Integrated Likelihood Inference in Count Distributions

Timothy Ruel

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Table of contents

1 Introduction

Consider a vector $\theta = (\theta_1, ..., \theta_n)$ in which each component represents the mean of a distinct Poisson process. The purpose of this paper is to discuss the task of conducting likelihood-based inference for a real-valued parameter of interest $\psi = \tau(\theta)$. In particular, we will examine the utility of the integrated likelihood function as a tool for obtaining interval and point estimates for ψ, using the performance of the more easily calculated profile likelihood as a benchmark.

We may obtain a sample of values from each Poisson process through repeated measurements of the number of events it generates over a fixed period of time. Suppose we have done so, and let X_{ij} represent the jth count from the ith sample, so that $X_{ij} \sim \text{Poisson}(\theta_i)$ for $i = 1, ..., n$ and $j = 1, ..., m_i$. The probability mass function (pmf) for a single observation $X_{ij} = x_{ij}$ is

$$p(x_{ij};\, \theta_i) = \frac{e^{-\theta_i}\theta_i^{x_{ij}}}{x_{ij}!},\;\; x_{ij} = 0, 1, 2, ...;\;\; \theta_i > 0. \tag{1.1}$$

Denote the sample of counts from the ith process by the vector $X_{i\bullet} = (X_{i1}, ..., X_{im_i})$, its associated mean by $\bar{X}_{i\bullet} = \frac{1}{m_i}\sum_{j=1}^{m_i} X_{ij}$, and assume that all of the counts both within and between samples are measured independently. The likelihood function for an individual component θ_i based on the data $X_{i\bullet} = x_{i\bullet}$ is then equal to the product of the individual probabilities of the observed counts, i.e.

$$\begin{aligned}
L(\theta_i; x_{i\bullet}) &= \prod_{j=1}^{m_i} p(x_{ij}; \theta_i) \\
&= \prod_{j=1}^{m_i} \frac{e^{-\theta_i}\theta_i^{x_{ij}}}{x_{ij}!} \\
&= \left(\prod_{j=1}^{m_i} e^{-\theta_i}\right)\left(\prod_{j=1}^{m_i} \theta_i^{x_{ij}}\right)\left(\prod_{j=1}^{m_i} x_{ij}!\right)^{-1} \\
&= \left(e^{-\sum_{j=1}^{m_i}\theta_i}\right)\left(\theta_i^{\sum_{j=1}^{m_i} x_{ij}}\right)\left(\prod_{j=1}^{m_i} x_{ij}!\right)^{-1} \\
&= e^{-m_i\theta_i}\theta_i^{m_i\bar{x}_{i\bullet}}\left(\prod_{j=1}^{m_i} x_{ij}!\right)^{-1}.
\end{aligned} \tag{1.2}$$

Since L is only useful to the extent that it informs our understanding of the value of θ_i, we are free to replace it with any other function differing from it by just a (nonzero) multiplicative term that is constant with respect to θ_i, provided that the result still satisfies the necessary regularity conditions, as this will not change any conclusions regarding θ_i that we draw from it. Hence, we may safely discard the term in parentheses on the final line of Equation 1.2 as it does not depend on θ_i and instead simply write

$$L(\theta_i; x_{i\bullet}) = e^{-m_i\theta_i}\theta_i^{m_i\bar{x}_{i\bullet}}. \tag{1.3}$$

It will generally be more convenient to work with the log-likelihood function, which is given by

$$\begin{aligned}\ell(\theta_i; x_{i\bullet}) &= \log L(\theta_i; x_{i\bullet}) \\ &= \log\left(e^{-m_i\theta_i}\theta_i^{m_i\bar{x}_{i\bullet}}\right) \\ &= -m_i\theta_i + m_i\bar{x}_{i\bullet}\log\theta_i \\ &= m_i(\bar{x}_{i\bullet}\log\theta_i - \theta_i).\end{aligned} \tag{1.4}$$

The sum of the log-likelihood functions for each component of θ then forms the basis of the log-likelihood function for θ itself:

$$\begin{aligned}\ell(\theta; x_{1\bullet}, ..., x_{n\bullet}) &= \log L(\theta; x_{1\bullet}, ..., x_{n\bullet}) \\ &= \log\left(\prod_{i=1}^{n} L(\theta_i; x_{i\bullet})\right) \\ &= \sum_{i=1}^{n} \log L(\theta_i; x_{i\bullet}) \\ &= \sum_{i=1}^{n} \ell(\theta_i; x_{i\bullet}) \\ &= \sum_{i=1}^{n} m_i(\bar{x}_{i\bullet}\log\theta_i - \theta_i).\end{aligned} \tag{1.5}$$

We can derive the maximum likelihood estimate (MLE) for θ_i by differentiating Equation 1.4 with respect to θ_i, setting the result equal to 0, and solving for θ_i. This gives the nice result that the MLE is simply equal to the mean of the sample of data $X_{i\bullet}$. That is,

$$\hat{\theta}_i = \bar{X}_{i\bullet}. \tag{1.6}$$

Similarly, the MLE for the full parameter θ is just the vector of MLEs for its individual components:

$$\hat{\theta} \equiv (\hat{\theta}_1, ..., \hat{\theta}_n) = (\bar{X}_{1\bullet}, ..., \bar{X}_{n\bullet}). \tag{1.7}$$

2 Pseudolikelihoods

Let $\Theta \subseteq \mathbb{R}_+^n$ represent the space of possible values for θ and suppose we have a real-valued *parameter of interest* $\psi = \tau(\theta)$, where $\tau : \Theta \to \Psi$ is a known function with at least two continuous derivatives. Though it is not strictly necessary, in order to align with the tendency of researchers to focus on one-dimensional summaries of vector quantities we will assume for our purposes that ψ is a scalar, i.e. $\Psi \subseteq \mathbb{R}$.

This reduced dimension of Ψ relative to Θ implies the existence of a *nuisance parameter* $\lambda \in \Lambda \subseteq \mathbb{R}^{n-1}$. As its name suggests, λ tends to obfuscate or outright preclude inference regarding ψ and typically must be eliminated from the likelihood before proceeding. The product of this elimination is called a *pseudolikelihood function*. Any function of the data and ψ alone could theoretically be considered a pseudolikelihood, though course in practice some are more useful than others. If we let $\Theta_\psi = \{\theta \in \Theta : \tau(\theta) = \psi\}$, then associated with each

$\psi \in \Psi$ is the set of likelihood values $\mathcal{L}_\psi = \{L(\theta) : \theta \in \Theta_\psi\}$. For a given value of ψ, there may exist multiple corresponding values of λ.

We can construct pseudolikelihoods for ψ through clever choices by which to summarize $\mathcal{L}_\psi$ over all possible values of λ. Among the most popular methods of summary are profiling (i.e. maximization), conditioning, and integration, each with respect to the nuisance parameter. These summaries do come at a cost, however; eliminating a model's nuisance parameter from its likelihood almost always sacrifices some information about its parameter of interest as well. One measure of a good pseudolikelihood, therefore, is the balance it strikes between the amount of information it retains about ψ and the ease with which it can be computed.

