AST 5731 Project 1 Synopsis

Authors: Sam Penders, Derek Perera, Kiet Pham

October 5, 2022

1 Introduction

For this project, we analyze the number of globular clusters in a galaxy (N_{GC}) in a specific range of absolute visual magnitude (M_V). Globular clusters are spheroidal, gravitationally-bound conglomerations of thousands to millions of stars in a galaxy. The distribution of globular clusters when limited by M_V is an active research question that can be studied with recent observations. As this is a counting problem where the counts can take on any integer value from zero to some very large number, limited by the size of the galaxy, we expect this distribution can be approximated as Poisson. Additionally, globular clusters in different locations in a galaxy will interact minimally, so we can consider the presence of globular clusters in spatially-separated regions to be independent events.

To test this hypothesis, we employ a Bayesian inference analysis with the goal of finding the mean number of globular clusters in a galaxy within a specific M_V range. We restrict our analysis to the range $-15 < M_V < -14$, where the range of counts of globular clusters is suitable to be described by a Poisson distribution. For our data set, we use a relatively recent catalog of globular cluster systems [1] with various physical characteristics. In Figure 1, we present the relevant observations of the catalog. Our analysis was conducted in Jupyter notebook using Python 3 with common packages like Pandas, Numpy, SciPy, and Matplotlib. This notebook was submitted alongside this report, and versions are listed inside. In Figure 2, we present the data points in our chosen magnitude range of $-15 < M_V < -14$.

2 Statistical Model

We assume that the number of globular clusters X_i in galaxy i of visual magnitude $-15 < M_V < -14$ is a Poisson random variable with an unknown mean $\theta \in \Theta = [0, \infty)$. For a sample X of 13 observations, $X = (X_1, X_2, \ldots, X_{13})$, we assume $X_i \,|\, \theta \overset{\text{ind}}{\sim} \text{Poisson}(\theta)$, as motivated in the Introduction. Our sample space $\mathcal{X}$ for X is $\mathcal{X} = \mathbb{N}_0^{13}$.

Our data set comes from the catalog [1] described in Section 1, of which we care about the observation $\mathbf{x} = (x_1, x_2, \ldots, x_{13})$ of the number of globular clusters in 13 galaxies within our range of M_V. Our statistical model employs Bayesian inference allowing us to obtain the posterior distribution $p(\theta \,|\, \mathbf{x})$ by multiplying the likelihood $p(\theta \,|\, \mathbf{x})$ by the prior $\nu(\theta)$:

$$p(\theta \,|\, \mathbf{x}) \propto p(\mathbf{x} \,|\, \theta)\nu(\theta).$$

Based on our assumptions about X, the likelihood is given by

$$p(\mathbf{x} \,|\, \theta) = \frac{\theta^{\sum_{i=1}^{13} x_i} e^{-13\theta}}{x_1! x_2! \ldots x_{13}!}.$$

Lastly, we choose a prior for θ. Given the domain of θ, a natural assumption is that $\theta \,|\, a, b \sim \text{Gamma}(a, b)$ for some shape and rate parameters $a, b > 0$ that we will pick. This choice of distribution gives us a conjugate prior where the posterior density for θ may be found analytically. Our prior density θ is

