Analysis of Optimal inferential models for a Poisson Mean [1]

Ashwin Bhargava, Meredith Saunders, Anna Tomkins
NC State University ST 502
(Dated: December 16, 2019)

In this paper, We aim to analyze the paper: "Optimal inferential models for a Poisson Mean" by Ryan Martin, et al. This paper aims to utilize Inferential models to analyze the issue of the Poisson, which is challenging for its discreteness. [1] We will begin by first demonstrating the challenges and interesting characteristics of the Poisson distribution. We will then go into defining properties of inference to then set up the foundation for inferential models. With this, we will then apply the general inferential model structure to the Poisson and replicate the results produced in Dr. Martins paper.

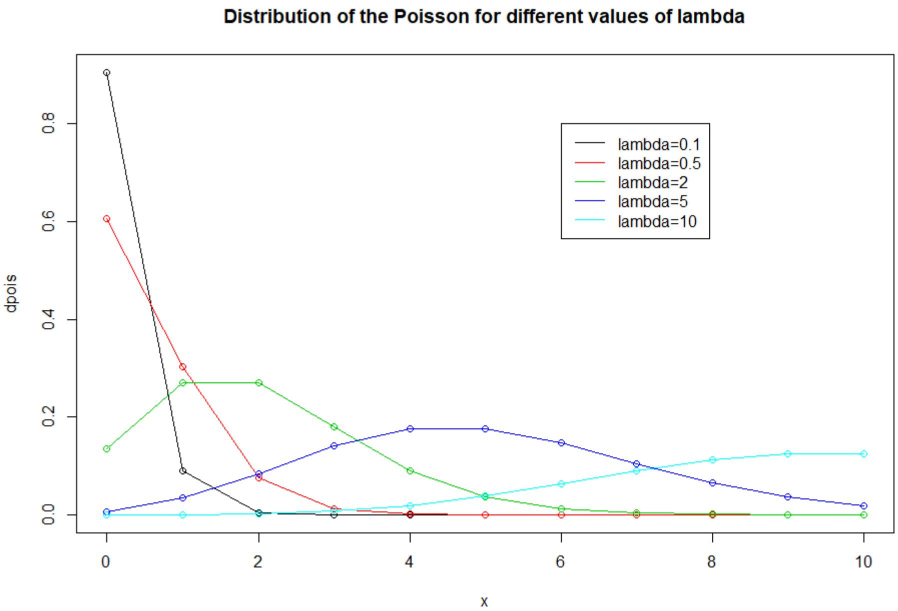
I. INTRODUCTION AND BACKGROUND

A. The Poisson Distribution

The Poisson Distribution presents some interesting characteristics. For one, it is a discrete distribution, meaning that it can only take integer values, further, it can only take positive integer values. Other interesting characteristics are that all occurrences are independent, and the mean is constant. The Poisson frequency distribution is [2]:

$$P(X=k) = \frac{\lambda^k}{k!}e^{-\lambda}, \qquad k = 0, 1, 2, \dots \tag{1}$$

Utilizing R,we can observe how the shape of the function varies with λ. (R code included in Appendix)



We also know that the Poisson distribution is related to the Binomial distribution. It is related in that the Poisson distribution approximates a Binomial distribution when the number of trials is large (n $\rightarrow \infty$) and the probability of success is small (p $\rightarrow$ 0), such that $np = \lambda$. We proved this during the semester of ST 502. [2]

The Poisson Distribution has a wide range of applications. For example, IBM conducted a study where it utilized the Poisson model combined with Bayesian statistics to understand employees previous experience and compared to success within the company.[3] In this study IBM treated their employees as counts, and drew inference about the mean of the distribution based on this data. In physics, this distribution is used to model several topics, such as scattering of alpha particles. This experiment observes a small number of particles, and is best modeled by the Poisson distribution for its discreteness and counting characteristics. Another example is understanding traffic trends during light and heavy traffic situations. [2] The Poisson distribution provides important analysis of count data and time until a measured "event" under fixed time or space conditions.

B. Inference

After understanding the basic principles and properties of the Poisson distribution, it is now important to understand how to draw inference from it. As stated in the textbook Inferential Models the definition of statistical inference is

> "Statistical inference provides meaningful probabilistic summaries of evidence available in the observed data concerning the truthfulness and falsity of any assertion of hypothesis about the unknown quantities of interest." [4]

Essentially, we want to gain insight to unknown parameters based on observed data. In order to further our understanding of inference, we have some basic functions we want to define, the *belief* and *plausibility* functions.