2.1 The Profile Likelihood

The most straightforward method we can use to construct a pseudolikelihood (or equivalently, a pseudo-log-likelihood) function for ψ is usually to find the maximum of $\ell(\theta)$ over all possible of values of θ for each value of ψ. This yields what is known as the *profile* log-likelihood function, formally defined as

$$\ell_p(\psi) = \sup_{\theta \in \Theta:\, \tau(\theta)=\psi} \ell(\theta), \;\; \psi \in \Psi. \tag{2.1}$$

Suppose that ψ and λ are both explicit model parameters, with associated parameter spaces Ψ and Λ, respectively, so that the full parameter θ may be decomposed as $\theta = (\psi, \lambda)$. Assuming our model of choice satisfies the appropriate regularity conditions (see the appendix for further discussion), it will be possible to derive what is called the *partial-MLE for λ given a particular value of ψ*. This estimator, which we denote using the symbol $\widehat{\lambda}_\psi$, will be a function of both the data and ψ.[1] By "a particular value of ψ", we mean that for any specific value of ψ that we choose, say ψ_1, the corresponding value $\widehat{\lambda}_{\psi_1}$ will be the value of λ that maximizes the function $\ell(\psi_1, \lambda)$ over all possible values of λ. Formally,

$$\widehat{\lambda}_\psi = \arg\max_{\lambda \in \Lambda} \; \ell(\psi, \lambda). \tag{2.2}$$

For models in which it exists and has a closed form expression, $\widehat{\lambda}_\psi$ can be used to derive the model's profile log-likelihood. Indeed, doing so is almost trivial as the task of solving Equation 2.1 reduces simply to plugging in $\widehat{\lambda}_\psi$ for λ in the log-likelihood:

$$\ell_p(\psi) = \ell(\psi, \widehat{\lambda}_\psi). \tag{2.3}$$

Historically, the efficiency with which the profile is capable of producing accurate estimates of ψ relative to its ease of computation has made it the method of choice for statisticians when performing likelihood-based inference regarding a parameter of interest. Examples of profile-based statistics are the MLE for ψ, i.e.,

$$\widehat{\psi} = \arg\sup_{\psi \in \Psi} \ell_p(\psi), \tag{2.4}$$

and the signed likelihood ratio statistic for ψ, given by

$$R_\psi = \text{sgn}(\widehat{\psi} - \psi)(2(\ell_p(\widehat{\psi}) - \ell_p(\psi)))^{\frac{1}{2}}. \tag{2.5}$$

[1] This is in contrast to the regular MLE for λ which, if it exists, would be a function of the data alone.

2.2 The Integrated Likelihood

The *integrated likelihood* for ψ seeks to summarize $\mathcal{L}_\psi$ by its average value with respect to some weight function π over the space Θ_ψ. From a theoretical standpoint, this is preferable to the maximization procedure found in the profile likelihood as it naturally incorporates our uncertainty regarding the nuisance parameter's true value into the resulting pseudolikelihood. The general form of an integrated likelihood function is given

$$\bar{L}(\psi) = \int_{\Theta_\psi} L(\theta)\pi(\theta;\psi)d\theta. \tag{2.6}$$

It is up to the researcher to choose the weight function $\pi(\cdot;\psi)$, which plays an important role in the properties of the resulting integrated likelihood. Severini (2007) developed a method for re-parameterizing λ that makes the integrated likelihood relatively insensitive to the exact weight function chosen. Using this new parameterization, we have great flexibility in choosing our weight function; as long as it does not depend on the parameter of interest, the integrated likelihood that is produced will enjoy many desirable frequency properties.

3 Applications

We now turn our attention to the task of using the ZSE parameterization to construct an integrated likelihood that can be used to make inferences regarding a parameter of interest derived from the Poisson model described in the introduction. We will

3.1 Estimating the Weighted Sum of Poisson Means

Consider the weighted sum

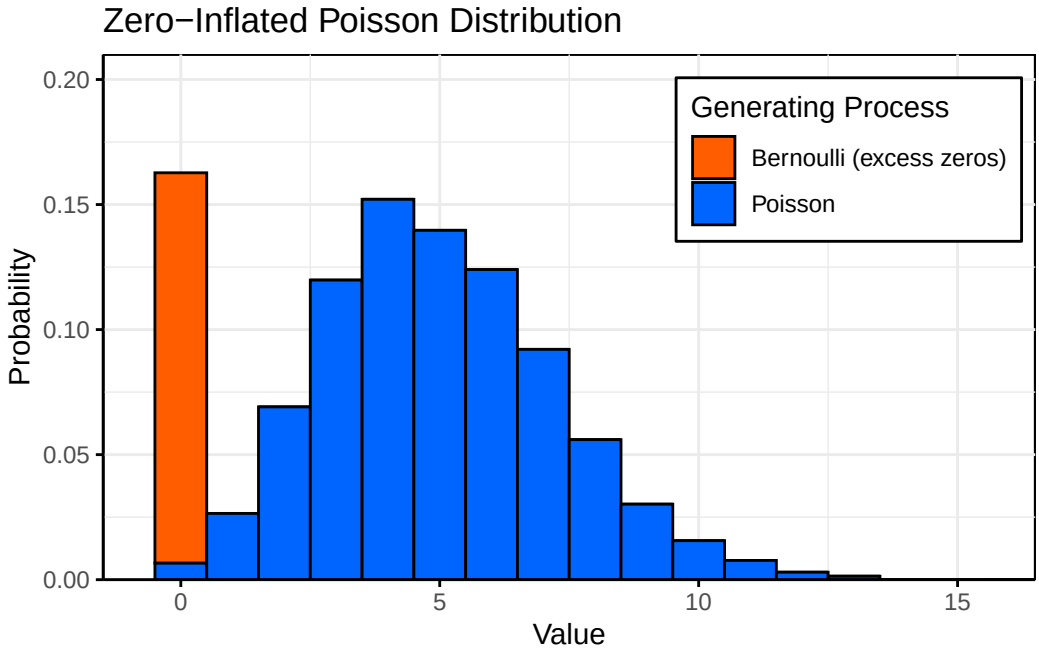
$$Y = \sum_{i=1}^{n} w_i X_i, \tag{3.1}$$

where each w_i is a known constant greater than zero. Suppose we take for our parameter of interest the expected value of this weighted sum, so that

$$\psi \equiv \mathrm{E}(Y) = \sum_{i=1}^{n} w_i \theta_i. \tag{3.2}$$

3.2 Zero-Inflated Poisson Regression

A sample of count data is called *zero-inflated* when it contains an excess amount of zero-valued observations. A common tactic to account for this excess is to model the data using a mixture of two processes, one that generates zeros and another that generates counts, some of which may also be zeros. When this count-generating process follows a Poisson distribution, we call the resulting mixture a zero-inflated Poisson (ZIP) model.