$$\nu(\theta) = \Gamma(a, b).$$

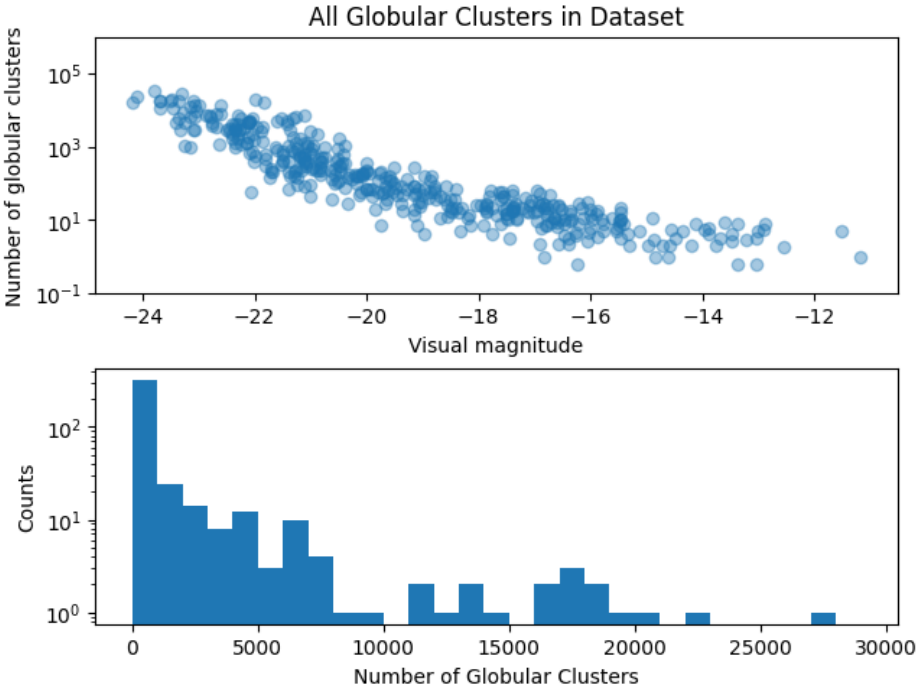

Figure 1: *Top:* Full dataset of galaxies plotted with their absolute visual magnitude M_V on the x-axis and the number of globular cluster N_{GC} within them on the y-axis. The overall trend shows that dimmer galaxies have fewer globular clusters. *Bottom:* Histogram of N_{GC} within the full sample. Most galaxies have a relatively small number of globular clusters, which exponentially falls off to larger N_{GC}. We restrict our analysis to $-15 < M_V < -14$, where the number of globular clusters is small.

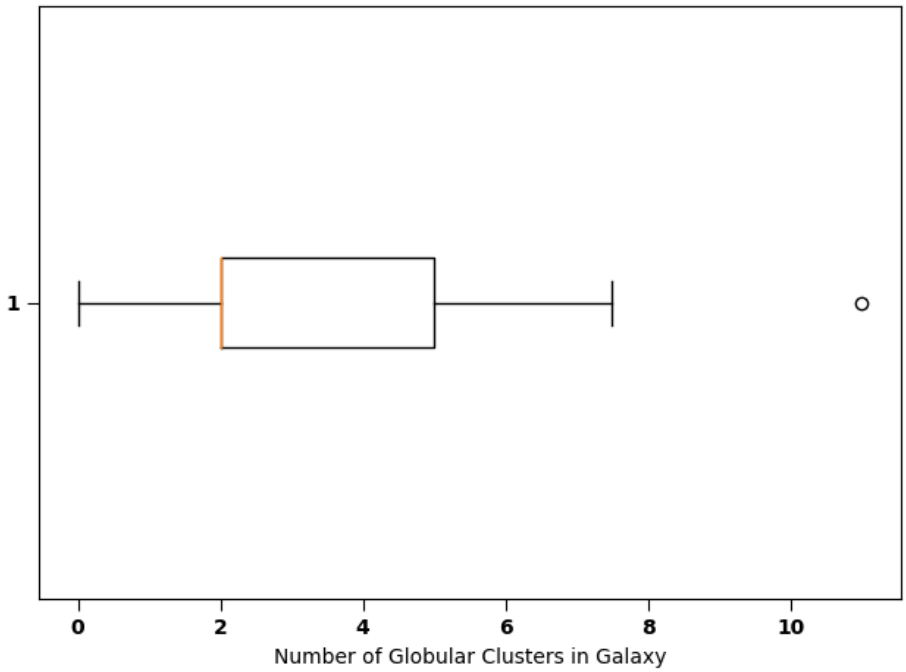

Figure 2: Box plot of the Number of Globular Clusters in a galaxy within the M_V range: $-15 < M_V < -14$. The vertical orange line indicates the median of the data. The sample mean and variance are 3.47 and 9.43 respectively. Although we assume this data to be a Poisson, the observation that the variance is larger than the mean, indicating overdispersion, suggests a negative binomial distribution is more appropriate.

