assuming an extra variable $Z_i$, where $Z_i = 0$ if the observation $Y_i$ is from the Poisson part and $Z_i = 1$ otherwise, the likelihood function of the model can easily be broken down into several parts. An E-M algorithm can then be used to obtain the maximum likelihood estimators for the coefficients of the ZIP model treating the $Z_i$'s as missing values. The details of the E-M algorithm will be presented later in the Chapter of toxicity testing.

In the case where $\underline{p}$ and $\underline{\lambda}$ are related, Lambert proposed the parameterization $\log(\underline{\lambda}) = \mathbf{B}\beta$ and $\text{logit}(\underline{p}) = -\tau\mathbf{B}\beta$, with an unknown, real-valued shape parameter $\tau$ which implies that $p_i = (1 + \lambda_i^{\tau})^{-1}$. Heilbron used the function $p_i = e^{-\tau\lambda_i}$ to relate p and $\lambda$ in his zero-altered Poisson regression model discussed previously.

In the terminology of generalized linear models, log($\underline{\lambda}$) and logit($\underline{p}$) are the natural links or transformations that linearize the Poisson means and the Bernoulli probability of success. The ZIP model with logit link for $\underline{p}$, log link for $\underline{\lambda}$, and shape parameter $\tau$ will be denoted by ZIP($\tau$). A Newton Raphson algorithm is used to estimate the coefficient for the ZIP($\tau$) regression model which is computationally more complex than the E-M algorithm.

**2.4 Summary**

The models discussed in section 2.2 and 2.3 are able to incorporate the extra zeroes in the data and reflect two possible mechanisms that produce extra zeroes: overdispersion and a subpopulation of zeroes. A summary of the reviewed models is listed in Table 2.3. The major difference between the two groups of models is the problem of "mixing" zeroes. The choice of models is really depending on the nature of the problems. Due to the nature of the problems, the three models mentioned above all use discrete distributions in their models, such as Poisson and negative binomial. However, these models can be applied to continuous cases with non-negative data without any conceptual difference. For example, Feuerverger (1979) proposes a mixture to include zero mass and a gamma distribution to model rainfall data. Aitchison (1955)'s delta-distribution model is a mixture of zero mass and a lognormal distribution.

**Table 2.3.** Summary of the reviewed models

| Model (Y= XZ) | Zero Altered Model | Added Zero Model | Prob(Z=0) related to the mean of X |
|---|---|---|---|
| Lachenbruch's model | √ | | |
| Heilbron's ZAP and ZANB | √ | | √ |
| Heilbron's Poisson with zero | | √ | √ |
| Lambert's ZIP | | √ | |
| Lambert's ZIP(τ) | | √ | √ |

## 2.5 Example: Multi-population Lachenbruch model with lognormal density

Since extra zeroes can be a very common problem in data analysis, the authors mentioned in the previous section also presented many applications and examples along with the mixture models that they propose. These applications and examples not only provide practical tools in data analysis, they also inspire some new methodologies for analyzing zero heavy data. In the following section, a multi-population extension of Lachenbruch's hypotheses testing model is presented. In general, hypotheses testing problems can be treated as model reduction problems with the null hypothesis as the reduced model and the alternative hypothesis as the full model. In terms of Lachenbruch's model, the hypotheses can be extended to a multi-population comparison with

$H_0$: $f_1(x) = f_2(x) = \ldots = f_n(x)$ or

$H_0'$: $h_1(x) = h_2(x) = \ldots = h_n(x)$ , given $x > 0$ and

$H_0''$: $p_1 = p_2 = \ldots = p_n$ .

Instead of only doing the one sample or two samples tests, an ANOVA-like analysis can be carried out with the likelihood ratio test. The data (shown in Appendix A) used in this example is taken from Koopmans (1981) p.107. It is from a biological study of the seasonal activity patterns of a species of field mice and consists of the average distances traveled between captures by those mice at least twice in a given month. The distances are rounded to the nearest meter. A simple one way ANOVA analysis can easily be done to find out if differences exist in the mean distance between the four seasons. However, from the boxplot (Figure 2.1), it is very obvious that fall and winter data have a large proportion of zeroes which skew the distribution towards zero. This definitely violates the normality assumption, and as a result, the normal ANOVA test would not be appropriate here. Instead, a likelihood ratio test based on Lachenbruch's hypotheses testing model can be used. Since the data consists of distances traveled by the mice, the assumption of a lognormal density for nonzero data would be appropriate. Hence,

$$h(x) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left(-\frac{1}{2\sigma^2}(\ln x - \mu)^2\right).$$

Along with the lognormal density, a probability p is used to model the zeroes. This lognormal mixture model is actually the delta distribution model of Aitchison and Brown (1957). The maximum likelihood estimators of the parameters are based on