Let $U \sim \text{Bernoulli}(1-\rho)$ and $V \sim \text{Poisson}(\mu)$. Suppose U and V are independent and let $Y = UV$. Then $Y \sim \text{ZIP}(\mu, \rho)$. To derive its distribution, we begin by recognizing that $Y = 0$ when either $U = 0$ or $V = 0$ so that

$$\begin{aligned}
\mathbb{P}(Y=0) &= \mathbb{P}(U=0 \cup V=0) \\
&= \mathbb{P}(U=0) + \mathbb{P}(V=0) - \mathbb{P}(U=0 \cap V=0) \\
&= \mathbb{P}(U=0) + \mathbb{P}(V=0) - \mathbb{P}(U=0)\mathbb{P}(V=0) \\
&= \rho + e^{-\mu} - \rho e^{-\mu} \\
&= \rho + (1-\rho)e^{-\mu}.
\end{aligned} \tag{3.3}$$

In order for Y to take on a value $y > 0$, we must have $U = 1$ and $V = y$. That is,

$$\begin{aligned}
\mathbb{P}(Y=y) &= \mathbb{P}(U=1 \cap V=y) \\
&= \mathbb{P}(U=1)\mathbb{P}(V=y) \\
&= (1-\rho)\frac{e^{-\mu}\mu^y}{y!}, \quad y = 1, 2, ...
\end{aligned} \tag{3.4}$$

Thus, the full probability mass function (pmf) for a ZIP random variable is given by

$$\mathbb{P}(Y=y) = \begin{cases} \rho + (1-\rho)e^{-\mu}, & y = 0 \\ (1-\rho)\frac{e^{-\mu}\mu^y}{y!}, & y = 1, 2, ... \end{cases} \tag{3.5}$$

Alternatively, we may write

$$\mathbb{P}(Y=y) = \left(\rho + (1-\rho)e^{-\mu}\right)^{\mathbb{1}_{y=0}} \left((1-\rho)\frac{e^{-\mu}\mu^y}{y!}\right)^{1-\mathbb{1}_{y=0}}, \quad y = 0, 1, 2, ..., \tag{3.6}$$

where $\mathbb{1}_{y=0}$ denotes an indicator function that assumes the value one when $y = 0$ and zero otherwise.

Suppose we observe multiple counts $Y_1, ..., Y_n$ generated independently and identically from Y.[2] The likelihood function for μ and ρ based on an individual observation $Y_i = y_i$ is simply equal to the pmf for Y evaluated at y_i. That is,

$$L(\mu, \rho; y_i) = \Big(\rho + (1-\rho)e^{-\mu}\Big)^{\mathbb{1}_{y_i=0}} \left((1-\rho)\frac{e^{-\mu}\mu^{y_i}}{y_i!}\right)^{1-\mathbb{1}_{y_i=0}}, \;\; \mu > 0, \;\; \rho \in [0, 1). \tag{3.7}$$

We can safely ignore any multiplicative constant in L without influencing our inferences regarding μ and ρ. In particular, we can discard the term $\left(\frac{1}{y_i!}\right)^{1-\mathbb{1}_{y_i=0}}$ as it depends only on the observation y_i and is constant with respect to the model parameters. Hence, we may write

$$L(\mu, \rho; y_i) = \Big(\rho + (1-\rho)e^{-\mu}\Big)^{\mathbb{1}_{y_i=0}} \Big((1-\rho)e^{-\mu}\mu^{y_i}\Big)^{1-\mathbb{1}_{y_i=0}}, \;\; \mu > 0, \;\; \rho \in [0, 1). \tag{3.8}$$

From this we may derive the corresponding log-likelihood function:

$$\begin{aligned}
\ell(\mu, \rho; y_i) &= \log L(\mu, \rho; y_i) \\
&= \log\left\{\Big(\rho + (1-\rho)e^{-\mu}\Big)^{\mathbb{1}_{y_i=0}} \Big((1-\rho)e^{-\mu}\mu^{y_i}\Big)^{1-\mathbb{1}_{y_i=0}}\right\} \\
&= \log\left\{\Big(\rho + (1-\rho)e^{-\mu}\Big)^{\mathbb{1}_{y_i=0}}\right\} + \log\left\{\Big((1-\rho)e^{-\mu}\mu^{y_i}\Big)^{1-\mathbb{1}_{y_i=0}}\right\} \\
&= \mathbb{1}_{y_i=0}\log\Big(\rho + (1-\rho)e^{-\mu}\Big) + (1-\mathbb{1}_{y_i=0})\log\Big((1-\rho)e^{-\mu}\mu^{y_i}\Big) \\
&= \mathbb{1}_{y_i=0}\log\Big(\rho + (1-\rho)e^{-\mu}\Big) + (1-\mathbb{1}_{y_i=0})\Big(\log(1-\rho) - \mu + y_i\log\mu\Big).
\end{aligned} \tag{3.9}$$

The log-likelihood for the full sample $y_\bullet \equiv (y_1, ..., y_n)$ is then obtained by summing over the

[2] Recall that the parameter of a generic Poisson random variable is defined relative to a fixed length of time during which observations may be recorded and added to the running total. By assuming that all counts come from the same ZIP-distributed random variable, we are implicitly assuming that each count was recorded over the same period of time and thus are on equal footing with one another.

index variable i as follows:

$$\begin{aligned}
\ell(\mu, \rho; y_\bullet) &= \sum_{i=1}^{n} \ell(\mu, \rho; y_i) \\
&= \sum_{i=1}^{n} \left[\mathbb{1}_{y_i=0} \log\Big(\rho + (1-\rho)e^{-\mu}\Big) + (1 - \mathbb{1}_{y_i=0})\Big(\log(1-\rho) - \mu + y_i \log\mu \Big) \right] \\
&= \sum_{i=1}^{n} \mathbb{1}_{y_i=0} \log\Big(\rho + (1-\rho)e^{-\mu}\Big) + \sum_{i=1}^{n} (1 - \mathbb{1}_{y_i=0})\Big(\log(1-\rho) - \mu + y_i \log\mu \Big) \\
&= \log\Big(\rho + (1-\rho)e^{-\mu}\Big) \underbrace{\sum_{i=1}^{n} \mathbb{1}_{y_i=0}}_{\text{A}} + \big(\log(1-\rho) - \mu \big) \underbrace{\sum_{i=1}^{n} (1 - \mathbb{1}_{y_i=0})}_{\text{B}} + \log\mu \underbrace{\sum_{i=1}^{n} (1 - \mathbb{1}_{y_i=0}) y_i}_{\text{C}}.
\end{aligned} \tag{3.10}$$