This model is summarized below:

1. θ: mean number of globular clusters observed per galaxy with visual magnitude between -15 and -14.
2. $\Theta = [0, \infty)$.
3. $\mathbf{x} = (x_1, x_2, \ldots, x_{13})$: observations of the number of globular clusters in the sample of 13 galaxies, where each $x \in \mathbb{N}_0$
4. $\mathcal{X}$: All possible vectors of 13 non-negative integers.
5. $X_i \,|\, \theta \overset{\text{ind}}{\sim} \text{Poisson}(\theta)$.
6. $\nu(\theta) = \text{Gamma}(a, b)$.

3 Bayesian Analysis

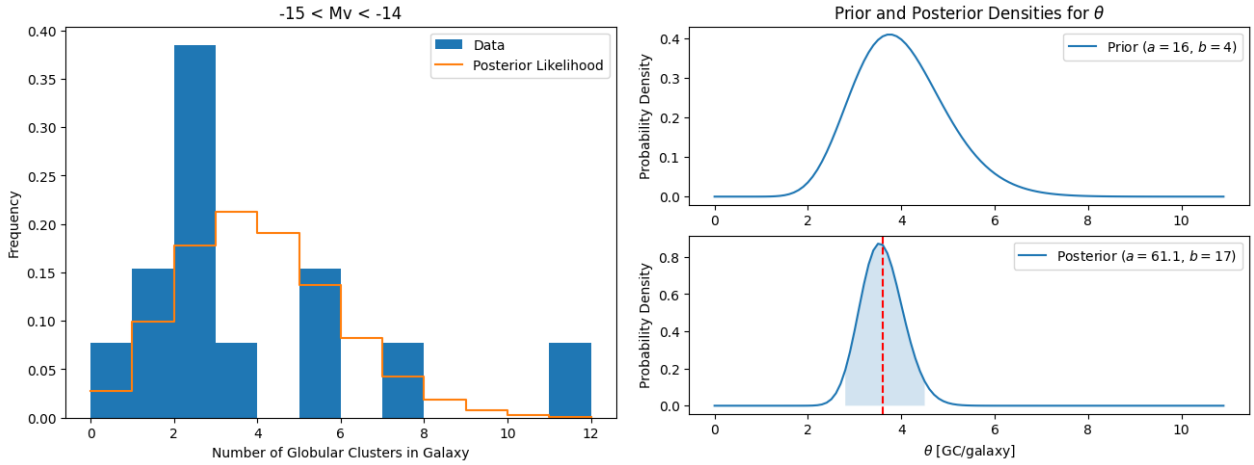

Figure 3: *Left:* Histogram of N_{GC} for galaxies with $-15 < M_V < -14$, labelled as data. In orange is a Poisson distribution with mean equal to the posterior mean of 3.59. From our analysis, the data roughly follows a Poisson distribution. *Right:* The upper panel shows the prior distribution we utilized, with $a = 16$ and $b = 4$. The lower panels shows the resulting posterior probability distribution. The vertical red dashed line indicates the posterior mean at 3.59. The light blue shaded region indicates the 95% confidence interval.

Following the statistical model, the use of $\Gamma(a, b)$ as our prior yields a Gamma distribution for our posterior:

$$p(\theta \,|\, \mathbf{x}) = \Gamma\left(a + \sum_{i=1}^{n} x_i, b + n\right) \tag{1}$$

Given our magnitude range is $-15 < M_V < -14$, x_i and $n = 13$ are the number of globular clusters in a galaxy with a magnitude within that range and the total sample of galaxies within that range, respectively. Since the posterior is a Gamma distribution, the posterior mean is:

$$\theta = \frac{a + \sum_{i=1}^{n} x_i}{b + n}. \tag{2}$$

Therefore, to calculate the posterior mean globular clusters per galaxy from this data set, we must define a and b for the prior. Based on data from [2], galaxies with $M_V = -15$ have approximately four globular clusters, so we assume $a = 16$ and $b = 4$ for the shape and rate parameters of the prior. This choice gives the prior mean $a/b = 4$ and the prior variance $a/b^2 = 1$.

