ALLOCATION PROBLEM IN UNDER-REPORTED COUNT DATA

NAME: RAKA SEN

ROLL NUMBER: 0445

REGISTRATION NUMBER: A01-2112-0760-17

SUPERVISOR: PROF. DR. SURUPA CHAKRABORTY

DECLARATION- I affirm that I have identified all my sources and that no part of my dissertation paper uses unacknowledged materials.

SIGNATURE: Raka Sen

TABLE OF CONTENTS

ABSTRACT

Statistical analyses involving count data may take several forms depending on the context of use, for example, simple counts, such as the number of occurrences of thunderstorms in a calendar year, number of arrivals in any service facility like ATM, bank, etc., number of suicides reported in a large population and categorical data in which the counts represent the numbers of items falling into each of several categories. The mostly adapted model for analysing count data is the Poisson Model.

Any counting system is prone to recording errors including under-reporting and over-reporting. Ignoring the misreporting pattern in count data can give rise to bias in the estimation of model parameters. The work here provides methods for correcting this bias. My work primarily deals with under-reported count data. Under-reporting occurs in survey data when there is a reason for participants to give a false negative response to a question, often due to a perceived social stigma, for example, maternal smoking in epidemiological studies. Any reporting or counting system is prone to such errors in recording. The reasons may be quite different in various fields of application like public health, criminology, actuarial science or production.

For the purpose of the study, the true count is assumed to follow Poisson distribution with mean λ. However, modelling underreported count data using a Poisson model poses some difficulties, particularly in estimation of the model parameter λ. When the data has not been contaminated by any reporting errors, estimate of the model parameter λ is the sample mean but in the presence of reporting errors, the sample mean is a biased estimate of λ.

Thus, we propose extension of the standard Poisson model so that under-reporting patterns can be captured. Also, a key assumption that is made for the under-reporting model is that the

reporting probability is constant and identical for all circumstances. However, in practise the reporting probability might change under different circumstances.

So, to summarize, the objective of this dissertation is three-fold:

a) Correct estimation of the true model parameter λ, based on a sample of size *n* where the error-prone data is observed and a subsample of size *r* where the true count is observed.

b) Optimum allocation of main study sample of size *n* and internal validation sample of size *r,* optimization in terms of increasing the precision of the corrected estimator while keeping the cost of survey at a fixed level.

c) Investigating the problem of allocation of resources between validation and non-validation units.

CHAPTER 1: INTRODUCTION

1.1: COUNT DATA:

In statistics, count data is a statistical data type, a type of data in which the observations can take only the non-negative integer values {0, 1, 2, 3,…} and where these integers arise from counting rather than ranking. The statistical treatment of count data is distinct from that of binary data, in which the observations can take only two values, usually represented by 0 and 1, and from ordinal data, which may also consist of integers but where the individual values fall on an arbitrary scale and only the relative ranking is important. Some examples of count data are illustrated below:

- Consumer demand: the number of products that a consumer buys on e-commerce platform like Amazon
- Recreational data: the number of trips taken per year
- Family economics: the number of children a couple has
- Health demand: the number of doctors' visits

1.2: THEORETICAL DISTRIBUTIONS FOR MODELLING COUNT DATA:

Count data is encountered on a daily basis and dealings. More understanding of such data and extraction of important information about the data needs some statistical analysis or modelling. Statistical analyses involving count data may take several forms depending on the context of use, that is, simple counts such as the number of plants in a particular field and categorical data in which counts represent the number of items falling in each of the several categories. Different count data may possess different characteristics and therefore cannot be used with a particular count data model. The most adapted model for analysing count data is the **Poisson model**[1]. Some of the alternative models that can be considered for modelling count data are **Negative Binomial**[2], **Hurdle or two-part models**[3] and **Zero-inflated models**[4]. However it is well known that the cases with small integral-valued observed data tend to fit the Poisson distribution. Poisson distribution is applied in situations where there are a large number of independent Bernoulli trials[5] with a very small probability of success in each trial. Thus, very commonly encountered situations of Poisson distribution are:

- The number of arrivals in a car wash in one hour
- The number of file server virus infection at a data centre during a 24-hour period
- The number of asthma patient arrivals in a given hour at a walk-in clinic
- The number of birth, deaths, marriages, divorces, suicides and homicides over a given period of time
- The number of visitors to a website per minute
- The number of calls to a consumer hot line in a 5-minute period

The above are only some typical examples of Poisson distribution. They are in no way exhaustive.

1.3: REPORTING ERRORS IN COUNT DATA:

Any counting system is prone to recording errors including under-reporting and over-reporting. Failing to correct this misreporting introduces biases and it may lead to misinformed decision making.

1.3.1: OVER-REPORTING IN COUNT DATA:

Over-reporting in registration systems occurs when the reported number of events is higher than the actual counts. Depending on the field of application, various factors might play a role in over-reporting of an event. In public health, a physician's unwitting mistakes in the diagnostic process could result in over-reporting of a specific disease. Two different explanations of over-reporting have been tendered with regard to survey responses. One explanation considers inaccurate memory function or recall errors and the second is social desirability which has been claimed to be the main cause of inflated self-reports. In general, research participants want to respond in a way that makes them look as good as possible. Thus, they tend to under-report behaviours deemed inappropriate by researchers or other observers and they tend to over-report behaviours viewed as appropriate.

1.3.2: UNDER-REPORTING IN COUNT DATA:

Underreporting is a problem in the collection that occurs when the counting of some event is for some reason incomplete; particularly occurring in survey data when there is a reason for participants to give a false negative response to a question, often due to a perceived social stigma. A famous example is maternal smoking during pregnancy, which is a key risk factor for adverse offspring outcomes including preterm birth and low birth weight (LBW). Any reporting or counting system is prone to such errors in recording. The reasons may be quite different in various fields of application like public health, criminology, actuarial science or production.in public health we have reporting systems for infectious diseases like HIV or chronic diseases like diabetes in which recording failures may occur as result of diagnostic errors or patients avoiding diagnosis. An example from industrial production is the number of products that are broken within a certain period, typically the warranty period. This number is important for quality management. Only the number of returned products is known, but the true total number includes also those goods that are not returned by customers. In all these cases, reporting systems give lower counts than the actual number of events.

My work here primarily deals with underreported count data.

1.4: PROBLEM ANALYSIS:

Modelling underreported count data using a Poisson model poses some difficulties, particularly in estimation of the model parameter λ. Ignoring the misreporting pattern (undercount) and carrying out naïve analysis results in biased inferences.

So, for the under-reported data, an adjusted model is set-up and various aspects of it are studied.

CHAPTER 2: METHODOLOGY

2.1: MODEL AND LIKELIHOOD:

To start with, let Y* denote the count variable of interest, which is assumed to follow a Poisson distribution with mean λ. The probability mass function of Y* is given by,

$$f(y *; \lambda) = e^{\lambda} \frac{\lambda^{y*}}{y*!} \text{ ; y*=0, 1, 2, ….. } \quad \text{---- (i)}$$

However, in many practical situations, the count data available are error-prone due to misreporting, faulty method of data collection or imperfect diagnosis. When the data has not been contaminated with any kind of reporting error, the sample mean ($\overline{Y *}$) is the Minimum Variance Unbiased Estimator (MVUE [6]) of λ. Nevertheless, when contaminated with error prone data (underreported data in this particular study), $\overline{Y *}$ is biased and can no longer be used as a viable unbiased estimator of the model parameter λ. We thus denote the observed count by Y and treat it as surrogate for true Y*.

Here, Y<=Y*, with probability 1.