The summation in A is counting the number of zero counts in the sample. Let $\bar{\rho}$ represent the proportion of these observed zero counts in the sample so that

$$n\bar{\rho} \equiv \sum_{i=1}^{n} \mathbb{1}_{y_i=0}. \tag{3.11}$$

Similarly, the summation in B is counting the number of nonzero counts in the sample. Since $1 - \bar{\rho}$ represents the proportion of the observed nonzero counts, it follows that

$$n(1-\bar{\rho}) \equiv \sum_{i=1}^{n} (1 - \mathbb{1}_{y_i=0}). \tag{3.12}$$

Finally, consider the summation in C. Whenever $y_i = 0$, $(1 - \mathbb{1}_{y_i=0}) y_i = (1-1) \cdot 0 = 0 = y_i$. Whenever $y_i > 0$, $(1 - \mathbb{1}_{y_i=0}) y_i = (1-0) y_i = y_i$. It follows that $(1 - \mathbb{1}_{y_i=0}) y_i = y_i$ for all values of y_i. Hence, the summation is simply adding all the observed counts in the sample. Let $\bar{y}$ denote the sample mean so that

$$n\bar{y} \equiv \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} (1 - \mathbb{1}_{y_i=0}) y_i \tag{3.13}$$

Substituting these three expressions on the left in Equation 3.11, Equation 3.12, and Equation 3.13 for their corresponding summations in the final line of Equation 3.10, we arrive at

$$\ell(\mu, \rho; y_\bullet) = n\bar{\rho} \log\Big(\rho + (1-\rho)e^{-\mu}\Big) + n(1-\bar{\rho})\big(\log(1-\rho) - \mu \big) + n\bar{y} \log\mu. \tag{3.14}$$

3.2.1 Profile Likelihood

Suppose that μ is our parameter of interest, and ρ is a nuisance parameter. The partial-MLE for ρ given a particular value of μ, as defined in Equation 2.2, is given by

$$\hat{\rho}_\mu = \begin{cases} \frac{\bar{\rho} - e^{-\mu}}{1 - e^{-\mu}}, & \bar{\rho} \geq e^{-\mu} \\ 0, & \bar{\rho} < e^{-\mu} \end{cases} \tag{3.15}$$

See Appendix A for a formal proof of this statement. Since ρ represents the probability of excess zero counts in the sample, it is unsurprising that our estimate of ρ when $\bar{\rho} \geq e^{-\mu}$ turns out simply to be 0.[3] Indeed, the observation that $\bar{\rho} \geq e^{-\mu}$ could be considered a bellwether for zero-inflation, as it indicates an empirical excess of zero counts relative to what we would expect for a non-zero-inflated Poisson random variable. Hence, we refer to it as the *zero-inflated condition*. Assuming no zero-inflation is present, the MLE of μ is the sample mean $\bar{y}$. When μ is unknown, as is typically the case, a sensible choice is therefore to replace it with $\bar{y}$ so that the veracity of $\bar{\rho} \geq e^{-\bar{y}}$ becomes our test of zero-inflation.

Assume that our observed sample of counts $y_1, ..., y_n$ satisfies the zero-inflated condition, so that $\hat{\rho}_\mu = \frac{\bar{\rho} - e^{-\mu}}{1 - e^{-\mu}}$. This implies

$$1 - \hat{\rho}_\mu = \frac{1 - e^{-\mu}}{1 - e^{-\mu}} - \frac{\bar{\rho} - e^{-\mu}}{1 - e^{-\mu}} = \frac{1 - e^{-\mu} - \bar{\rho} + e^{-\mu}}{1 - e^{-\mu}} = \frac{1 - \bar{\rho}}{1 - e^{-\mu}}. \tag{3.16}$$

and

$$\hat{\rho}_\mu + (1 - \hat{\rho}_\mu)e^{-\mu} = \frac{\bar{\rho} - e^{-\mu}}{1 - e^{-\mu}} + \frac{1 - \bar{\rho}}{1 - e^{-\mu}} e^{-\mu} = \frac{\bar{\rho}(1 - e^{-\mu})}{1 - e^{-\mu}} = \bar{\rho}. \tag{3.17}$$

Per Equation 2.3, the profile log-likelihood for μ can be obtained by plugging this partial-MLE for μ into the full log-likelihood:

$$\begin{aligned}
\ell_p(\mu) &= \ell(\mu, \hat{\rho}_\mu) \\
&= n\bar{\rho} \log\left(\hat{\rho}_\mu + (1 - \hat{\rho}_\mu)e^{-\mu}\right) + n(1 - \bar{\rho})(\log(1 - \hat{\rho}_\mu) - \mu) + n\bar{y} \log \mu \\
&= n\bar{\rho} \log(\bar{\rho}) + n(1 - \bar{\rho})\left(\log\left(\frac{1 - \bar{\rho}}{1 - e^{-\mu}}\right) - \mu\right) + n\bar{y} \log \mu \\
&= n\Big[\bar{y} \log \mu - (1 - \bar{\rho})(\log(1 - e^{-\mu}) + \mu)\Big] + n\bar{\rho} \log(\bar{\rho}) + n(1 - \bar{\rho}) \log(1 - \bar{\rho}).
\end{aligned} \tag{3.18}$$

As usual, we can discard the terms not depending on μ, yielding

$$\ell_p(\mu) = n\Big[\bar{y} \log \mu - (1 - \bar{\rho})(\log(1 - e^{-\mu}) + \mu)\Big]. \tag{3.19}$$

3.2.2 Integrated Likelihood

$$\begin{aligned}
L(\mu, \rho; y_\bullet) &= \exp\Big\{\ell(\mu, \rho; y_\bullet)\Big\} \\
&= \exp\Big\{n\bar{\rho} \log\left(\rho + (1 - \rho)e^{-\mu}\right) + n(1 - \bar{\rho})(\log(1 - \rho) - \mu) + n\bar{y} \log \mu\Big\} \\
&= \Big[\rho + (1 - \rho)e^{-\mu}\Big]^{n\bar{\rho}} \Big[(1 - \rho)e^{-\mu}\Big]^{n(1-\bar{\rho})} \mu^{n\bar{y}}.
\end{aligned}$$

$$\frac{\partial \ell(\mu, \pi)}{\partial \pi} = n\left[\frac{\bar{\pi}(1 - e^{-\mu})}{\pi + (1 - \pi)e^{-\mu}} - \frac{1 - \bar{\pi}}{1 - \pi}\right].$$

[3] Recall that $e^{-\mu}$ is the probability of observing a zero under a non-zero-inflated Poisson random variable with rate parameter μ.

$$
\begin{aligned}
\mathrm{E}(\bar{\rho}) &= \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{Y_i=0}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(\mathbb{1}_{Y_i=0}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathrm{P}(Y_i=0) \\
&= \frac{1}{n}\sum_{i=1}^{n}\rho+(1-\rho)e^{-\mu} \\
&= \frac{1}{n}\cdot n\cdot(\rho+(1-\rho)e^{-\mu}) \\
&= \rho+(1-\rho)e^{-\mu}.
\end{aligned}
$$

$$
\mathrm{E}(\bar{\rho};\hat{\mu},\phi) \equiv \mathrm{E}(\bar{\rho};\mu_0,\rho_0)\Big|_{(\mu_0,\rho_0)=(\hat{\mu},\phi)} = \phi+(1-\phi)e^{-\hat{\mu}}.
$$