Before we define these functions, we first want to set up the problem and define our data:

$$X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, \sigma^2) \tag{2}$$

θ is unknown, and is the mean of our data. Thus θ is the quantity that we want to make inference on. With this, we want to make assertions of interest about θ. This assertion, denoted A, is $A = [\theta_0, \theta_1]$, where $\theta_0 \leq \theta_1$ [4].

The *belief* is written $bel_x(A)$. 'Belief' and 'probability' can be thought of similarly, although beliefs will have different mathematical properties than an ordinary probability measure. Belief is the "truthfulness of the assertion A given data set X." The mathematical difference from probability is that it does not properly satisfy the complement rule, which defines a probability measure.[4] This simply means the belief function is sub-additive. The definition of belief now sets us up to understand *plausibility*, which is:

$$pl_x(a) = 1 - bel_x(A^c) \tag{3}$$

Equation 3 denotes the plausibility of the assertion A - a measure of the evidence in data X that does not support the falsity of A. In order to better understand plausibility, we can think of it similar to that of a p-value. Both belief and plausibility are important to summarize the evidence in x supporting A. With these definitions, we then want to formalize this to an Inferential Model.

C. Inferential Models

1. *Basic Construction*

An Inferential Model (IM) follows a similar beginning framework to that of a fiducial framework, which we will lay out below: we have an auxilary variable, U, which takes values in a space $\mathbb{U}$ with probability measure $U \sim P_U$, which is associated with X and θ. This then characterizes the sampling distribution $X \sim P_{X|\theta}$. After this, the IM approach takes a different path to the fiducial approach. The IM approach instead of interpreting U as a random variable, we now treat u^*, the unobserved value, which is related to the data and the true value of θ as the fundamental quantity. [1] This is different because the fiducial method would continue to intepret U as a random variable. The IM approach is able to correct for possible bias that is introduced in a fiducial framework because it is not entirely a prior-free model [4]. We now follow three steps that define an Inferential Model, which is defined in [1]:

1. A-step. Associate X, θ, and $U \sim P_U$ in a way consistent with the sampling distribution $X \sim P_{X|\theta}$ such that for all $x \in \mathbb{X}$ and all $u \in \mathbb{U}$, it defines a unique subset $\Theta_x(u) \subseteq \Theta$, possibly empty, containing all possible candidate values of θ given (x, u)

2. P-step. Predict the unobserved value u^* of U associated with the observed data by an admissible predictive random set $\mathcal{S}$.

3. C-step. Combine $\mathcal{S}$ and the association $\Theta_x(u)$ specified in the A-step to obtain [1]

$$\Theta_x(\mathcal{S}) = \bigcup_{u \in \mathcal{S}} \Theta_x(u) \tag{4}$$

The A-step uses the predictable auxiliary variables, which allow us to introduce posterior probability-like quantities without a prior distribution for θ. [4] With this algorithm, we can then compute the belief function:

$$bel_x(A; \mathcal{S}) = P_{\mathcal{S}}\{\Theta_x(\mathcal{S}) \subseteq A\} \tag{5}$$

The plausibility function then follows as: [1]

$$pl_x(A;\mathcal{S}) = bel_x = P_{\mathcal{S}}\{\Theta_x(\mathcal{S}) \cap A \neq \emptyset\} \tag{6}$$

The P-step relies upon an admissible predictive random set $\mathcal{S}$ that is often restricted to sets which have a nested support. This allows for a upper bound to be placed on the belief function to uphold the validity requirement. [1]

2. Validity and Optimality

After defining an Inferential Model, we now want to understand the properties that ensure these models are valid and optimal. We want to define *valid*: An IM is valid of the $bel_X(A,\mathcal{S})$ is no larger than Unif(0,1) when $X \sim P_{x|\theta}$ with $\theta \notin A$ [1]. The uniform distribution is used due to its familiarity and simplicity as a fixed objective scale.

$$sup_{\theta \in A^c} P_{X|\theta}\{bel_X(A;\mathcal{S}) \geq 1-\alpha\} \leq \alpha, \alpha \in (0,1) \tag{7}$$

We can achieve an optimal IM by maximizing the belief function while maintaining the validity of the IM using the ability to generate an upper bound.