Let (1-π) denote the probability of missing a count. Thus, conditional on the true count, the distribution of the surrogate is given by,

$$f(y|y *, \pi) = \binom{y*}{y} \pi^{y}(1-\pi)^{y*-y} \text{ ; y=0, 1, 2 …y*. } \quad \text{--- (ii)}$$

That is, $(Y|Y *= y *) \sim Binomial(y *, \pi)$. []

The marginal distribution of the surrogate Y follows from (i) and (ii) and is given by,

$$f(y;\lambda,\pi) = \sum_{y*=y}^{\infty} f(y|y*)f(y*)$$

$$= \sum_{y*=y}^{\infty} \binom{y*}{y} \pi^{y}(1-\pi)^{y*-y}\, e^{-\lambda}\frac{\lambda^{y*}}{y*!}$$

$$= \sum_{y*=y}^{\infty} \frac{\pi^{y}(1-\pi)^{y*-y}\, e^{-\lambda\pi}e^{-\lambda(1-\pi)}\lambda^{y*-y}\lambda^{y}}{y!\,(y*-y)!}$$

$$= \frac{(\lambda\pi)^{y}}{y!}e^{-\lambda\pi}\sum_{y*=y}^{\infty}\frac{\{(1-\pi)\lambda\}^{y*-y}\, e^{-\lambda(1-\pi)}}{(y*-y)!}$$

$$= e^{-\lambda\pi}\frac{(\lambda\pi)^{y}}{y!}\sum_{z=0}^{\infty}\frac{\{(1-\pi)\lambda\}^{z}\, e^{-\lambda(1-\pi)}}{z!}$$

$$= e^{-\lambda\pi}\frac{(\lambda\pi)^{y}}{y!} \qquad \text{, where } Z{\sim}Poisson((1-\pi)\lambda) \qquad \text{---- (iii)}$$

From (iii), it follows that the surrogate Y is also distributed as Poisson but with a modified mean given by λπ, that is, Y~ Poisson (λπ). In case π is known, the maximum likelihood estimator (MLE[7]) of λ can simply be obtained as $\frac{\overline{Y}}{\pi}$, where $\overline{Y}$ is the mean of the surrogate observations. However, for unknown π, simultaneous estimation of λ and π is carried out based on a sample of size n of the surrogate observations, i.e., Y_1, Y_2, …,Y_n as follows:

The likelihood equation[8] can be written as,

$$L = \prod_{i=1}^{n} f(yi)$$

The log-likelihood equation[9] is given by,

$$l = \sum_{i=1}^{n} \log_e f(y_i)$$

$$= \sum_{i=1}^{n} \log_e(e^{-\lambda\pi} \frac{(\lambda\pi)^{y_i}}{y_i!})$$

$$= \ln(\lambda\pi) \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \ln y_i! - n\lambda\pi$$

We obtain the Score Equations as,

$\frac{\partial l}{\partial \lambda} = 0 \Rightarrow -n\pi + \frac{1}{\lambda}\sum_{i=1}^{n} y_i = 0$ ….. (1)

$\frac{\partial l}{\partial \pi} = 0 \Rightarrow -n\lambda + \frac{1}{\pi}\sum_{i=1}^{n} y_i = 0$ ….. (2)

Using (1), we get, $\hat{\lambda} = \frac{\overline{y}}{\pi}$ …… (3)

Using (2), we get, $\hat{\pi} = \frac{\overline{y}}{\lambda}$ …… (4)

But, (3) and (4) are identical equations. Hence, we cannot get separate estimates of λ and π. This is called the identifiability problem. We can only estimate the function λπ, whose estimate is given by $\widehat{\lambda\pi} = \overline{y}$. Thus, for unknown π, simultaneous estimation of λ and π clearly falls through. To resolve the identifiability issue, additional data is thus essential. Here, we make use of validation data in internal setting for simultaneous estimation of λ and the nuisance parameter π.

A validation sample is a collection of study units, where the variable of interest is accurately observed by the use of a so called gold standard method. In this mode of data collection, first a main study sample of size n is considered where observations on the surrogate (Y) are collected. An internal validation sample of size r (r<n) is then chosen as a subsample from the main study sample where the true variable (Y*) is enumerated by special effort. We assume that the selection of the subsample is by a random process (for the purpose of my study, I have considered SRSWOR) that is independent of the observed data though more general

study design plans could be used. For notational simplicity, we assume that the first r units are selected in the validation sample. The observed data is then denoted by:

DATA:

$Y_1, Y_{2, \ldots\ldots}, Y_n$: Surrogate count (observed data)

$Y^*_1, Y^*_2, \ldots, Y^*_r$: True count (validation data)

2.2: ESTIMATION OF PARAMETERS BY THE METHOD OF JOINT LIKELIHOOD:

Given the data, the likelihood function for the validation and non-validation set can be factored as:

$$L = \prod_{i=1}^{r} f(y_i, y_i *) \prod_{i=r+1}^{n} f(y_i)$$

$$= \prod_{i=1}^{r} f(y_i | y_i *)\, f(y_i *) \prod_{i=r+1}^{n} f(y_i)$$

$$= \prod_{i=1}^{r} \binom{y_i *}{y_i} \pi^{y_i} (1-\pi)^{y_{i*}-y_i}\, e^{-\lambda} \frac{\lambda^{y_{i*}}}{y_i * !} \prod_{i=r+1}^{n} e^{-\lambda\pi} \frac{(\lambda\pi)^{y_i}}{y_i!}$$

$$= \prod_{i=1}^{r} \frac{\pi^{y_i}((1-\pi)\lambda)^{y_{i*}-y_i} \lambda^{y_i} e^{-\lambda(1-\pi)} e^{-\lambda\pi}}{y_i!\,(y_i * - y_i)!} \prod_{i=r+1}^{n} e^{-\lambda\pi} \frac{(\lambda\pi)^{y_i}}{y_i!}$$

$$= \prod_{i=1}^{r} \{(1-\pi)\lambda\}^{z_i} \frac{e^{-\lambda(1-\pi)}}{z_i!} \prod_{i=1}^{n} e^{-\lambda\pi} \frac{(\lambda\pi)^{y_i}}{y_i!}, where\ y_i * - y_i = z_i\,, i = 1,2, \ldots, r$$

Thus,

$$\left.\begin{array}{l} Z_1, Z_2, \ldots, Z_r \sim \text{Poisson}(\lambda(1-\pi)), iid. \\ Y_1, Y_2, \ldots, Y_n \sim \text{Poisson}(\lambda\pi), iid. \end{array}\right\rangle \text{Independently.} \qquad \ldots..\ (5)$$

Let $\alpha = \lambda(1\text{-}\pi)$ and $\beta = \lambda\pi$.

Then, MLE of $\alpha = \overline{Z_r} = \frac{1}{r}\sum_{i=1}^{r} Z_i$ and the MLE of $\beta = \overline{Y_n} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

We have, $\alpha + \beta = \lambda$.

Therefore,

$$\boxed{\begin{gathered}\hat{\lambda} = \hat{\alpha} + \hat{\beta} = \overline{Z_r} + \overline{Y_n} \\ \hat{\pi} = \frac{\overline{Y_n}}{\overline{Z_r} + \overline{Y_n}}\end{gathered}}$$

Here, $\hat{\lambda}\ and\ \hat{\pi}$ are in fact the Maximum Likelihood Estimates of λ and π respectively.

2.3: VARIANCE OF THE MODEL PARAMETER $\hat{\lambda}$:

Further, we obtain the variance of $\hat{\lambda}$ as,

$$Var(\hat{\lambda}) = Var(\overline{Z_r}) + Var(\overline{Y_n}) \quad [from\ (5)]$$
$$= \frac{\lambda(1-\pi)}{r} + \frac{\lambda\pi}{n} \quad [from\ (5)]$$

An alternative derivation for the variance of the model parameter has been provided in the Appendix (Ref. Section 2-Appendix).