4 Importance Sampling

Let $p(\theta)$ denote a prior distribution for a parameter $\theta \in \Theta \subseteq \mathbb{R}^d$ and $L(\theta;X)$ the likelihood function of our model based on data X. The posterior distribution for θ is given by $\pi(\theta|X) = cL(\theta;X)p(\theta)$, where $c = \left(\int_\Theta L(\theta;X)p(\theta)d\theta\right)^{-1} < \infty$. Suppose we have another function $f(\theta) > 0$ for all $\theta \in \Theta$ and we are interested in estimating the expectation of this function with respect to the distribution of p. Call this value μ. Then we have

$$
\begin{aligned}
\mu &= \mathrm{E}_p(f(\theta)) \\
&= \int_\Theta f(\theta)p(\theta)d\theta \\
&= \int_\Theta \frac{f(\theta)}{cL(\theta;X)}cL(\theta;X)p(\theta)d\theta \\
&= \int_\Theta \frac{f(\theta)}{cL(\theta;X)}\pi(\theta|X)d\theta \\
&= \mathrm{E}_\pi\left(\frac{f(\theta)}{cL(\theta;X)}\right).
\end{aligned}
$$

The *importance sampling estimator* for μ is

$$
\hat{\mu}_\pi = \frac{1}{R}\sum_{i=1}^{R}\frac{f(\theta_i)}{cL(\theta_i;X)},\ \ \theta_i \sim \pi.
$$

Note that $\widehat{\mu}_\pi$ is unbiased, i.e.

$$\begin{aligned}
\mathrm{E}_\pi(\widehat{\mu}_\pi) &= \mathrm{E}_\pi\left(\frac{1}{R}\sum_{i=1}^{R}\frac{f(\theta_i)}{cL(\theta_i;X)}\right) \\
&= \frac{1}{R}\sum_{i=1}^{R}\mathrm{E}_\pi\left(\frac{f(\theta_i)}{cL(\theta_i;X)}\right) \\
&= \frac{1}{R}\sum_{i=1}^{R}\mu \\
&= \frac{1}{R}R\mu \\
&= \mu,
\end{aligned}$$

and by the law of large numbers converges in distribution to μ, i.e.

$$\widehat{\mu}_\pi \to \mu \text{ as } R \to \infty.$$

The variance of $\widehat{\mu}_\pi$ is given by

$$\begin{aligned}
\mathrm{Var}_\pi(\widehat{\mu}_\pi) &= \mathrm{Var}_\pi\left(\frac{1}{R}\sum_{i=1}^{R}\frac{f(\theta_i)}{cL(\theta_i;X)}\right) \\
&= \frac{1}{R^2}\sum_{i=1}^{R}\mathrm{Var}_\pi\left(\frac{f(\theta_i)}{cL(\theta_i;X)}\right) \\
&= \frac{1}{R^2}\sum_{i=1}^{R}\mathrm{Var}_\pi\left(\frac{f(\theta)}{cL(\theta;X)}\right) \\
&= \frac{1}{R^2}R\cdot\mathrm{Var}_\pi\left(\frac{f(\theta)}{cL(\theta;X)}\right) \\
&= \frac{1}{R}\mathrm{Var}_\pi\left(\frac{f(\theta)}{cL(\theta;X)}\right) \\
&= \frac{1}{R}\left\{\mathrm{E}_\pi\left[\left(\frac{f(\theta)}{cL(\theta;X)}\right)^2\right] - \left[\mathrm{E}_\pi\left(\frac{f(\theta)}{cL(\theta;X)}\right)\right]^2\right\} \\
&= \frac{1}{R}\left\{\int_\Theta\left(\frac{f(\theta)}{cL(\theta;X)}\right)^2\pi(\theta|X)d\theta - \mu^2\right\} \\
&= \frac{1}{R}\left\{\int_\Theta\frac{f(\theta)^2}{c^2L(\theta;X)^2}cL(\theta;X)p(\theta)d\theta - \mu^2\right\} \\
&= \frac{1}{R}\left\{\int_\Theta\frac{f(\theta)^2p(\theta)}{cL(\theta;X)}d\theta - \mu^2\right\} \\
&= \frac{1}{R}\left\{\int_\Theta\frac{f(\theta)^2p(\theta)^2}{cL(\theta;X)p(\theta)}d\theta - \mu^2\right\} \\
&= \frac{1}{R}\left\{\int_\Theta\frac{\left(f(\theta)p(\theta)\right)^2}{\pi(\theta|X)}d\theta - \mu^2\right\} \\
&= \frac{\sigma_\pi^2}{R},
\end{aligned}$$

where

$$\sigma_\pi^2 = \int_\Theta\frac{\left(f(\theta)p(\theta)\right)^2}{\pi(\theta|X)}d\theta - \mu^2.$$

Some clever rearranging and substituting allows us to rewrite it as

$$\begin{aligned}
\sigma_\pi^2 &= \int_\Theta \frac{\left(f(\theta)p(\theta)\right)^2}{\pi(\theta|X)}d\theta - \mu^2 \\
&= \int_\Theta \frac{\left(f(\theta)p(\theta)\right)^2}{\pi(\theta|X)}d\theta - 2\mu^2 + \mu^2 \\
&= \int_\Theta \frac{\left(f(\theta)p(\theta)\right)^2}{\pi(\theta|X)}d\theta - 2\mu\int_\Theta f(\theta)p(\theta)d\theta + \mu^2\int_\Theta \pi(\theta|X)d\theta \\
&= \int_\Theta \left(\frac{\left(f(\theta)p(\theta)\right)^2}{\pi(\theta|X)} - 2\mu f(\theta)p(\theta) + \mu^2\pi(\theta|X)\right)d\theta \\
&= \int_\Theta \frac{\left(f(\theta)p(\theta)\right)^2 - 2\mu f(\theta)p(\theta)\pi(\theta|X) + \mu^2\pi(\theta|X)^2}{\pi(\theta|X)}d\theta \\
&= \int_\Theta \frac{\left(f(\theta)p(\theta) - \mu\pi(\theta|X)\right)^2}{\pi(\theta|X)}d\theta.
\end{aligned}$$

We can also write

$$\begin{aligned}
\sigma_\pi^2 &= \int_\Theta \frac{\left(f(\theta)p(\theta) - \mu\pi(\theta|X)\right)^2}{\pi(\theta|X)}d\theta \\
&= \int_\Theta \left(\frac{f(\theta)p(\theta) - \mu\pi(\theta|X)}{\pi(\theta|X)}\right)^2 \pi(\theta|X)d\theta \\
&= \mathrm{E}_\pi\left[\left(\frac{f(\theta)p(\theta) - \mu\pi(\theta|X)}{\pi(\theta|X)}\right)^2\right].
\end{aligned}$$