$$\mathbb{U}_x(A) = \{u \in \mathbb{U} : \theta_x(u) \subseteq A\} \tag{8}$$

Proposition 1. *Given an assertion A, suppose that $\{\mathbb{U}_x(A) : x \in \mathbb{X}\}$ defined in (8) forms a nested collection of sets. Then there exists an admissible predictive random set $\mathcal{S}^*$ such that $bel_X(A;\mathcal{S}^* = bel_x(A;\mathcal{S}_0)$ for all x.*

If the validity of the IM holds while maintaining an upper bound on any predictive random set, $\mathcal{S}$, and the set $\mathbb{U}_x(A)$ is nested, we can conclude the corresponding IM is optimal [1]. We can then use the belief and plausibility functions to generate plausibility regions for the unknown parameter. Often sets are discovered that are not nested, and we have to take additional steps to correct for the issue to allow for the optimal IM to be found.

II. APPLICATION TO THE POISSON

A. Association with Gamma and step construction

Given a Poi(θ) process, integration-by-parts shows that F_θ satisfies $F_\theta(x) = 1 - G_{x+1}(\theta)$, where G_x is a Gamma(x,1) distribution function. For this discrete problem, the association for X, given θ, may be written as:

$$G_{x+1}(\theta) < u < G_x(\theta) \tag{9}$$

Equation 9 links the data, parameter, and chosen auxiliary variable. This u-interval can be inverted to find the θ interval.

$$\Theta_x(u) = (G_x^{-1}(u), G_{x+1}^{-1}(u)] \tag{10}$$

Recall that inference on u^* is performed since we know more about u^* than θ. This is because we know x $= \theta + u^*$, and X=x is observed. Therefore, if we can do inference on u^*, we also know θ in the IM framework.

In order to be a credible predictive random set, S is defined:

$$S(u) = [1/2 - |u - 1/2|], 1/2 + |u + 1/2|] \tag{11}$$

This satisfies the criteria to be an efficient predictive random set for predicting U from a uniform distribution [0,1]. This is not the only set we could choose, there may be other choices for S(u), which may perform differently based on the assertion of interest. The random set is thus given as:

$$\Theta_x(S) = \bigcup_{u \in s} (G_x^{-1}(u), G_{x+1}^{-1}(u)] = (G_x^{-1}(0.5 - |u-0.5|), G_{x+1}^{-1}(0.5 + |u-0.5|)) = (\underline{\Theta}_x(U), \overline{\Theta}_x(U)) \tag{12}$$

Where u is a random draw from a Uniform distribution from 0 to 1. For an assertion A $= \{\theta\}$, the plausibility function is:

$$pl_x(\theta) = 1 - P_U\{\underline{\Theta}_x(U) > \theta\} - P_U\{\overline{\Theta}_x(U) < \theta\} = 1 - max\{1 - 2G_x(\theta), 0\} - max\{2G_{x+1}(\theta) - 1, 0\} \tag{13}$$

Graphing this plausibility function for some X=x as a function of θ should indicate a maximum of θ values around X=x which are plausible. For example, for the assertion $\{\theta{=}3\}$, when X $=$ 3 is observed, $pl_3(3) = 1$. For comparison the Poisson PMF with these two values returns a value of 0.224 [4].

B. Nesting Predictive Random Sets via Intersections

A predictive random set must be nested so that we can find the optimal belief function while maintaining validity. In order to modify the sets, we reorder the sets via intersections that are constructed based on the generated ranking, ρ.

1. Choose a ranking ρ on $\mathbb{X}$
2. Let $\mathbb{T} = \{1, 2, ...\}$ and define $\mathcal{S}_\rho = \{\mathcal{S}_t^\rho : t \in \mathbb{T}$ as follows. Set $\mathcal{S}_0 = \emptyset$ and $t = 1, 2, ...$ [1]

$$\mathcal{S}_t^\rho = \bigcap_{x:\rho(x)>t} \mathbb{U}_x(\{_0\}^c) = \bigcup_{x:\rho(x)>t} (G_{x+1}(\theta_0), G_x(\theta_0)] \tag{14}$$

We achieve an admissible predictive random set when the sets are nested which allows us to find the largest subset that contains the natural measure, $P_\mathcal{S}\{\mathcal{S} \subseteq S\} = P_U(S)$.[4] The largest subset that does not break the validity theorem is used to optimize the corresponding belief and plausibility functions. The belief function is generated via:

$$bel_x(\{\theta_0\}^c; \mathcal{S}_\rho) = P_U\{S^\rho_{\rho(x)-1}\} = \sum_{x':\rho(x')<\rho(x)} [G_{x'}(\theta_0) - G_{x'+1}(\theta_0)] = \sum_{x':\rho(x')<\rho(x)} f_{\theta_0}(x') \tag{15}$$