2.4: ESTIMATION OF PARAMETERS BY THE METHOD OF PSEUDO LIKELIHOOD:

With the known data in hand, the likelihood function for the validation and non-validation set can be factored as:

$$L = \prod_{i=1}^{r} f(y_i, y_i *) \prod_{i=r+1}^{n} f(\,y_i)$$

$$= \prod_{i=1}^{r} f(y_i|y_i *)\, f(y_i *) \prod_{i=r+1}^{n} f(\,y_i)$$

Taking logarithm on both sides of the above equation, we get the log-likelihood equation as,

$$\ln L = \sum_{i=1}^{r} \ln f(y_i|y_i *) + \sum_{i=1}^{r} \ln f(y_i *) + \sum_{i=r+1}^{n} \ln f(\,y_i)$$

$$= L_1 + L_2 + L_3$$

Where, $L_1 = \sum_{i=1}^{r} \ln f(y_i|y_i *)$

$L_2 = \sum_{i=1}^{r} \ln f(y_i *)$

$L_3 = \sum_{i=r+1}^{n} \ln f(\,y_i)$

Now,

$$L_1 = \sum_{i=1}^{r} \ln f(y_i|y_i *)$$

$$= \sum_{i=1}^{r} \ln\{\binom{y_i *}{y_i} \pi^{y_i}(1-\pi)^{y_i*-y_i}\}$$

$$= \sum_{i=1}^{r} [\ln\binom{y_i *}{y_i} + y_i \ln\pi + (y_i * - y_i)\ln(1-\pi)]$$

Differentiating *L*$_1$ with respect to π, we get,

$$\frac{\partial L_1}{\partial \pi} = 0$$

$$\Rightarrow \frac{1}{\pi}\sum_{i=1}^{r} y_i - \frac{1}{(1-\pi)}\sum_{i=1}^{r}(y_i * - y_i) = 0$$

$$\Rightarrow \frac{1}{\pi}\sum_{i=1}^{r} y_i = \frac{1}{(1-\pi)}\sum_{i=1}^{r}(y_i * - y_i)$$

$$\Rightarrow \left(\frac{1}{\pi} + \frac{1}{(1-\pi)}\right)\sum_{i=1}^{r} y_i = \frac{1}{(1-\pi)}\sum_{i=1}^{r} y_i *$$

$$\Rightarrow \frac{1}{\pi(1-\pi)}\sum_{i=1}^{r} y_i = \frac{1}{(1-\pi)}\sum_{i=1}^{r} y_i *$$

$$\Rightarrow \hat{\pi} = \frac{\sum_{i=1}^{r} y_i}{\sum_{i=1}^{r} y_i *} \qquad \ldots\ldots..(6)$$

$$\boxed{\therefore \hat{\pi} = \frac{\sum_{i=1}^{r} y_i}{\sum_{i=1}^{r} y_i *}}$$

Again,

$$L_2 = \sum_{i=1}^{r} \ln f(y_i *)$$

$$= \sum_{i=1}^{r} \ln\left(e^{-\lambda} \frac{\lambda^{y_i *}}{y_i * !}\right)$$

$$= \sum_{i=1}^{r} (-\lambda + y_i * \ln \lambda - \ln y_i * !)$$

Also,

$$L_3 = \sum_{i=r+1}^{n} \ln f(\, y_i)$$

$$= \sum_{i=r+1}^{n} e^{-\lambda\pi} \frac{(\lambda\pi)^{y_i}}{y_i!}$$

$$= \sum_{i=r+1}^{n} (-\lambda\pi + y_i \ln(\lambda\pi) - \ln(y_i!))$$

$$\therefore, \qquad \frac{\partial L_2}{\partial \lambda} + \frac{\partial L_3}{\partial \lambda} = 0$$

$$\Rightarrow -r + \frac{1}{\lambda}\sum_{i=1}^{r} y_i * - \pi(n-r) + \frac{1}{\lambda}\sum_{i=r+1}^{n} y_i = 0$$

$$\Rightarrow \frac{1}{\lambda}\left(\sum_{i=1}^{r} y_i * + \sum_{i=r+1}^{n} y_i\right) = r + \pi(n-r)$$

$$\Rightarrow \hat{\lambda} = \frac{\sum_{i=1}^{r} y_i * + \sum_{i=r+1}^{n} y_i}{r + \hat{\pi}(n-r)}$$

, where $\hat{\pi}$ is as obtained in (6).

$$\boxed{\therefore, \hat{\lambda} = \frac{\sum_{i=1}^{r} y_i * + \sum_{i=r+1}^{n} y_i}{r + \hat{\pi}(n-r)}}$$

CHAPTER 3: OPTIMAL SAMPLE SIZE SELECTION

From Section 2.3, we see that the Variance of $\hat{\lambda}$ is obtained as,

$$Var(\hat{\lambda}) = \frac{\lambda(1-\pi)}{r} + \frac{\lambda\pi}{n}$$

Let us define the **cost function**[10] as follows:

$$C = rC_0 + nC_1 \qquad \ldots\ldots(1)\,,$$

where, C is the total cost incurred in the survey,

C_0 is the cost of surveying one unit of the internal validation sample of size *r*,

C_1 is the cost of surveying one unit of the main sample of size *n*.

In practise, C_0 is likely to be larger than C_1. Hence, a unit increase in *r* increases the cost more than a unit increase in *n*, and this is just the reverse of the situation in respect of the variance, which decreases at a more rapid rate for increases in *r* than for increases in *n*.

Because of the behaviour of the variance and cost functions in opposite directions for increases in *r* and *n*, it is necessary in practice to evolve an optimum (or near optimum) solution for the problem of determination of the values of r and n such that the efficiency per unit cost is maximum. In other words, I wish to obtain the optimum values of the main sample size(*n*) and the number of validation units (*r*) so as to minimise the sampling variance of the estimator ($\hat{\lambda}$) for a fixed cost.

Here we optimise the values of *r* and *n* by the method of **Lagrange multipliers**[11].

Let us consider the following Lagrangian Function:

$F = V(\hat{\lambda}) + k(C - C^*),$

$where\ k\ is\ the\ Lagrange\ multiplier\ and\ C^* is\ the\ fixed\ total\ cost.$

The Lagrangian Function *F* is a function of the main sample size (*n*), number of validation units (*r*) and the Lagrange multiplier *k*.

$$\therefore, \qquad F = \frac{\lambda(1 - \pi)}{r} + \frac{\lambda\pi}{n} + k(rC_0 + nC_1 - C^*) \qquad \ldots..(2)$$

Partially obtaining the derivative of *F* with respect to *r* and equating to 0, we get,

$$\frac{\partial F}{\partial r} = 0 \Rightarrow \frac{-\lambda}{r^2} + \frac{\lambda\pi}{r^2} + kC_0 = 0$$

$$\Rightarrow kC_0 = \frac{\lambda(1-\pi)}{r^2}$$

$$\Rightarrow r = \sqrt{\frac{\lambda(1-\pi)}{kC_0}} \qquad \ldots\ldots\ldots. (3)$$

Partially obtaining the derivative of *F* with respect to *n* and equating to 0, we get,

$$\frac{\partial F}{\partial n} = 0 \Rightarrow \frac{-\lambda\pi}{n^2} + kC_1 = 0$$

$$\Rightarrow kC_1 = \frac{\lambda\pi}{n^2}$$

$$\Rightarrow n = \sqrt{\frac{\lambda\pi}{kC_1}} \qquad \ldots\ldots\ldots. (4)$$

Putting the values of *r* and *n* from (3) and (4) in (1), we get,

$$C^* = \sqrt{\frac{\lambda(1-\pi)}{kC_0}}\,C_0 + \sqrt{\frac{\lambda\pi}{kC_1}}\,C_1$$

$$\Rightarrow \sqrt{k} = \frac{\sqrt{\lambda(1-\pi)C_0} + \sqrt{\lambda\pi C_1}}{C^*}$$

$$\Rightarrow k = \frac{\lambda(1-\pi)C_0+\lambda\pi C_1+2\lambda\sqrt{\{(1-\pi)\pi C_0 C_1\}}}{C^{*2}} = k', say.$$

Putting the value of *k'* in (3) and (4), we get,

$$r_{opt} = \sqrt{\frac{\lambda(1-\ \pi)}{k' C_0}}$$

And,

$$n_{opt} = \sqrt{\frac{\lambda\pi}{k' C_1}}$$

It is to be noted that the expression for the optimum values of *r* and *n* involve the population parameter λ and the nuisance parameter π. Two courses of action can be undertaken here.

- Either one can use true values of the model parameter λ and the nuisance parameter π.

- Or, one can use the estimates of the parameters, that is, $\hat{\lambda}$ and $\hat{\pi}$. In accordance with the expressions of $\hat{\lambda}$ and $\hat{\pi}$ (as derived in Section 2), it is evident that both these expressions involve *r* and *n*. Thus, in practise, the parameters $\hat{\lambda}$ and $\hat{\pi}$ are estimated by conducting a **pilot survey**[12] or small-scale survey with a few observations.

CHAPTER 4: SIMULATION STUDY

As mentioned earlier, the objective of the study is three-fold:

(a) Estimation of the model parameter λ and the nuisance parameter π.

(b) Optimum allocation of sample of size n and internal validation sample of size r.