Since there are fewer galaxies in this M_V range than the entire sample, this choice yields a prior distribution that roughly spans the relevant parameter space (see right panel of Figure 3) while skewing slightly to smaller N_{GC} as we should expect with a Poisson distribution.

With this choice, we calculate the posterior mean N_{GC} per galaxy using Equation (2) to be 3.59 with a 95% confidence interval of (2.75, 4.55), obtained using the posterior density from Equation (1). The posterior variance of θ is found to be 0.21. The posterior probability density distribution is shown in the right panel of Figure 3.

4 Prior Predictive Check

To assess whether out choice of prior generates reasonable data points roughly consistent with the data, we perform a prior predictive check. For this, we draw 1000 random variables distributed as a gamma distribution with $a = 16$ and $b = 4$. From these we generate a simulated data point from a Poisson distribution with a mean at the random variable from before. This gives us 1000 simulated data points with which we can compare with the observed data as shown in Figure 4. This analysis finds that the simulated data is roughly consistent with the observed data, with the discrepancies largely being a part of the smaller sample size seen with the data. Therefore, we conclude that our choice of prior is reasonable.

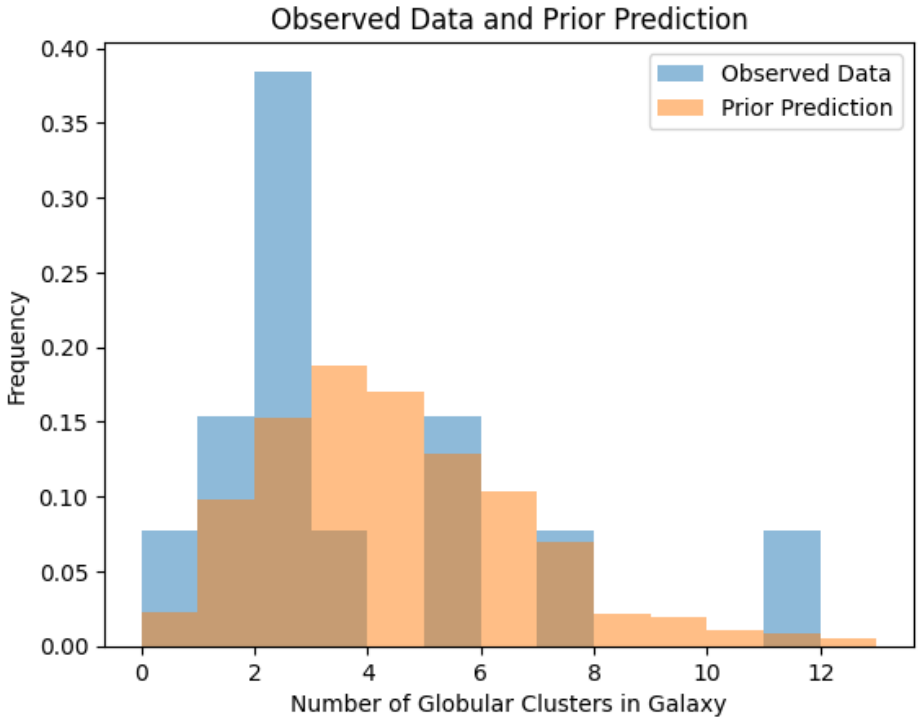

Figure 4: Histogram of the observed data in light blue and the simulated data points from the prior predictive check in light orange. The prior predictive check generates data that is roughly consistent with those observed, hence allowing us to conclude that our prior is reasonable.