The plausibility then follows as: [1]

$$pl_x(\theta_0; \mathcal{S}_\rho) \equiv pl_x(\{\theta_0\}; \mathcal{S}_\rho) = 1 - \sum_{x':\rho(x')<\rho(x)} f_{\theta_0}(x') \tag{16}$$

The ranking, ρ, selected for the nesting of the intersections will be optimal when the belief function is maximized under $X \sim Pois(\theta)$ for $\theta \neq \theta_0$.

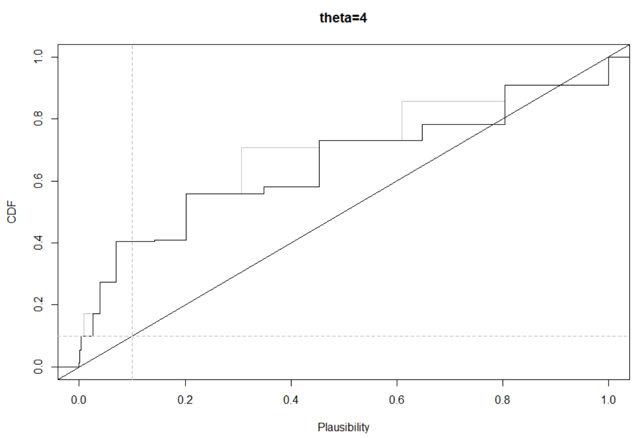
Additional ordering operations are demonstrated in Figure I based on the algorithm provided.

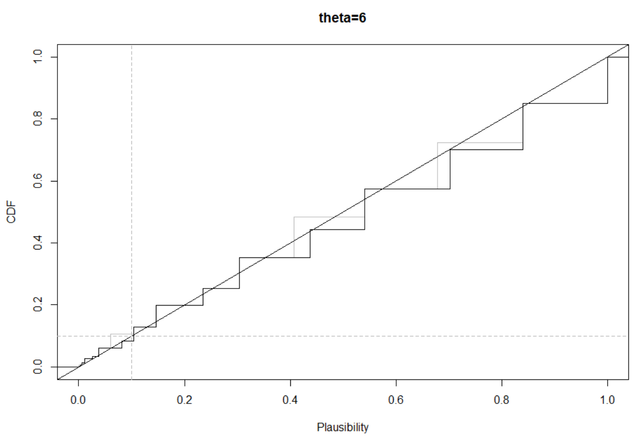
III. R CODE AND PLOTS

A. Figure I

In figure I, we replicated Figure 2 from [1]. This is a plot of the CDF of $pl_x(\theta)$. We used 100,000 Monte Carlo samples from $X \sim Pois(\theta_0)$, where $\theta_0 = 7$. We plotted various values of θ and can observe the deviation from the straight line as we change the values of θ. The further the chosen assertion θ is from θ_0, the more deviation there is from the uniform distribution CDF, which is represented by the straight, diagonal line in each plot of Figure 1. The plausibility distribution functions are above the uniform CDF, which is what we would like to observe since we are assigning smaller plausibility to assertions that are not true [4].

The validity requirement is of interest in the middle-left graph, where the assertion $\theta = \theta_0$. Ideally we would observe the Optimal IM to be the only method that satisfies the validity principle - the two frequentist methods, denoted with grey lines, should be exceed the Uniform CDF. This is because these methods do not produce probabilistic interference in the sense that the inferential models do (a confidence interval and power calculation can be done without seeing X, they do not depend on data). Our reproduced plot $\theta = \theta_0 = 7$ was unable to show this property.