(c) Investigating the problem of allocation between validation and non-validation units.

Owing to the non-availability of primary[13] and secondary[14] data on a suitable Poisson variable, in my dissertation, I have proceeded to fulfil the objectives based on artificial data generated under a simulation set-up. The data is generated from a Poisson population with varied choices of the model parameter and as the data has been assumed to be under-reported, varied choices of the nuisance parameter have also been taken into consideration.

The following combinations of parametric values have been considered for the simulation study:

i. $\lambda = 5, \pi = 0.9$

ii. $\lambda = 5, \pi = 0.8$

iii. $\lambda = 2, \pi = 0.8$

iv. $\lambda = 10, \pi = 0.8$

4.1: ALGORITHM:

This study utilises a simulation technique in R to generate the data that is used for validating the findings and attaining the objectives.

The following algorithm is used for simulation purpose.

Step 1: Initialise λ and π to their respective values.

Step 2: Define the sample size *n* for the data to be simulated.

- The choices of *n* are *n = 50, n = 100, n = 150.*

Step 3: A single random sample of size n is drawn without replacement from a Poisson population with parameter λ. Let this be stored in a variable labelled *y*.

Step 4: Set the number of simulations. In this case, the data sets were simulated 1000 times, that is, repeat Step 3 ***'R=1000'*** times.

Step 5: A single random sample of unit size is drawn from a Binomial(y, π) population. Let this be stored in a variable labelled *ycurl*.

Step 6: Repeat Step 5 ***'R=1000'*** times.

Step 7: Sample *r* units from *n* units without replacement.

- Here, we take *r* as *10%*, *20%* and *30%* of *n*.

Step 8: Repeat Step 7 ***'R=1000'*** times.

Step 9: Obtain the sample observations from *y* corresponding to the *r* units sampled in Step 7. Store these in a matrix named *ydata.WOR*.

Step 10: Obtain the sample observations from *ycurl* corresponding to the *r* units sampled in Step 7. Store these in a matrix named *ycurldata.WOR*.

Step 11: Define z_i = *ydata.WOR – ycurldata.WOR, i=1, 2, …, r*.

Step 12: Obtain the estimates of λ and π according to the formulae derived in *Chapter 2 (Section 2.2 and Section 2.4 for the joint likelihood estimates and pseudo likelihood estimates respectively).*

Step 13: Obtain the Variance of $\hat{\lambda}$ in accordance with the formula derived in *Chapter 2 (Section 2.3)*.

Step 14: Note the optimum values of *r* and *n* according to the formula derived in *Chapter 3*.

Step 15: Summarize the results obtained in the form of a Table (*ref. Table 1 and Table 2*).

The R Code for the above algorithm is provided in the Appendix (ref. Section 1- Appendix).

TABLE 1: ESTIMATION OF PARAMETERS BY THE METHOD OF JOINT LIKELIHOOD AND OPTIMAL SAMPLE SIZE SELECTION BY MINIMISING VARIANCE SUBJECT TO COST CONSTRAINT [C = rC$_0$ + nC$_1$]

$$\hat{\lambda} = \hat{\alpha} + \hat{\beta} = \overline{Z_r} + \overline{Y_n} \ ; \quad \hat{\pi} = \frac{\overline{Y_n}}{\overline{Z_r} + \overline{Y_n}} ; \quad \hat{V}(\hat{\lambda}) = \frac{\hat{\lambda}(1-\hat{\pi})}{r} + \frac{\hat{\lambda}\hat{\pi}}{n}$$

Simulation number R=1000

Table 1.1: λ = 5, π = 0.9

n	R	$\hat{\lambda}$	$\hat{\pi}$	$\hat{V}(\hat{\lambda})$	n_{opt}	r_{opt}
50	5	4.98078	0.9032141	0.1863881	43	6
50	10	4.98308	0.9012674	0.1390210	57	9
50	15	4.975913	0.9017238	0.1223389	72	11
100	10	4.9918	0.9022614	0.0938282	86	13
100	20	4.99255	0.9011847	0.0696591	115	17
100	30	4.997833	0.9000399	0.0616353	143	21
150	15	5.00056	0.9003758	0.0632277	129	19
150	30	5.003893	0.8992396	0.0468045	172	26
150	45	5.000804	0.8996007	0.0411488	215	32

Table 1.2: λ = 5, π = 0.8

n	**R**	$\hat{\lambda}$	$\hat{\pi}$	$\hat{V}(\hat{\lambda})$	n_{opt}	r_{opt}
50	5	4.98272	0.8045132	0.2749845	36	8
50	10	4.98712	0.8013945	0.178980	47	11
50	15	4.97545	0.8016418	0.1455654	59	13
100	10	4.98670	0.8042465	0.1377218	71	16
100	20	4.99305	0.8014938	0.08957656	95	21
100	30	5.00193	0.7997537	0.07339044	118	26
150	15	5.00451	0.8008596	0.09315933	106	24
150	30	5.00157	0.8002621	0.05998713	142	32
150	45	5.00104	0.8000483	0.0488953	177	40

Table 1.3: λ = 2, π = 0.8

n	**R**	$\hat{\lambda}$	$\hat{\pi}$	$\hat{V}(\hat{\lambda})$	n_{opt}	r_{opt}
50	5	1.99346	0.8107551	0.1077746	36	8
50	10	1.98886	0.8047360	0.07084542	48	10
50	15	1.98653	0.8031547	0.05797899	59	13
100	10	2.00063	0.8059179	0.05495208	72	16
100	20	1.99653	0.8030117	0.03569702	95	21
100	30	1.99736	0.8013419	0.02923212	118	26
150	15	2.00198	0.8039966	0.03689023	107	24
150	30	2.00335	0.8000921	0.02403524	142	32
150	45	1.99927	0.8009475	0.01951894	177	40

Table 1.4: λ = 10, π = 0.8

n	R	$\hat{\lambda}$	$\hat{\pi}$	$\hat{V}(\hat{\lambda})$	n_{opt}	r_{opt}
50	5	10.0025	0.8002081	0.5597653	35	8
50	10	9.98600	0.7998565	0.3596107	47	11
50	15	9.98230	0.7996003	0.2930003	59	13
100	10	9.99222	0.8023432	0.2776749	71	16
100	20	10.00127	0.8008422	0.1796859	95	21
100	30	9.992353	0.8013616	0.1462370	118	26
150	15	10.01908	0.7997995	0.1871434	106	24
150	30	10.00055	0.8007595	0.1198040	142	32
150	45	10.00501	0.8002287	0.09779121	177	40

Inferences from Table 1

1. We obtain better estimates of the parameter λ as the sample size n increases; better in a sense that the estimate of variance of the parameter λ decreases with increase in n.

2. For a fixed sample size n, we observe that more accurate estimates of the parameters λ and π are obtained when size of the internal validation sample r is 30% of the main sample size. In other words, larger the size of the internal validation sample, better are the estimates.

3. The optimum values of the internal validation sample (r_{opt}) do not differ widely from the values of r obtained by pilot survey.

4. Also, the optimum value of n is closest to the value of n obtained by pilot survey when the true choice of r is 20% of the true choice of n. However, this observation pertains only to the particular cases observed in the study and their applicability for other choice of (n, r) is yet to be tested.

5. n_{opt} is an increasing function of π while r_{opt} is a decreasing function of π. No such conclusion can be drawn regarding the functional relationship of n_{opt} *and* r_{opt} with λ.

