F79MA – Assessed Project 2 – 2025-26

Student: Michaelangelo dela Paz HW-ID: H00443289

Campus: Dubai

1. Introduction

This report will consider the modelling of the number of counts of a genetic mutation in a fixed length of an RNA sequence using an alternative distribution to Poisson and perform Bayesian analysis to determine whether the fitted model of said distribution can predict future observations.

2. Poisson vs. Negative binomial model of the count data

In theory, the simplest model to fit count data is the Poisson distribution with fixed rate λ; however, in many cases, λ is unknown. In Bayesian statistics, λ can be modelled as a Gamma distributed random variable (i.e. $\lambda \sim Gamma(\alpha, \beta)$). However, this would cause each observation to have a Poisson distribution with random λ. To find the actual distribution of the observed data, we find the marginal distribution

$$p_X(x) = \int_0^\infty \Pr(X = x \mid \lambda)\, \pi(\lambda)\; d\lambda.$$

Substituting $\Pr(X = x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$, and $\pi(\lambda) = \frac{\beta^\alpha}{\tau(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$ results

$$p_X(x) = \int_0^\infty \frac{\lambda^x e^{-\lambda}}{x!} \times \frac{\beta^\alpha}{\tau(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}\; d\lambda \;=\; \frac{\beta^\alpha}{x!\,\tau(\alpha)} \int_0^\infty \lambda^{x+\alpha-1} e^{-(1+\beta)\lambda}\, d\lambda.$$

Note that the integral $\int_0^\infty \lambda^{x+\alpha-1} e^{-(1+\beta)\lambda} d\lambda$ is equal to the Gamma integral identity $\frac{\tau(x+\alpha)}{(1+\beta)^{x+\alpha}}$ (Pishro-Nik, n.d.). Thus, this results to

$$p_X(x) = \frac{\beta^\alpha}{x!\,\tau(\alpha)} \frac{\tau(x+\alpha)}{(1+\beta)^{x+\alpha}}$$

which simplifies to

$$p_X(x) = \binom{x + \alpha - 1}{x} \left(\frac{1}{\beta+1}\right)^x \left(\frac{\beta}{\beta+1}\right)^\alpha = \binom{x-1}{\alpha-1} \beta^k (1-\beta)^{x-\alpha},$$

which is the probability mass function of a negative binomial distribution. As a result, this may be more effective in modelling count data, as the negative binomial distribution is able to capture overdispersion in the data, where the variance is greater than the mean. This is not possible with the Poisson distribution, as it is bound by the constraint that the variance must be equal to the mean. (Gelman et al., 2013)

3. Analysis of the count data

3.1 Overview

The following analysis will be based on mathematical derivation and utilizing R to carry out calculations and simulations. For replicability, the seed has been set using the line `set.seed(last_four_digits)`, with `last_four_digits` = 3289. This analysis will use the sample `My_Data`, which has been taken from `Full_Data` from the dataset `count_data.csv`. This analysis is based on the belief that `My_Data` can be modelled as independent and identically distributed realisations $\underline{x} = (x_1, \dots, x_{1000})$ from a negative binomial distribution given by *NegBin*(k, p), where *k* = *5* and *p* is unknown.

Before carrying out Bayesian analysis, given $\bar{x} = 23.041$, the maximum likelihood estimator (MLE) $\hat{p}$ given by $\hat{p} = {}^{k}\!/_{\bar{x}}$ is calculated and stored in variable `Quantity1` = 0.217. This estimate based on the observed data will serve as a non-Bayesian benchmark against the findings of the Bayesian analysis.

3.2 Bayesian analysis

3.2.1 Jeffrey's prior

Firstly, the prior distribution will need to be determined; however, not enough prior information has been given. In cases of vague priors, the Jeffrey's prior is derived, which is given by

$$\pi_J(p) \propto \sqrt{I(p)},$$

where $I(p)$ is the Fisher information. Thus, we derive $I(p)$ mathematically, starting with the likelihood function $L(p; \underline{x})$ given by

$$\begin{aligned} L(p; \underline{x}) &= \prod_{i=1}^{n} \binom{x_i - 1}{k - 1} p^k (1-p)^{x_i - k} \\ &\propto p^{nk} (1-p)^{\sum_{i=1}^{n}(x_i - k)} \end{aligned}$$

Then, the log-likelihood function $\ell(p; \underline{x})$ is given by

$$\ell(p; \underline{x}) = nk \log(p) + \left(\sum_{i=1}^{n} (x_i - k) \right) \log(1-p).$$

Then its first and second derivative are given by

$$\ell'(p;\underline{x}) = \frac{nk}{p} + \frac{\sum_{i=1}^{n}(x_i - k)}{1-p}, \qquad \ell''(p;\underline{x}) = -\frac{nk}{p^2} - \frac{\sum_{i=1}^{n}(x_i - k)}{(1-p)^2}.$$