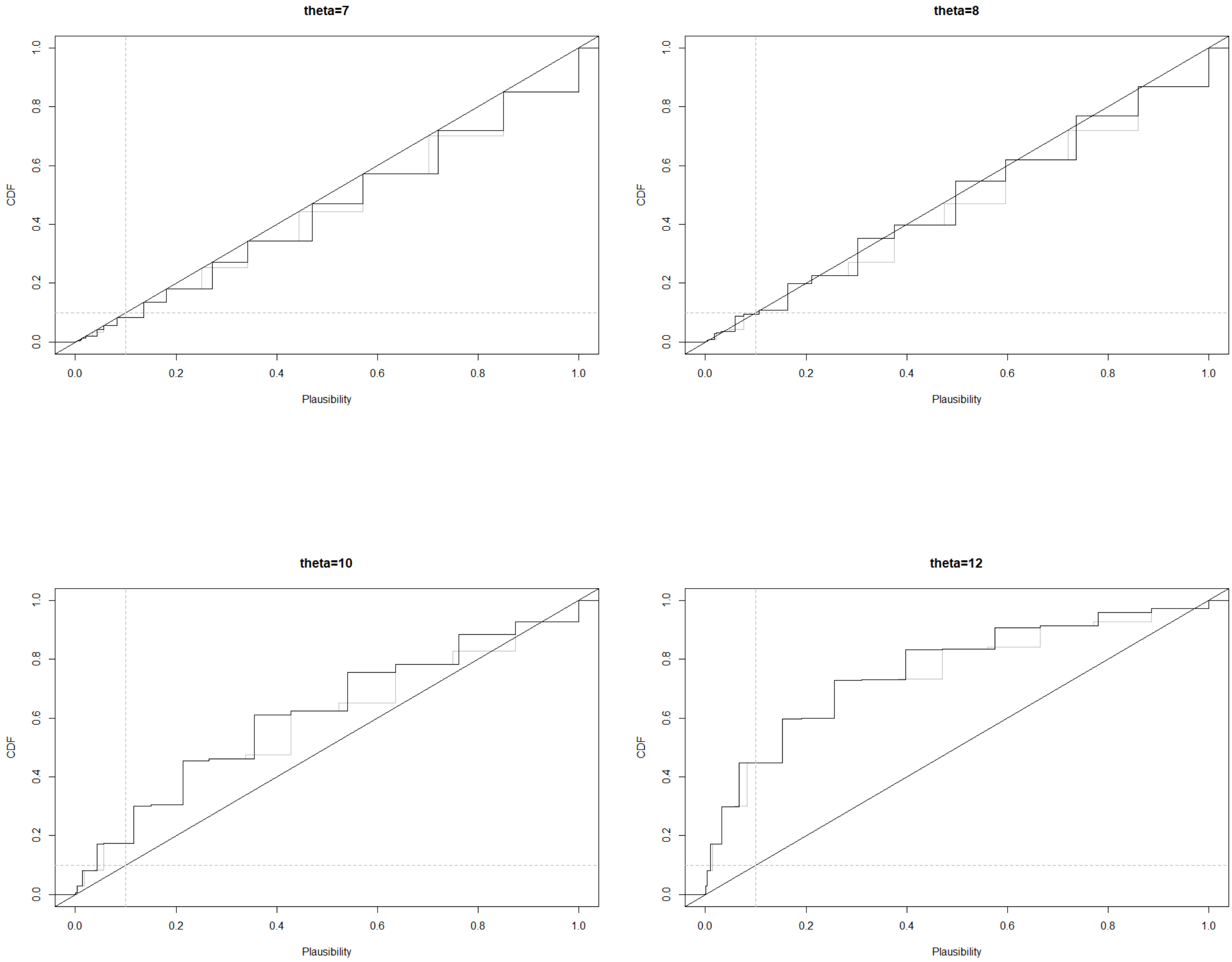


FIG. 1: Plots of CDF of plausibility using Monte Carlo samples for various θ's. $\theta_0 = 7$

B. Figure II

In figure II, we replicated Figure 3 from [1]. These plots are of the plausibility, $pl_x(\theta)$, for various x values. We can see that all of the plots are centered at our true x value, and both the IM and normal plausibility functions peak at $\theta = x$. This means that as we move further and further from our true value, $pl_x(\theta)$ decreases. The horizontal line is a "cut-off" to capture where the 90% plausibility intervals start and end. The normal plausibility intervals are slightly shorter, however, since there is no satisfaction of the validity theorem, the coverage is not guaranteed in this framework.