<u>TABLE 2</u>: COMPARISON BETWEEN THE METHOD OF JOINT LIKELIHOOD AND THE METHOD OF PSEUDO LIKELIHOOD.

$$\hat{\lambda}_{JL} = \hat{\alpha} + \hat{\beta} = \overline{Z_r} + \overline{Y_n}\ ;\quad \hat{\lambda}_{pseudo} = \frac{\sum_{i=1}^{r} y_{i*} + \sum_{i=r+1}^{n} y_i}{r + \hat{\pi}(n-r)}$$

$$\hat{\pi}_{JL} = \frac{\overline{Y_n}}{\overline{Z_r} + \overline{Y_n}}\ ;\quad \hat{\pi}_{pseudo} = \frac{\sum_{i=1}^{r} y_i}{\sum_{i=1}^{r} y_{i*}}$$

Simulation number R=1000

Table 2.1: λ = 5, π = 0.9

n	r	$\hat{\lambda}_{JL}$	$\hat{\lambda}_{pseudo}$	$\hat{\pi}_{JL}$	$\hat{\pi}_{pseudo}$
50	5	4.98078	4.99789	0.9032141	0.9005454
50	10	4.98308	4.98933	0.9012674	0.8999703
50	15	4.97591	4.97997	0.9017238	0.9007096
100	10	4.99180	5.00091	0.9022614	0.9006504
100	20	4.99255	4.99658	0.9011847	0.9003248
100	30	4.99783	4.99839	0.9000399	0.8999133
150	15	5.00056	5.00748	0.9003758	0.8991280
150	30	5.00389	5.00597	0.8992396	0.8988070
150	45	5.00080	5.00244	0.8996007	0.8991929

Table 2.2: λ = 5, π = 0.8

n	r	$\hat{\lambda}_{JL}$	$\hat{\lambda}_{pseudo}$	$\hat{\pi}_{JL}$	$\hat{\pi}_{pseudo}$
50	5	4.98272	5.01704	0.8045132	0.8005259
50	10	4.98712	5.00279	0.8013945	0.7986739
50	15	4.97545	4.98336	0.8016418	0.7999032
100	10	4.98670	5.00508	0.8042466	0.8016133

100	20	4.99305	5.00149	0.8014938	0.7999663
100	30	5.00193	5.00369	0.7997537	0.7994436
150	15	5.00451	5.02093	0.8008596	0.7984111
150	30	5.00184	5.00568	0.8002621	0.7996277
150	45	5.00104	5.00456	0.8000483	0.7992771

Table 2.3: $\lambda = 2$, $\pi = 0.8$

n	r	$\hat{\lambda}_{JL}$	$\hat{\lambda}_{pseudo}$	$\hat{\pi}_{JL}$	$\hat{\pi}_{pseudo}$
50	5	1.99346	2.03798	0.8107551	0.7971408
50	10	1.98886	2.00456	0.8047360	0.7990107
50	15	1.98653	1.99289	0.8031547	0.8001434
100	10	2.00063	2.02263	0.8039966	0.7987559
100	20	1.99653	2.00295	0.8000921	0.8001458
100	30	1.99736	2.00202	0.8009475	0.7988077
150	15	2.00198	2.02073	0.8039966	0.7973379
150	30	2.00335	2.00807	0.8000921	0.7980266
150	45	1.99927	2.00197	0.8009475	0.7996105

Table 2.4: $\lambda = 10$, $\pi = 0.8$

n	r	$\hat{\lambda}_{JL}$	$\hat{\lambda}_{pseudo}$	$\hat{\pi}_{JL}$	$\hat{\pi}_{pseudo}$
50	5	10.00250	10.05399	0.8002081	0.7963818
50	10	9.98600	9.99830	0.7998565	0.7987837
50	15	9.98230	9.99021	0.7996003	0.7987078
100	10	9.99222	10.01043	0.8023432	0.8011157
100	20	10.00127	10.00778	0.8008422	0.8002721
100	30	9.992353	9.99354	0.8013616	0.8012735
150	15	10.01908	10.03033	0.7997995	0.7990119
150	30	10.00055	10.00147	0.8007595	0.8007531
150	45	10.00501	10.00713	0.8002287	0.8000085

Inferences from table 2

1. On visual inspection, the estimates of λ, obtained by the method of pseudo likelihood is slightly more accurate than the estimates λ, obtained by the method of joint likelihood.
2. However, the results are just the opposite for the estimates of π. We obtain more accurate estimate of π by the method of joint likelihood.
3. The results hold good for any true choices of π and λ.

However, the gain in precision is not significant enough to prefer one method over another.

Query:

WHETHER ALLOCATING THE TOTAL RESOURCES FOR ASCERTAINING FULLY VALIDATED DATA IS ADVANTAGEOUS COMPARED TO SHARING ALLOCATION BETWEEN VALIDATED AND NON-VALIDATED UNITS?

Let us now consider two methodologies:-

Method 1: where the resources are shared between validation and non-validation data

A sample of size *n* of observations on the count variable of interest is drawn from the population where the data is known to have been contaminated by under-reporting error. An internal validation sample of size *r* (*r*<*n*) is then obtained as a subsample of the main study sample where the true count variable is enumerated by special effort. We obtain the estimate of the model parameter λ based on these two samples (*ref. Section 2.2*).

We get, $\widehat{\lambda_1} = \overline{Z_r} + \overline{Y_n}$, where, $\overline{Z_r} = \frac{1}{r}\sum_{i=1}^{r} Z_i \; and \; \overline{Y_n} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, and,

$Z_1, Z_2, \ldots, Z_r \sim$ Poisson$(\lambda(1-\pi))$
$>$ Independently.
$Y_1, Y_2, \ldots, Y_n \sim$ Poisson$(\lambda\pi)$

Also, we obtain the variance of the estimate $\widehat{\lambda_1}$ as, $Var\left(\widehat{\lambda_1}\right) = \frac{\lambda(1-\pi)}{r} + \frac{\lambda\pi}{n}$

Therefore, the estimate of variance is, $\hat{V}\left(\widehat{\lambda_1}\right) = \frac{\hat{\lambda}(1-\hat{\pi})}{r} + \frac{\hat{\lambda}\hat{\pi}}{n} = V_1, say.$

Method 2: where total resources are allocated for ascertaining fully validated data

Consider a fully validated data of size r_1 of fully validated data obtained by special efforts. Here, the data has not been exposed to any form of reporting errors and hence, we simply obtain the estimate of the model parameter as, $\widehat{\lambda_2} = \overline{Y\,*} = \frac{1}{r_1}\sum_{i=1}^{r_1} Y_i^{\,*}$. It is to be noted that Y_i^* is generated from *Poisson (λ)*.Then, in this case, we obtain the variance of the estimate $\widehat{\lambda_2}$ as, $V\left(\widehat{\lambda_2}\right) = \frac{\lambda}{r_1}$. Therefore, the estimate of variance is, $\hat{V}\left(\widehat{\lambda_2}\right) = \frac{\hat{\lambda}}{r_1} = V_2, say.$

The natural query that arises in this situation is that which of the above methods leads to more efficient estimation of the model parameter by minimising the variance, while keeping the cost of survey at a fixed level. Since data on the true count variable can be obtained only by employing higher quality resources and better trained personnel, the cost of surveying one unit of the internal validation sample is quite higher than the per unit cost of survey of the main study sample. In similar lines with the query, I simply want to examine whether the estimate of the model parameter based on the validation and non-validation samples is more accurate than that obtained based on a validation sample of smaller size, owing to the fact that the cost of survey for the methods is kept fixed.

PROCEDURE:

1. We have, $\boldsymbol{V_1} = \frac{\hat{\lambda}(1-\hat{\pi})}{r} + \frac{\hat{\lambda}\hat{\pi}}{n}$;

   $$\boldsymbol{V_2} = \frac{\hat{\lambda}}{r_1}$$

2. Suppose Cost per unit for surveying validation units (*C₀*) is *5k*

   And Cost per unit for surveying non-validation units (*C₁*) is *k*.

   Then, the Total Cost is given by, $C = rC_0 + nC_1 = r \times 5k + n \times k = k(5r + n)$

   Let, without loss of generality, *k = 1*. Then, $C = 5r + n$.

3. Now, I consider various choices of (*n, r*). Compute $\mathbf{V_1}$. Suppose, *n = 50, r = 2*. Then, *C = 60*. Now, while computing $\mathbf{V_2}$, keeping cost same, I use $r_1 = 12$. Compare the variances $\mathbf{V_1}$ and $\mathbf{V_2}$.