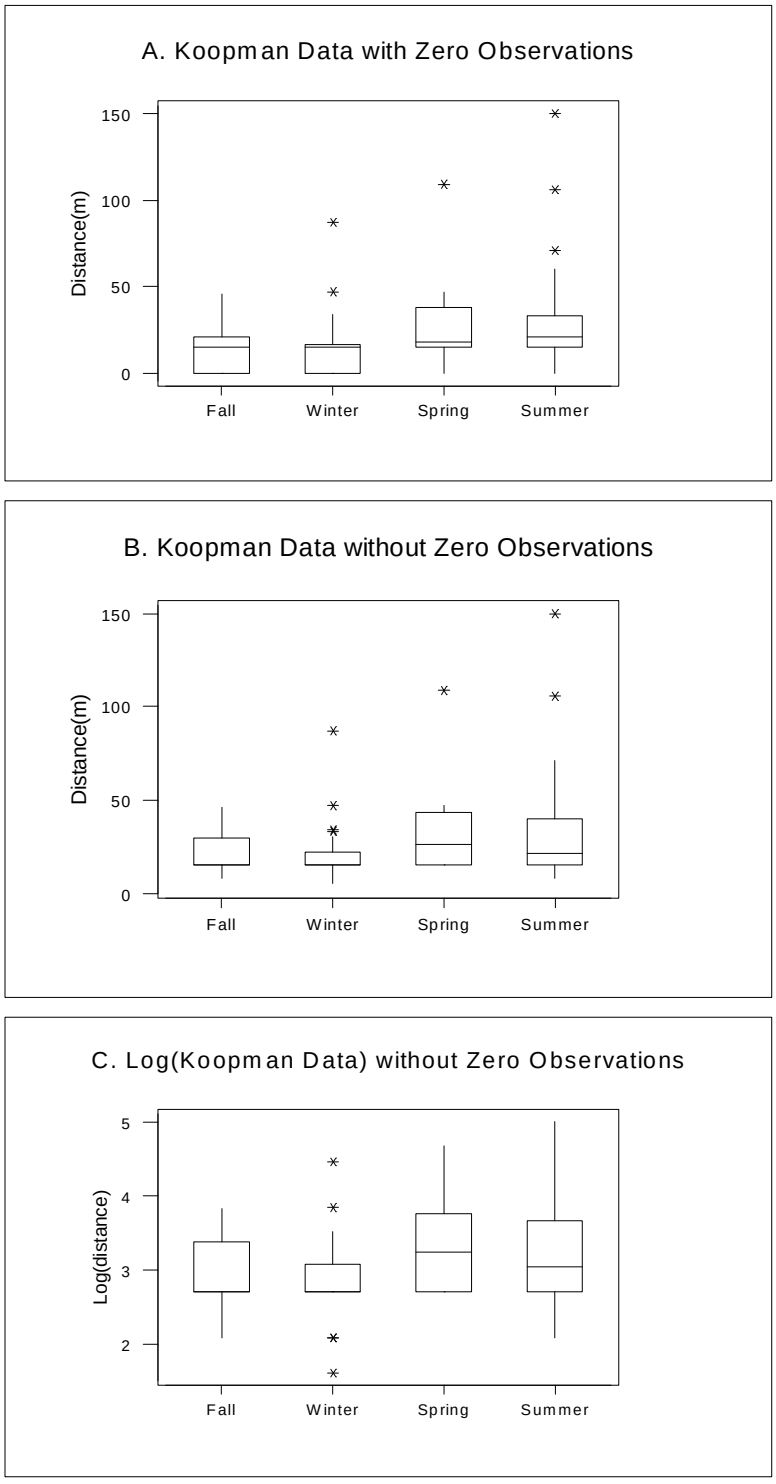


**Figure 2.1.** Boxplots of Koopman Data

Aitchison and Brown and they are shown in Table 2.4. The $\hat{\mu}$'s and the $\hat{\sigma}$'s are just the sample means and the sample variances of the logarithm of the nonzero data.

**Table 2.4.** Estimated Parameters of Koopmans' Data

| Seasons | n | n-$n_0$ | #zero($n_0$) | $\hat{\mu}$ | $\hat{\sigma}$ | $\hat{p}$(= $n_0$/n) |
|---|---|---|---|---|---|---|
| Fall | 27 | 17 | 10 | 2.91 | 0.513 | 0.370 |
| Winter | 34 | 24 | 10 | 2.87 | 0.586 | 0.294 |
| Spring | 17 | 14 | 3 | 3.29 | 0.608 | 0.177 |
| Summer | 27 | 24 | 3 | 3.25 | 0.731 | 0.111 |
| All | 105 | 79 | 26 | 3.07 | 0.641 | 0.248 |

The likelihood ratio can be calculated as follows:

Under the null hypothesis, there are three parameters. A probability p for the zero data, and an overall mean $\mu$ and variance $\sigma^2$ for the nonzero data. Hence, the log likelihood is

$$\ln L_0 = n_0 \ln \hat{p} + (n-n_0) \ln (1-\hat{p}) - (n-n_0)\ln \hat{\sigma} - \sum_{j=1}^{n-n_0} \{\ln x_j + \frac{1}{2}(\ln x_j - \hat{\mu})^2 / \hat{\sigma}^2\}$$

$$= -304.98.$$

Under the alternative hypothesis, there are 12 parameters $\mu_i$, $\sigma_i$, $p_i$, $i = 1,2,3,4$. A mean and variance for the nonzero data for each season, and a probability for the zero part for each season. Hence the loglikelihood is:

$$\ln L = \sum_{i=1}^{4} \ln L_i$$

$$= \sum_{i=1}^{4} \{n_{0i}\ln \hat{p}_i + (n_i - n_{0i}) \ln (1 - \hat{p}_i) - (n_i - n_{0i})\ln \hat{\sigma}_i - \sum_{j=1}^{n_i - n_{0i}} [\ln x_j + \frac{1}{2}(\ln x_j - \hat{\mu}_i)^2 / \sigma_i{}^2]\}$$

$$= -295.81.$$

Two times the log of the likelihood ratio ($H_A/H_0$) gives an asymptotic chi-square distribution with k degrees of freedom, where k is the difference in the number of parameters estimated in the two hypotheses. In this example, k = 12 – 3 = 9 and the likelihood ratio gives a chi-square with 9 degrees of freedom. The test statistic is

$\chi_9^2 = 2(-295.81 - (-304.98)) = 2(9.17) = 18.34.$

The p-value is 0.03, and hence, indicates that the zero proportion, variance and mean of distances traveled by the field mouse are in fact significantly different between the four seasons. As noticed from the boxplot of the complete data and the non-zero data in Figure 2.1, the difference mainly comes from the zeroes part of the data. The different hypotheses testing procedures on the three different parameters of the likelihood can also be carried out. The results are presented in Table 2.5. The null hypothesis ($H_0$) and the alternative hypotheses ($H_A$) are shown in the first two columns. In order to ease the calculation, the log likelihood is broken up into 3 components, as shown in the table, i.e. $n_0 \log p + (n-n_0)\log (1-p)$, $(n-n_0)\log\sigma$, and $\Sigma \log x_i + (\log x_i - \mu)^2/2\sigma^2$. These values are calculated correspond to the alternative hypotheses ,where p, $\mu$, and $\sigma$ can be $p_i$, $\mu_i$, and $\sigma_i$ depending on the different hypotheses. The column of log likelihood is the calculated log likelihood of each model based on the alternative hypothesis using the equation:

$$\ln L_0 = n_0 \ln \hat{p} + (n-n_0) \ln (1 - \hat{p}) - (n-n_0)\ln \hat{\sigma} - \sum_{j=1}^{n-n_0} \{\ln x_j + \frac{1}{2}(\ln x_j - \hat{\mu})^2 / \hat{\sigma}^2\}$$

The Chi-squared values are obtained by two times the difference of log likelihood with the degree of freedom shown in (). Finally the p-values are presented in the last column. The

test of different mean, variance, and zero proportion was shown significantly different previously with a p-value of 0.03. The test of different zero proportion and the test of different mean are both significant with p-values 0.04 and 0.07 respectively. However, the hypothesis of testing different variance is highly insignificant with p-value 0.5. This further confirms what was observed in boxplot, that the difference comes from the zeroes.

**Table 2.5**. Results of the different hypotheses testing on different combination of parameters based on the likelihood ratio test for the Koopmans data

| $H_0$ | $H_A$ (# of parameters) | $n_0 \log p + (n-n_0)\log(1-p)$ | $(n-n_0)\log\sigma$ | $\Sigma \log x_i + (\log x_i-\mu)^2/2\sigma^2$ | log likelihood | Chi-squared (df) | p-values |
|---|---|---|---|---|---|---|---|
| $p, \mu, \sigma$ (3) | ------ | -58.77 | -35.10 | 281.31 | -304.98 | ------ | ------ |
| | $p_i, \mu_i, \sigma_i$ (12) | -54.67 | -38.65 | 279.81 | -295.83 | 18.3 (9) | 0.0318* |
| | $p, \mu_i, \sigma_i$ (9) | -58.77 | -38.65 | 279.81 | -299.93 | 10.1 (6) | 0.1205 |
| | $p, \mu, \sigma_i$ (6) | -58.77 | -38.65 | 283.70 | -303.82 | 2.32 (3) | 0.5087 |
| | $p, \mu_i, \sigma$ (6) | -58.77 | -35.10 | 277.83 | -301.50 | 6.96 (3) | $0.0732^+$ |
| | $p_i, \mu, \sigma$ (6) | -54.67 | -35.10 | 281.31 | -300.88 | 8.20 (3) | 0.0421* |
| | $p_i, \mu_i, \sigma$ (9) | -54.67 | -35.10 | 277.83 | -297.40 | 15.16 (6) | 0.0190* |
| | $p_i, \mu, \sigma_i$ (9) | -54.67 | -38.65 | 283.70 | -299.72 | 10.52 (6) | 0.1044 |

* statistical significant at 0.05 level, $^+$statistical significant at 0.1 level