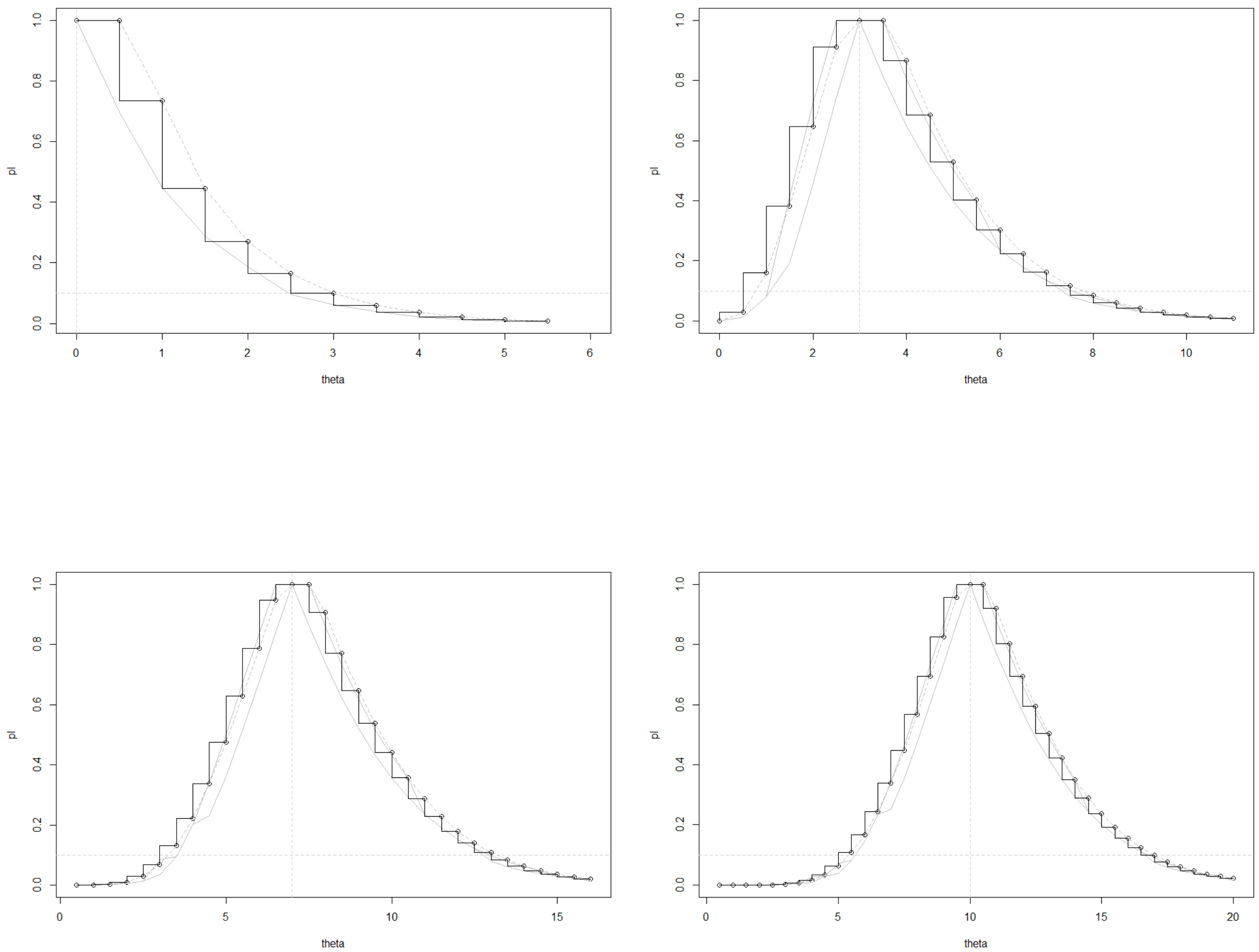


FIG. 2: Plots of plausibility as a function of θ, for x values of 0, 3, 7, 10

IV. CONCLUSION

As presented, we see that the Poisson Distribution has some interesting, but challenging characteristics. Due to its discreteness, independence, and constant mean, we face a challenge with wanting to perform any inference. The Poisson is only defined by one parameter, which is its rate. The optimal IM allows for us to draw inference on the unknown parameter by creating an admissible predictive random set from any observed data. Our generated set should contain all possible values of the unknown rate value. We use plausibility, combined with a belief function, to determine the optimal IM function that provides insight for the true rate value.

[1] R. M. et. al, 06/30/2012 p. 20 (2012).
[2] J. Rice, *Mathematical Statistics and Data Analysis* (Thomson Higher Education, 2007), 3rd ed.
[3] IBM, *Using bayesian one sample inference: Poisson models to draw inference about employees' previous experience* (Unknown), URL `https://www.ibm.com/support/knowledgecenter/en/SSLVMB_subs/statistics_casestudies_project_ddita/spss/tutorials/bayesian_poisson_intro.html`.
[4] R. Martin, *Inferential Models* (CRC Press, 2016), 1st ed.