4. Repeat this for various choices of *n* and *r* such that *C = 60*.

5. Next, we change the total cost to *C=100* and proceed as earlier.

TABLE 3: TABLE TO DEMONSTRATE WHETHER ALLOCATING THE TOTAL RESOURCES FOR ASCERTAINING FULLY VALIDATED DATA IS ADVANTAGEOUS COMPARED TO SHARING ALLOCATION BETWEEN VALIDATED AND NON-VALIDATED DATA.

$$\widehat{\boldsymbol{\lambda_1}} = \overline{Z_r} + \overline{Y_n}\ ;\ \ \widehat{\boldsymbol{\lambda_2}} = \overline{Y*} = \frac{1}{r_1}\sum_{i=1}^{r_1} Y_i^*$$

$$\boldsymbol{V_1} = \hat{V}(\widehat{\lambda_1}) = \frac{\hat{\lambda}(1-\hat{\pi})}{r} + \frac{\hat{\lambda}\hat{\pi}}{n};\ \ \boldsymbol{V_2} = \hat{V}(\widehat{\lambda_2}) = \frac{\hat{\lambda}}{r_1}$$

$$Cost\ Function\ C = k \times n + pk \times r = k(n + pr), where\ C_0 = pC_1.$$

WLG, k=1

Simulation number R=1000

Table 3.1: C = 5r + n, λ = 5, π = 0.9

Total Cost And r_1	N	r	$\widehat{\lambda_1}$	$\widehat{\lambda_2}$	V_1	V_2
C=60 **r_1=12**	50	2	4.97178	4.98162	0.315179	0.415135
	35	5	4.99206	4.97917	0.228113	0.414931
	25	7	4.99163	4.98552	0.249949	0.415460
	10	**10**	**4.97460**	**4.97460**	**0.497460**	**0.414550**
C=100 **r_1=20**	**90**	**2**	**4.99544**	**4.99248**	**0.281566**	**0.249624**
	75	5	4.98056	4.98729	0.155787	0.249365
	50	10	4.98308	4.98162	0.139021	0.249081
	40	12	4.97802	4.97780	0.153230	0.248890
	25	15	4.98880	4.98520	0.212707	0.249276
	20	**16**	**4.97410**	**4.97510**	**0.254867**	**0.248755**
	10	**18**	**4.48550**	**4.97460**	**0.448550**	**0.248730**
C=125 **r_1=25**	**115**	**2**	**5.00038**	**4.99800**	**0.270734**	**0.19992**
	100	5	4.98930	4.99615	0.140479	0.19985
	75	10	4.98836	4.98729	0.109375	0.19949
	50	15	4.97591	4.98162	0.122338	0.19927
	25	**20**	**4.98270**	**4.98552**	**0.204199**	**0.19921**
	15	**22**	**4.48700**	**4.97973**	**0.299133**	**0.19919**
C=200 **r_1=40**	175	5	5.00937	5.00257	0.123705	0.12506
	150	10	5.01103	4.99934	0.080539	0.12498
	100	20	4.99255	4.99615	0.069659	0.12490
	75	25	4.98464	4.98729	0.079701	0.12468
	50	30	4.98215	4.98162	0.106282	0.12454
	25	**35**	**4.49220**	**4.98552**	**0.179688**	**0.12464**
	15	**37**	**4.48700**	**4.97973**	**0.299133**	**0.12449**
C=300 **r_1=60**	**275**	**5**	**4.99208**	**5.00245**	**0.110888**	**0.08337**
	250	10	4.99687	5.00286	0.066561	0.08338
	210	18	4.99608	4.99871	0.048839	0.08331
	150	30	5.00389	4.99934	0.046805	0.08332
	125	35	5.00071	4.99828	0.050352	0.08330
	100	40	4.99678	4.99615	0.057457	0.08327
	50	**50**	**4.98162**	**4.98162**	**0.099632**	**0.08303**

Table 3.2: C = 2r + n, λ = 5, π = 0.9

Total Cost And r_1	N	r	$\widehat{\lambda_1}$	$\widehat{\lambda_2}$	V_1	V_2
C=60 **r_1=30**	**50**	**5**	**4.98078**	**4.98162**	**0.186388**	**0.166054**
	44	8	4.98158	4.97893	0.163373	0.165964
	40	10	4.97483	4.97780	0.160962	0.165927
	24	**18**	**4.99006**	**4.98683**	**0.214796**	**0.166228**
	20	**20**	**4.97510**	**4.97510**	**0.248755**	**0.165837**
C=100 **r_1=50**	**90**	**5**	**4.99394**	**4.99249**	**0.147578**	**0.099850**
	80	**10**	**4.98486**	**4.98909**	**0.105174**	**0.099782**
	64	18	4.99485	4.98725	0.098276	0.099745
	50	**25**	**4.98098**	**4.98162**	**0.109546**	**0.099632**
	36	**32**	**4.97971**	**4.97786**	**0.140056**	**0.099557**
C=150 **r_1=75**	**140**	**5**	**5.00039**	**4.99899**	**0.129458**	**0.066653**
	120	**15**	**5.00215**	**4.99911**	**0.070772**	**0.066655**
	90	30	4.98958	4.99249	0.066499	0.066567
	70	**40**	**4.98930**	**4.98916**	**0.076648**	**0.066522**
	50	**50**	**4.98162**	**4.98162**	**0.099632**	**0.066422**
C=200 **r_1=30**	**180**	**10**	**4.99738**	**5.00134**	**0.073986**	**0.050013**
	150	25	4.99855	4.99340	0.049942	0.049993
	100	**50**	**4.99892**	**4.99615**	**0.055008**	**0.049962**
	80	**60**	**4.98750**	**4.98909**	**0.064427**	**0.049891**

Table 3.3: C = 8r + n, λ = 5, π = 0.9

Total Cost And r_1	N	r	$\widehat{\lambda_1}$	$\widehat{\lambda_2}$	V_1	V_2
C=80 **r_1=10**	64	2	4.98041	4.98725	0.297478	0.498725
	40	5	4.97323	4.97780	0.208067	0.497780
	24	7	4.98064	4.98683	0.255892	0.498683
C=160 **r_1=20**	120	5	5.01535	4.99911	0.137684	0.249955
	80	10	4.98463	4.98909	0.105174	0.249454
	40	15	4.98116	4.97780	0.145243	0.248890
	32	16	4.98466	4.98313	0.171239	0.249156
	16	18	4.48438	4.97625	0.280273	0.248814

FOOTNOTE: In table 3, the highlighted figures are the cases where allocating the total resources for ascertaining fully validated data is advantageous compared to sharing allocation between validated and non-validated data.

Inferences from table 3

1. The glaring point that is evident from this table is that as the difference between C_0 and C_1 increases, that is, greater the value of C_0 compared to C_1, Method-2 becomes less efficient than Method-1. In particular, we see that if $C = 2r + n$, that is, if cost per unit for surveying validation units is twice the cost per unit for surveying non-validation units, then for most choices of (n, r), allocating total resources for ascertaining fully validated data proves to be more efficient than sharing resources between validation and non-validation data. If $C = 5r + n$, that is, cost per unit for surveying validation units is five times the cost per unit for surveying non-validation units, the efficiency of the second method decreases and if $C = 8r + n$, that is, if $C_0=8C_1$, the second method becomes completely useless.

2. As the difference between n and r decreases, that is, as the size of the internal validation sample increases, allocating total resources for ascertaining fully validated data proves to be more useful than sharing resources between validation and non-validation data.
3. A general observation from the table is that if $n<r<r_{1,}$ the second method turns out to be more efficient than the first method. However, this is not a desirable circumstance as we have assumed the size of the internal validation sample (r) to be less than the size of the main error contaminated sample (n).
4. Comparing the values of V_1 and V_2 for all choices of (n, r) in the *Tables 3.1, 3.2 and 3.3*, we observe that on an average sharing the total resources between validation and non-validation data turns out to be more efficient than allocating the total resources for ascertaining fully validated data.

CHAPTER 4: CONCLUDING REMARKS

Data collection often involves reporting errors. Under-reporting, a more common problem in counting systems happens when the reporting of some events is not complete. As a consequence of under-reporting, the mean of the observed counts is smaller than the true mean. Ignoring the under-reporting pattern of count data could result in biased estimates of the effects of interest, which ultimately leads to misleading inferences. Extension of the standard Poisson model has been proposed so that the under-reporting patterns can be captured. In the course of the previous chapters, it is well established that the corrected estimators are rather satisfactory in terms of having low variance. We have also seen that larger samples are more efficient in estimating the population parameters since they offer better representation of the population.