Because the θ_i are sampled from π, the natural variance estimate is

$$\widehat{\sigma}_\pi^2 = \frac{1}{R}\sum_{i=1}^R \left(\frac{f(\theta_i)}{cL(\theta_i;X)} - \widehat{\mu}_\pi\right)^2 = \frac{1}{R}\sum_{i=1}^R (w_i f(\theta_i) - \widehat{\mu}_\pi)^2,$$

where $w_i = \frac{1}{cL(\theta_i;X)}$.

$$\begin{aligned}
\sigma_\pi^2 + \mu &= \int_\Theta \frac{\left(f(\theta)p(\theta)\right)^2}{\pi(\theta|X)}d\theta \\
&= \int_\Theta \frac{\left(f(\theta)p(\theta)\right)^2}{cL(\theta;X)p(\theta)}d\theta \\
&= \int_\Theta \frac{f(\theta)^2}{cL(\theta;X)}p(\theta)d\theta \\
&= \mathrm{E}_p\left(\frac{f(\theta)^2}{cL(\theta;X)}\right) \\
&= \mathrm{E}_\pi\left(\frac{f(\theta)^2}{c^2L(\theta;X)^2}\right).
\end{aligned}$$

4.1 Self-normalized importance sampling

$\pi(\theta|X) = cL(\theta; X)p(\theta)$, $c > 0$ unknown. $p_u(\theta) = ap(\theta)$, $a > 0$ unknown. $p_u(\theta) = bp(\theta)$, $a > 0$ unknown.

$$\tilde{\mu}_\pi =$$

References

1. Beckett S, Jee J, Ncube T, et al (2014) Zero-inflated Poisson (ZIP) distribution: Parameter estimation and applications to model data from natural calamities. Involve 7(6):751–767. https://doi.org/10.2140/involve.2014.7.751

2. Dencks S, Piepenbrock M, Schmitz G (2020) Assessing Vessel Reconstruction in Ultrasound Localization Microscopy by Maximum Likelihood Estimation of a Zero-Inflated Poisson Model. IEEE Trans Ultrason, Ferroelect, Freq Contr 67(8):1603–1612. https://doi.org/10.1109/TUFFC.2020.2980063

3. Feng CX (2021) A comparison of zero-inflated and hurdle models for modeling zero-inflated count data. J Stat Distrib App 8(1):8. https://doi.org/10.1186/s40488-021-00121-4

4. Schwartz J, Giles DE (2016) Bias-reduced maximum likelihood estimation of the zero-inflated Poisson distribution. Communications in Statistics - Theory and Methods 45(2):465–478. https://doi.org/10.1080/03610926.2013.824590

5. Severini TA (2007) Integrated likelihood functions for non-Bayesian inference. Biometrika 94(3):529–542. https://doi.org/10.1093/biomet/asm040

6. Severini TA (2018) Integrated likelihoods for functions of a parameter. Stat 7(1):e212. https://doi.org/10.1002/sta4.212

A Theorems

Theorem 1 *Suppose we observe a sample of counts $y_1, ..., y_n$ independently and identically distributed according to a zero-inflated Poisson random variable with parameters μ and ρ, where $\mu > 0$ is the expected Poisson count and $\rho \in [0, 1)$ is the probability of excess zero counts. Let $\bar{\rho} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{y_i=0}$ represent the proportion of observed zero counts in the sample. Then the partial-MLE for ρ given a particular value of μ is given by*

$$\hat{\rho}_\mu = \begin{cases} \frac{\bar{\rho}-e^{-\mu}}{1-e^{-\mu}}, & e^{-\mu} \leq \bar{\rho} \leq 1 \\ 0, & 0 \leq \bar{\rho} < e^{-\mu} \end{cases}$$

Proof *The log-likelihood function of the model is*

$$\ell(\mu, \rho) = n\bar{\rho} \log\Big(\rho + (1-\rho)e^{-\mu}\Big) + n(1-\bar{\rho})(\log(1-\rho) - \mu) + n\bar{y} \log \mu;\ \mu > 0,\ \rho \in [0, 1),$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ denotes the sample mean. This definition artificially restricts the domain of the input ρ to the interval $[0, 1)$ to reflect the interpretation of ρ as a probability. Since there is no way to interpret a complex-valued log-likelihood, the $\log(1-\rho)$ expression in the second term above explicitly prohibits values of $\rho \geq 1$ from being passed to ℓ. However, it is still possible to plug in some values of $\rho < 0$ to ℓ and obtain a real number. All that is needed to ensure ℓ remains real-valued is for the $\rho + (1-\rho)e^{-\mu}$ expression inside the logarithm in the first term of the formula to remain positive; there are some negative values of ρ which satisfy this condition.

Claim 1.1 *$\rho + (1-\rho)e^{-\mu} > 0$ if and only if $\rho > \frac{1}{1-e^{\mu}}$.*

Proof of Claim 1.1 *Note that $\mu > 0 \implies 0 < e^{-\mu} < 1$. Then we have*

$$\begin{aligned} \rho + (1-\rho)e^{-\mu} > 0 &\iff \rho - \rho e^{-\mu} + e^{-\mu} > 0 \\ &\iff \rho(1 - e^{-\mu}) + e^{-\mu} > 0 \\ &\iff \rho > -\frac{e^{-\mu}}{1-e^{-\mu}} \\ &\iff \rho > \frac{e^{-\mu}}{e^{-\mu}-1} \\ &\iff \rho > \frac{e^{-\mu}}{e^{-\mu}-1} \cdot \frac{e^{\mu}}{e^{\mu}} \\ &\iff \rho > \frac{1}{1-e^{\mu}}. \end{aligned}$$

Define the function l^* by extending ℓ to include values of ρ in the interval $\Big(\frac{1}{1-e^{\mu}},\ 1\Big)$, i.e.

$$\ell^*(\mu, \rho) = n\bar{\rho} \log\Big(\rho + (1-\rho)e^{-\mu}\Big) + n(1-\bar{\rho})(\log(1-\rho) - \mu) + n\bar{y} \log \mu;\ \mu > 0,\ \frac{1}{1-e^{\mu}} < \rho < 1.$$

Claim 1.2 *For a given value of μ, the value of ρ that maximizes $\ell^*(\mu, \rho)$ is $\frac{\bar{\rho}-e^{-\mu}}{1-e^{-\mu}}$.*

Proof of Claim 1.2 *Let $\widehat{\rho}^*_\mu$ denote the the value of ρ that maximizes $\ell^*(\mu,\rho)$ for a given value of μ. Under suitable regularity conditions, easily satisfied in this case since the Poisson distribution belongs to the well-behaved exponential family of distributions, $\widehat{\rho}^*_\mu$ is guaranteed to exist and will be the unique value of ρ (as a function of μ and the data) that solves the critical point equation*

$$\left.\frac{\partial \ell^*(\mu,\rho)}{\partial \rho}\right|_{\rho=\widehat{\rho}^*_\mu} \equiv 0.$$