5 Sensitivity Analysis

Lastly, to test how much our posterior changes with the prior assumptions, we perform a sensitivity check. Since we are utilizing conjugate priors, the posterior density will always be a Gamma distribution. Therefore, our sensitivity check need only look at how the prior uses of a and b change the posterior. We check a variety of a and b values that are different from our prior assumption of $a = 16$ and $b = 4$. These are shown in Figure 5. As we can see, the posterior mean does change dramatically for different priors. The most drastic changes are for the use of $a = 100$, where the mean jumped up to 6.31 and 10.36 for $b = 10$ and $b = 1$, respectively. While these are significant changes, the priors for these are not representative of the data. For the rest of the permutations, the posterior mean does vary but remains at ~ 3 globular clusters per galaxy. Therefore, we conclude that our prior assumption is sensitive but to within a reasonable degree of consistency.

6 Conclusion

We used Bayesian methods to predict the mean number of globular clusters in galaxies with $-15 < M_V < -14$, assuming that the number of globular clusters is a Poisson random variable, and using a gamma distribution as the prior for the mean. The posterior mean was found to be 3.59 with a 95% confidence interval of (2.75, 4.55). Even though we assumed a Poisson distribution for N_{GC}, Figure 2 indicates that our magnitude range depicts overdispersion, suggesting that a negative binomial distribution would be more appropriate. Therefore, we conclude that while our analysis with a Poisson yielded reasonable results, a better analysis would be to use a negative binomial distribution.

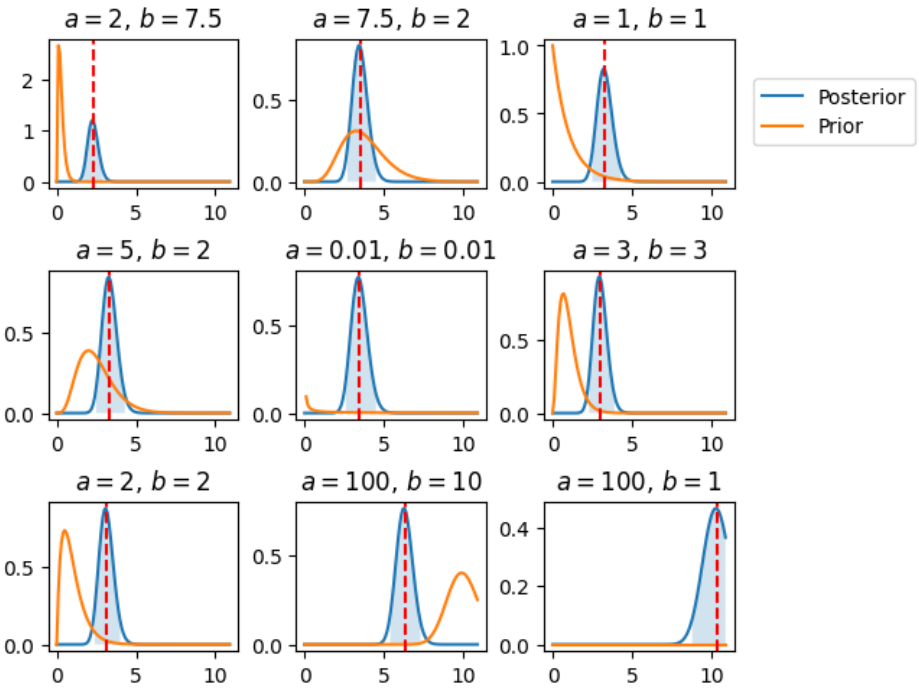

Figure 5: Sensitivity check results for a variety of priors (orange) with varying a and b. The resulting posteriors for each are shown in blue, with the red vertical line and shaded blue region indicating the resulting posterior's mean and 95% confidence interval, respectively.

References

[1] William E. Harris, Gretchen L. H. Harris, and Matthew Alessi. A CATALOG OF GLOBULAR CLUSTER SYSTEMS: WHAT DETERMINES THE SIZE OF a GALAXY's GLOBULAR CLUSTER POPULATION? *The Astrophysical Journal*, 772(2):82, jul 2013.

[2] W. E. Harris. Globular cluster systems in galaxies beyond the local group. *Ann. Rev. Astron. Astrophys.*, 29:543–579, 1991.