In an ideal world, it would always be possible to use statistically sound sampling techniques to produce price indices with a high degree of accuracy and within given resource constraints. Reality, however, is usually very far away from this ideal situation. It is almost always impossible to achieve efficient samples because sampling frames are always deficient to some extent, missing some key information or the response rates are unpredictable and may prove to be deficient, which affects the accuracy of the price index levels and measured price changes. In the study, we have come across methods to optimise the sampling frame while ensuring there in no undue wastage of resources.

As stated earlier, larger samples are more advantageous for estimation of model parameters because they offer a better representation of the population. However, when under-reported data is under scrutiny, there are quite a few numbers of instances when larger samples containing both validated and non-validated data are less efficient than smaller samples comprising fully validated data, under certain cases of cost constraints.

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to my dissertation advisor, Prof. Dr. Surupa Chakraborty for the continuous support of my dissertation, for her patience and immense knowledge. As, my professor and mentor, she has taught me more than I can give her credit for.

I am grateful to all those with whom I had the pleasure to work with during my dissertation course.

I am also grateful to St. Xavier's College (Autonomous), Kolkata for giving me an opportunity to enrich my knowledge and allowing me to test the skills I have garnered over the last three years in my under-graduate course.

BIBLIOGRAPHY

1. Dealing with under-reported variables: An information theoretic solution- Konstantinos Sechidis, Gavin Brown.
2. Adjusting for MisReporting in count data- Gelareh Rahimighazikalayeh
3. www.semanticscholar.org

APPENDIX

SECTION 1: R CODES

1.1:

```
rm(list=ls())
set.seed(seed=123)

R=1000
n=50
r=10
lambda=5
pi=0.9
r1=20

y=matrix(0,nrow=n,ncol=R)
y=replicate(R,rpois(n,lambda))
y
ysum=apply(y,2,sum)
ysum
ymean=apply(y,2,mean)
ymean

A=matrix(0,nrow=n,ncol=R)
for(i in 1:n)
{
for(j in 1:R)
{

  A[i,j]=rbinom(1,y[i,j],pi)
}
}
ycurl=A
ycurl
ycurl.total=apply(ycurl,2,sum)
ycurl.total
```

```
ycurlmean=apply(ycurl,2,mean)
ycurlmean

place.WOR=matrix(0,nrow=r,ncol=R)
for(j in 1:R)
{
 place.WOR[,j]=sample(1:n,r)
}

ydata.WOR=matrix(0,nrow=r,ncol=R)
for(i in 1:r)
{
 for(j in 1:R)
 {
  ydata.WOR[i,j]=y[place.WOR[i,j],j]
}
}
ydata.WOR
ydata.sum.WOR=apply(ydata.WOR,2,sum)
ydata.sum.WOR

ycurl.WOR=matrix(0,nrow=r,ncol=R)
for(i in 1:r)
{
 for(j in 1:R)
 {
   ycurl.WOR[i,j]=ycurl[place.WOR[i,j],j]
 }
}
ycurl.data.WOR=ycurl.WOR
ycurl.data.WOR
ycurl.sum.WOR=apply(ycurl.WOR,2,sum)
ycurl.sum.WOR

zi=ydata.WOR-ycurl.data.WOR
```

```
zi
zmean=apply(zi,2,mean)
zmean

pi.hat.WOR=ycurlmean/(ycurlmean+zmean)
pi.hat.WOR
pi.hat=sum(pi.hat.WOR)/R
pi.hat

lambda.hat.WOR=ycurlmean+zmean
lambda.hat.WOR

pi.hat=sum(pi.hat.WOR)/R
pi.hat

lambda.hat=sum(lambda.hat.WOR)/R
lambda.hat

V1=((lambda.hat*pi.hat)/n)+((lambda.hat*(1-pi.hat))/r)
V1

c0=8
c1=1
c=r*c0+n*c1
c
k=(lambda.hat*c0-lambda.hat*pi.hat*c0+lambda.hat*pi.hat*c1+2*lambda.hat*sqrt((1-
pi.hat)*pi.hat*c0*c1))/(c*c)
ropt=sqrt((lambda.hat*(1-pi.hat))/(k*c0))
ropt
nopt=sqrt((lambda.hat*pi.hat)/(k*c1))
nopt

ynew=matrix(0,nrow=r1,ncol=R)
ynew=replicate(R,rpois(r1,lambda))
ynew
```

```
ynewmean=apply(y,2,mean)
ynewmean
lambda.est.alt=sum(ynewmean)/R
lambda.est.alt
v2.alt=lambda.est.alt/r1
v2.alt
```

1.2:

```
rm(list=ls())
set.seed(seed=123)

R=1000
n=50
r=5
lambda=5
pi=0.9

y=matrix(0,nrow=n,ncol=R)
y=replicate(R,rpois(n,lambda))
y
ysum=apply(y,2,sum)
ysum
ymean=apply(y,2,mean)
ymean

A=matrix(0,nrow=n,ncol=R)
for(i in 1:n)
{
for(j in 1:R)
{

  A[i,j]=rbinom(1,y[i,j],pi)
}
}
ycurl=A
```

```
ycurl
ycurl.total=apply(ycurl,2,sum)
ycurl.total
ycurlmean=apply(ycurl,2,mean)
ycurlmean

place.WOR=matrix(0,nrow=r,ncol=R)
for(j in 1:R)
{
 place.WOR[,j]=sample(1:n,r)
}

ydata.WOR=matrix(0,nrow=r,ncol=R)
for(i in 1:r)
{
 for(j in 1:R)
 {
  ydata.WOR[i,j]=y[place.WOR[i,j],j]
}
}
ydata.WOR
ydata.sum.WOR=apply(ydata.WOR,2,sum)
ydata.sum.WOR

ycurl.WOR=matrix(0,nrow=r,ncol=R)
for(i in 1:r)
{
 for(j in 1:R)
 {
   ycurl.WOR[i,j]=ycurl[place.WOR[i,j],j]
 }
}
ycurl.data.WOR=ycurl.WOR
ycurl.data.WOR
ycurl.sum.WOR=apply(ycurl.WOR,2,sum)
ycurl.sum.WOR

zi=ydata.WOR-ycurl.data.WOR
zi
zmean=apply(zi,2,mean)
zmean

pi.hat.WOR=ycurl.sum.WOR/ydata.sum.WOR
pi.hat.WOR
pi.hat.pseudo=sum(pi.hat.WOR)/R
pi.hat.pseudo

lambda.hat.WOR=(ydata.sum.WOR+ycurl.total-ycurl.sum.WOR)/(r+(n-r)*pi.hat.WOR)
lambda.hat.WOR

lambda.hat.pseudo=sum(lambda.hat.WOR)/R

lambda.hat.pseudo
pi.hat.pseudo
```

SECTION 2: ALTERNATIVE DERIVATION OF VARIANCE OF THE PARAMETERS

With the known data in hand, the likelihood function for the validation and non-validation set can be factored as:

$$L = \prod_{i=1}^{r} f(y_i, y_i *) \prod_{i=r+1}^{n} f(\, y_i)$$

$$= \prod_{i=1}^{r} f(y_i|y_i *)\, f(y_i *) \prod_{i=r+1}^{n} f(\, y_i)$$

Taking logarithm on both sides of the above equation, we get the log-likelihood equation as,

$$\ln L = \sum_{i=1}^{r} \ln f(y_i|y_i *) + \sum_{i=1}^{r} f(y_i *) + \sum_{i=r+1}^{n} \ln f(\, y_i)$$

$$\therefore, \quad \frac{\partial \ln L}{\partial \pi} = \frac{1}{\pi}\sum_{i=1}^{r} y_i - \frac{1}{(1-\pi)}\sum_{i=1}^{r}(y_i * - y_i) - \sum_{i=r+1}^{n} \lambda + \sum_{i=r+1}^{n} \frac{y_i}{\lambda\pi}.\lambda$$

$$= \frac{1}{\pi}\sum_{i=1}^{r} y_i - \frac{1}{(1-\pi)}\sum_{i=1}^{r}(y_i * - y_i) - (n-r)\lambda + \sum_{i=r+1}^{n} \frac{y_i}{\pi}$$

$$\therefore, \quad \frac{\partial^2 \ln L}{\partial \pi^2} = -\sum_{i=1}^{n} \frac{y_i}{\pi^2} - \sum_{i=1}^{r} \frac{(y_i * - y_i)}{(1-\pi)^2}$$