Hence, we must differentiate ℓ^* with respect to ρ, evaluate the derivative at $\rho = \widehat{\rho}^*_\mu$, and set the result equal to 0. Doing so and then solving for $\widehat{\rho}^*_\mu$ yields

$$0 \equiv \left.\frac{\partial \ell^*(\mu,\rho)}{\partial \rho}\right|_{\rho=\widehat{\rho}^*_\mu} = \frac{n\bar{\rho}}{\widehat{\rho}^*_\mu + (1-\widehat{\rho}^*_\mu)e^{-\mu}}(1-e^{-\mu}) - \frac{n(1-\bar{\rho})}{1-\widehat{\rho}^*_\mu} = n\left[\frac{\bar{\rho}(1-e^{-\mu})}{\widehat{\rho}^*_\mu + (1-\widehat{\rho}^*_\mu)e^{-\mu}} - \frac{1-\bar{\rho}}{1-\widehat{\rho}^*_\mu}\right].$$

$$\begin{aligned}
&\implies \frac{\bar{\rho}(1-e^{-\mu})}{\widehat{\rho}^*_\mu + (1-\widehat{\rho}^*_\mu)e^{-\mu}} - \frac{1-\bar{\rho}}{1-\widehat{\rho}^*_\mu} = 0 \\
&\implies \frac{\bar{\rho}(1-e^{-\mu})}{\widehat{\rho}^*_\mu + (1-\widehat{\rho}^*_\mu)e^{-\mu}} = \frac{1-\bar{\rho}}{1-\widehat{\rho}^*_\mu} \\
&\implies \frac{(1-\widehat{\rho}^*_\mu)(1-e^{-\mu})}{\widehat{\rho}^*_\mu + (1-\widehat{\rho}^*_\mu)e^{-\mu}} = \frac{1-\bar{\rho}}{\bar{\rho}} \\
&\implies \frac{\widehat{\rho}^*_\mu + (1-\widehat{\rho}^*_\mu)e^{-\mu}}{(1-\widehat{\rho}^*_\mu)(1-e^{-\mu})} = \frac{\bar{\rho}}{1-\bar{\rho}} \\
&\implies \frac{\widehat{\rho}^*_\mu}{(1-\widehat{\rho}^*_\mu)(1-e^{-\mu})} + \frac{(1-\widehat{\rho}^*_\mu)e^{-\mu}}{(1-\widehat{\rho}^*_\mu)(1-e^{-\mu})} = \frac{\bar{\rho}}{1-\bar{\rho}} \\
&\implies \frac{\widehat{\rho}^*_\mu}{(1-\widehat{\rho}^*_\mu)(1-e^{-\mu})} = \frac{\bar{\rho}}{1-\bar{\rho}} - \frac{e^{-\mu}}{1-e^{-\mu}} \\
&\implies \frac{\widehat{\rho}^*_\mu}{1-\widehat{\rho}^*_\mu} = \frac{\bar{\rho}}{1-\bar{\rho}}(1-e^{-\mu}) - e^{-\mu} \\
&\implies \widehat{\rho}^*_\mu = \frac{\frac{\bar{\rho}}{1-\bar{\rho}}(1-e^{-\mu}) - e^{-\mu}}{\frac{\bar{\rho}}{1-\bar{\rho}}(1-e^{-\mu}) - e^{-\mu} + 1} \qquad \left(\frac{A}{1-A} = B \iff A = \frac{B}{B+1}\right) \\
&\implies \widehat{\rho}^*_\mu = \frac{\frac{\bar{\rho}}{1-\bar{\rho}}(1-e^{-\mu}) - e^{-\mu}}{\frac{\bar{\rho}}{1-\bar{\rho}}(1-e^{-\mu}) + (1-e^{-\mu})} \\
&\implies \widehat{\rho}^*_\mu = \frac{\frac{\bar{\rho}}{1-\bar{\rho}}(1-e^{-\mu}) - e^{-\mu}}{(\frac{\bar{\rho}}{1-\bar{\rho}} + 1)(1-e^{-\mu})} \\
&\implies \widehat{\rho}^*_\mu = \frac{\frac{\bar{\rho}}{1-\bar{\rho}}(1-e^{-\mu}) - e^{-\mu}}{\frac{1}{1-\bar{\rho}}(1-e^{-\mu})} \\
&\implies \widehat{\rho}^*_\mu = \frac{\bar{\rho}(1-e^{-\mu}) - (1-\bar{\rho})e^{-\mu}}{1-e^{-\mu}} \\
&\implies \widehat{\rho}^*_\mu = \frac{\bar{\rho} - e^{-\mu}}{1-e^{-\mu}}.
\end{aligned}$$

Claim 1.3 *For a given value of μ, if $\bar{\rho} \geq e^{-\mu}$, the value of ρ that maximizes $\ell(\mu, \rho)$ is $\widehat{\rho}^*_\mu = \frac{\bar{\rho} - e^{-\mu}}{1 - e^{-\mu}}$.*

Proof of Claim 1.3 *Note that $\bar{\rho} \geq e^{-\mu} \implies \widehat{\rho}^*_\mu \geq 0$. Since $\ell(\mu, \rho) = \ell^*(\mu, \rho)$ for any value of $\rho \in [0, 1)$, the claim follows from the previous one.*

Claim 1.4 *$\rho + (1 - \rho)e^{-\mu} \geq e^{-\mu}$ for all $\rho \in [0, 1)$ and $\mu > 0$.*

Proof of Claim 1.4 *Let $h(\rho) = \rho + (1 - \rho)e^{-\mu}$ for $\rho \in [0, 1)$. Note that $h'(\rho) = 1 - e^{-\mu}$ for all values of ρ. $\mu > 0 \implies e^{-\mu} > 1 \implies 1 - e^{-\mu} > 0$. Thus, $h'(\rho) > 0$ and so $h(\rho)$ is a strictly increasing function everywhere in its domain. It follows that $h(\rho) \geq h(0) = e^{-\mu}$ for all $\rho \in [0, 1)$ and $\mu > 0$.*

Claim 1.5 *When $\bar{\rho} < e^{-\mu}$, $\ell(\mu, \rho)$ is strictly decreasing with respect to ρ for a fixed value of μ.*

Proof of Claim 1.5 *The derivative of ℓ with respect to ρ is given by*

$$\frac{\partial \ell(\mu, \rho)}{\partial \rho} = n \left[\frac{\bar{\rho}(1 - e^{-\mu})}{\rho + (1 - \rho)e^{-\mu}} - \frac{1 - \bar{\rho}}{1 - \rho} \right].$$

Note that

$$\begin{aligned}
\frac{\bar{\rho}(1 - e^{-\mu})}{\rho + (1 - \rho)e^{-\mu}} - \frac{1 - \bar{\rho}}{1 - \rho} &= \frac{\bar{\rho}(1 - e^{-\mu})}{\rho + (1 - \rho)e^{-\mu}} + \frac{\bar{\rho} - 1}{1 - \rho} \\
&< \frac{e^{-\mu}(1 - e^{-\mu})}{\rho + (1 - \rho)e^{-\mu}} + \frac{e^{-\mu} - 1}{1 - \rho} \quad \textit{(by assumption)} \\
&\leq \frac{e^{-\mu}(1 - e^{-\mu})}{e^{-\mu}} + \frac{e^{-\mu} - 1}{1 - \rho} \quad \textit{(by Claim 1.4)} \\
&= 1 - e^{-\mu} + \frac{e^{-\mu} - 1}{1 - \rho} \\
&= (1 - e^{-\mu}) \left(1 - \frac{1}{1 - \rho} \right) \\
&= (1 - e^{-\mu}) \left(- \frac{\rho}{1 - \rho} \right) \\
&\leq 0.
\end{aligned}$$

Hence, $\frac{\partial \ell(\mu, \rho)}{\partial \rho} < 0$ for all $\rho \in [0, 1)$ when $\bar{\rho} < e^{-\mu}$, and so the claim is proved.