$$\therefore, \quad \frac{\partial \ln L}{\partial \lambda} = -r + \frac{1}{\lambda}\sum_{i=1}^{r} y_i * - \pi(n-r) + \sum_{i=r+1}^{n} \frac{y_i}{\lambda}$$

$$\therefore, \quad \frac{\partial^2 \ln L}{\partial \lambda^2} = -\frac{1}{\lambda^2}\left(\sum_{i=1}^{r} y_i * + \sum_{i=r+1}^{n} y_i\right)$$

$$\therefore, \quad \frac{\partial^2 \ln L}{\partial \lambda \partial \pi} = -(n-r)$$

$$\therefore, \quad \frac{\partial^2 \ln L}{\partial \pi \partial \lambda} = -(n-r)$$

$$E\left(\frac{\partial^2 \ln L}{\partial \lambda^2}\right) = -\tfrac{1}{\lambda^2} E\left(\textstyle\sum_{i=1}^{r} y_i * + \sum_{i=r+1}^{n} y_i\right)$$

$$= -\tfrac{1}{\lambda^2}(r\lambda + (n-r)\lambda\pi)$$

$$= -\tfrac{1}{\lambda}[r + (n-r)\pi]$$

$$E\left(\frac{\partial^2 \ln L}{\partial \lambda \partial \pi}\right) = E\left(\frac{\partial^2 \ln L}{\partial \pi \partial \lambda}\right) = -(n-r)$$

$$E\left(\frac{\partial^2 \ln L}{\partial \pi^2}\right) = -\tfrac{1}{\pi^2} E\left(\textstyle\sum_{i=1}^{n} y_i\right) - \tfrac{1}{(1-\pi)^2}\left[E\left(\textstyle\sum_{i=1}^{r} y_i *\right) - E\left(\textstyle\sum_{i=1}^{r} y_i\right)\right]$$

$$= -\tfrac{n\lambda\pi}{\pi^2} - \tfrac{1}{(1-\pi)^2}(r\lambda - r\lambda\pi)$$

$$= -\lambda\left(\frac{n}{\pi} + \frac{r}{1-\pi}\right)$$

The Information matrix is given by,

$$I = \begin{bmatrix} -E\left(\frac{\partial^2 \ln L}{\partial\lambda^2}\right) & -E\left(\frac{\partial^2 \ln L}{\partial\lambda\partial\pi}\right) \\ -E\left(\frac{\partial^2 \ln L}{\partial\pi\partial\lambda}\right) & -E\left(\frac{\partial^2 \ln L}{\partial\pi^2}\right) \end{bmatrix}$$

$$= \begin{bmatrix} \frac{r+(n-r)\pi}{\lambda} & (n-r) \\ (n-r) & \lambda\left(\frac{n}{\pi} + \frac{r}{1-\pi}\right) \end{bmatrix}$$

$$\therefore, \quad I^{-1} = \frac{1}{|I|} Adj\, I$$

Now, we obtain, $|I| = \frac{nr}{\pi(1-\pi)}$

$$\therefore, \quad I^{-1} = \frac{\pi(1-\pi)}{nr} \begin{bmatrix} \lambda\left(\frac{n}{\pi} + \frac{r}{1-\pi}\right) & -(n-r) \\ -(n-r) & \frac{r+(n-r)\pi}{\lambda} \end{bmatrix}$$

$$\therefore,\ V(\hat{\lambda}) = \frac{\pi(1-\pi)}{nr}.\lambda\left(\frac{n}{\pi} + \frac{r}{1-\pi}\right)$$

$$= \frac{\pi(1-\pi)}{nr}.\lambda\left[\frac{n(1-\pi)+r\pi}{\pi(1-\pi)}\right]$$

$$= \frac{\lambda(n-\pi(n-r))}{nr}$$

$$= \frac{\lambda(1-\pi)}{r} + \frac{\lambda\pi}{n}$$

$$\therefore,\ V(\hat{\pi}) = \frac{\pi(1-\pi)}{nr}.\frac{r+(n-r)\pi}{\lambda}$$

$$\therefore,\ Cov(\hat{\lambda},\hat{\pi}) = -\frac{\pi(1-\pi)(n-r)}{nr}$$

SECTION 3: KEYWORDS

1. ***POISSON MODEL:***

   The Poisson distribution is popular for modelling the number of times an event occurs in an interval of time or space.

2. ***NEGATIVE BINOMIAL MODEL:***

   In probability theory and statistics, the negative binomial distribution is a discrete probability distribution that models the number of failures in a sequence of independent and identically distributed Bernoulli trials before a specified (non-random) number of successes (denoted r) occurs.

3. ***HURDLE OR 2 PART MODEL:***

   The two-part model relaxes the assumption that the zeros (whether or not there are events) and positives (how many events) come from the same data generating processes.

   Example: different factors may affect whether or not you practice a particular sport and how many times you practice your sport in a month. We can estimate two-part models similar to the truncated regression models.

4. ***ZERO INFLATED MODEL:***

   The zero-inflated model is used with count data when there is an excess zeros problem. The zero-inflated model lets the zeros occur in two different ways: as a realization of the binary process (z=0) and as a realization of the count process when the binary variable z=1. Example: you either like hiking or you do not. If you like *hiking, the number of hiking trips you can take is 0, 1, 2, 3, etc. So you may like hiking, but may not take a trip this year. We are able to generate more zeros in the data.*

5. ***BERNOULLI TRIALS:***

   In the theory of probability and statistics, a Bernoulli trial (or binomial trial) is a random experiment with exactly two possible outcomes, "success" and "failure", in which the probability of success is the same every time the experiment is conducted.

6. ***MVUE:***

   In statistics a minimum-variance unbiased estimator (MVUE) or uniformly minimum-variance unbiased estimator (UMVUE) is an unbiased estimator that has lower variance than any other unbiased estimator for all possible values of the parameter.

7. ***MLE:***

   In statistics, maximum likelihood estimation (MLE) is a method of estimating the parameters of a probability distribution by maximizing a likelihood function, so that under the assumed statistical model the observed data is most probable. The point in the parameter space that maximizes the likelihood function is called the maximum likelihood estimate. The logic of maximum likelihood is both intuitive and flexible, and as such the method has become a dominant means of statistical inference.

8. ***LIKELIHOOD EQUATION:***

In statistics, the likelihood function (often simply called the likelihood) measures the goodness of fit of a statistical model to a sample of data for given values of the unknown parameters. It is formed from the joint probability distribution of the sample, but viewed and used as a function of the parameters only, thus treating the random variables as fixed at the observed values.

9. LOG-LIKELIHOOD EQUATION:

Obtaining the logarithm of the likelihood equation results in the log-likelihood equation.

10. COST FUNCTION:

A cost function is a mathematical formula used to chart how production expenses will change at different output levels. In other words, it estimates the total cost of production given a specific quantity produced.

11. LAGRANGE MULTIPLIER METHOD:

In mathematical optimization, the method of Lagrange multipliers is a strategy for finding the local maxima and minima of a function subject to equality constraints (i.e., subject to the condition that one or more equations have to be satisfied exactly by the chosen values of the variables). The basic idea is to convert a constrained problem into a form such that the derivative test of an unconstrained problem can still be applied. Once stationary points have been identified from the first-order necessary conditions, the definiteness of the bordered Hessian matrix determines whether those points are maxima, minima, or saddle points.

The great advantage of this method is that it allows the optimization to be solved without explicit parameterization in terms of the constraints. As a result, the method *of Lagrange multipliers is widely used to solve challenging constrained optimization problems. The method can be summarized as follows: in order to find the stationary points of a function f(x) subjected to the equality constraint g(x) = 0, form the Lagrangian function, L(x ,λ) = f(x) + λg(x).*

12. PILOT SURVEY:

A pilot study, pilot project, pilot test, or pilot experiment is a small scale preliminary study conducted in order to evaluate feasibility, duration, cost, adverse events, and improve upon the study design prior to performance of a full-scale research project.

13. PRIMARY DATA:

Primary data is data that is collected by a researcher from first-hand sources, using methods like surveys, interviews, or experiments. It is collected with the research project in mind, directly from primary sources.

14. SECONDARY DATA:

The term 'primary data' is used in contrast with the term secondary data. Secondary data is data gathered from studies, surveys, or experiments that have been run by other people or for other research.