Claim 1.6 *When $\bar{\rho} < e^{-\mu}$, the value of ρ that maximizes $\ell(\mu, \rho)$ for a given value of μ is 0.*

Proof of Claim 1.6 *Assume $\bar{\rho} < e^{-\mu}$. This implies $\widehat{\rho}^*_\mu < 0$ and therefore is not a valid input for ρ to $\ell(\mu, \rho)$. By Claim 1.5, $\ell(\mu, \rho)$ is strictly decreasing with respect to ρ for a fixed value of μ. Thus, $\ell(\mu, 0) \geq \ell(\mu, \rho)$ for all $\rho \in [0, 1)$ with equality if and only if $\rho = 0$. Hence, $\rho = 0$ maximizes $\ell(\mu, \rho)$ for a given value of μ, and so the claim is proved.*

Together, Claim 1.3 and Claim 1.6 establish the result in the theorem.

Theorem 2 *Suppose we observe a sample of counts $y_1, ..., y_n$ independently and identically distributed according to a zero-inflated Poisson random variable with parameters μ and ρ, where $\mu > 0$ is the expected Poisson count and $\rho \in [0, 1)$ is the probability of excess zero counts. Let $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ denote the sample mean, $\bar{\rho} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{y_i=0}$ the proportion of observed zero counts in the sample, and $\gamma = \frac{\bar{y}}{1-\bar{\rho}}$. Then the maximum likelihood estimators for μ and ρ are given by*

$$\hat{\mu} = \begin{cases} W_0(-\gamma e^{-\gamma}) + \gamma, & \bar{\rho} > 0 \\ \bar{y}, & \bar{\rho} = 0 \end{cases}$$

and

$$\hat{\rho} = \begin{cases} \frac{\bar{\rho} - e^{-\hat{\mu}}}{1 - e^{-\hat{\mu}}}, & \bar{\rho} \geq e^{-\hat{\mu}} \\ 0, & \bar{\rho} < e^{-\hat{\mu}} \end{cases}$$

Proof *Theorem 1 gave us the formula to find $\hat{\rho}_\mu$, the value of ρ that maximizes the log-likelihood function for a given value of μ. Evaluating ℓ at this value of ρ yields what is called the profile log-likelihood for μ, $\ell_p(\mu)$. The MLE for μ will be the value of μ that maximizes $\ell_p(\mu)$. Once it has been found, the corresponding MLE for ρ can be calculated by plugging $\hat{\mu}$ into our formula for $\hat{\rho}_\mu$.*

Claim 2.1 $\bar{\rho} \geq e^{-\mu} \implies \ell_p(\mu) = n\Big[\bar{y}\log\mu - (1-\bar{\rho})(\log(1-e^{-\mu}) + \mu)\Big]$.

Proof of Claim 2.1 *From Theorem 1, $\bar{\rho} \geq e^{-\hat{\mu}} \implies \hat{\rho}_\mu = \frac{\bar{\rho} - e^{-\mu}}{1 - e^{-\mu}}$. It follows that*

$$1 - \hat{\rho}_\mu = \frac{1 - e^{-\mu}}{1 - e^{-\mu}} - \frac{\bar{\rho} - e^{-\mu}}{1 - e^{-\mu}} = \frac{1 - e^{-\mu} - \bar{\rho} + e^{-\mu}}{1 - e^{-\mu}} = \frac{1 - \bar{\rho}}{1 - e^{-\mu}}.$$

and

$$\hat{\rho}_\mu + (1 - \hat{\rho}_\mu)e^{-\mu} = \frac{\bar{\rho} - e^{-\mu}}{1 - e^{-\mu}} + \frac{1 - \bar{\rho}}{1 - e^{-\mu}}e^{-\mu} = \frac{\bar{\rho}(1 - e^{-\mu})}{1 - e^{-\mu}} = \bar{\rho}.$$

Thus,

$$\begin{aligned} \ell_p(\mu) &= \ell(\mu, \hat{\rho}_\mu) \\ &= n\bar{\rho}\log\left(\hat{\rho}_\mu + (1 - \hat{\rho}_\mu)e^{-\mu}\right) + n(1-\bar{\rho})(\log(1 - \hat{\rho}_\mu) - \mu) + n\bar{y}\log\mu \\ &= n\bar{\rho}\log(\bar{\rho}) + n(1-\bar{\rho})\left(\log\left(\frac{1-\bar{\rho}}{1-e^{-\mu}}\right) - \mu\right) + n\bar{y}\log\mu \\ &= n\Big[\bar{y}\log\mu - (1-\bar{\rho})(\log(1-e^{-\mu}) + \mu)\Big] + n\bar{\rho}\log(\bar{\rho}) + n(1-\bar{\rho})\log(1-\bar{\rho}). \end{aligned}$$

Discarding the final two terms as they don't depend on μ yields the expression for $\ell_p(\mu)$ given in the claim.

Claim 2.2

Claim 2.3 $\bar{\rho} < e^{-\mu} \implies \ell_p(\mu) = n(\bar{y}\log\mu - \mu)$.

Proof of Claim 2.3 *From Theorem 1,* $\bar{\rho} < e^{-\mu} \implies \hat{\rho}_\mu = 0$. *Thus,*

$$\begin{aligned}
\ell_p(\mu) &= \ell(\mu, \hat{\rho}_\mu) \\
&= n\bar{\rho}\log\left(\hat{\rho}_\mu + (1-\hat{\rho}_\mu)e^{-\mu}\right) + n(1-\bar{\rho})(\log(1-\hat{\rho}_\mu) - \mu) + n\bar{y}\log\mu \\
&= n\bar{\rho}\log(e^{-\mu}) + n(1-\bar{\rho})(\log(1) - \mu) + n\bar{y}\log\mu \\
&= -n\bar{\rho}\mu - n(1-\bar{\rho})\mu + n\bar{y}\log\mu \\
&= n(\bar{y}\log\mu - \mu).
\end{aligned}$$