Thus, the Fisher information $I(p)$ is given by:

$$\begin{aligned} I(p) &= -E\left[\ell''(p;\underline{x})\right] \\ &= -E\left[-\frac{nk}{p^2} - \frac{\sum_{i=1}^{n}(x_i - k)}{(1-p)^2}\right] \\ &= \frac{nk}{p^2} + \frac{nk}{p(1-p)} \\ &= \frac{nk}{p^2(1-p)}. \end{aligned}$$

Therefore, the Jeffrey's prior $\pi_J(p)$ is given by

$$\pi_J(p) \propto p^{-1}(1-p)^{-\frac{1}{2}}.$$

The prior is improper Beta distribution, meaning that it cannot be normalized as a proper density function on (0, 1). Thus, Bayes' theorem does not hold (*Week 8 Lecture Notes*). However, given we have enough information from the likelihood, this can still be used to derive the posterior density, as seen in the next section.

3.2.2 Posterior density

The posterior density is defined by

$$\begin{aligned} \pi(p|\underline{x}) &\propto \pi_J(p)L(p;\underline{x}) \\ &\propto p^{-1}(1-p)^{-\frac{1}{2}} \times p^{nk}(1-p)^{\sum_{i=1}^{n}(x_i-k)}, \end{aligned}$$

Given $n = 1000$ and $k = 5$, and using `My_Data`, it results to

$$\begin{aligned} &\pi(p|\underline{x}) \propto p^{5000-1}(1-p)^{\sum_{i=1}^{1000}(x_i-5)-\frac{1}{2}} \\ &\Rightarrow \pi(p|\underline{x}) \sim Beta(5000, 18041.5). \end{aligned}$$

Based on the posterior, the posterior mean is given by $\frac{\alpha}{\alpha+\beta}$, which is calculated and stored in variable `Quantity2` = 0.217, rounded to three decimal places. Note that the posterior mean and the MLE are approximately equal, which is the result when the prior is improper, as it gives minimal additional information outside the data. This implies that the posterior closely reflects the information from the sample `My_Data`. Figure 1 shows the posterior density plotted with Jeffrey's prior, showing the posterior distribution is a proper density despite the improper prior.

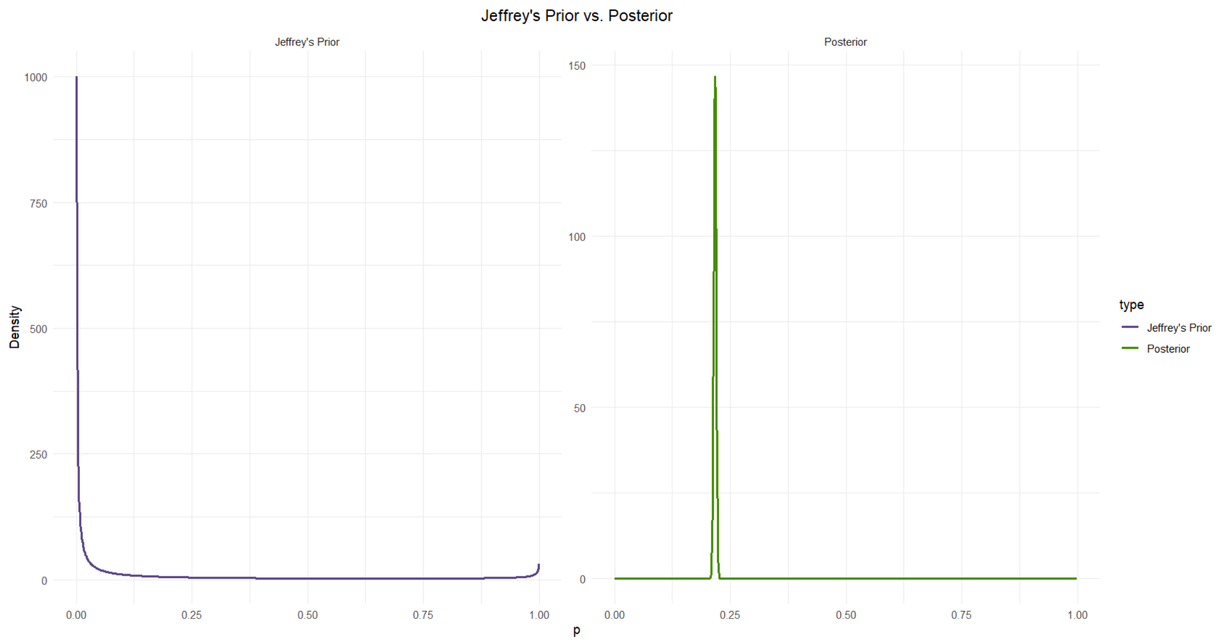

Figure 1: Jeffrey's Prior (left) vs. Posterior Density (right)

The prior diverges at $p = 0$ and flattens as p increases and therefore does not have a proper density. The shape of posterior distribution seems to be symmetric and very heavily concentrated around $p \approx 0.217$, which is consistent with the posterior mean. The very narrow spike indicates substantially low amount of variance. The posterior density shows that the best estimate for p is 0.217, which is the same as the MLE.

3.2.3 Posterior predictive distribution

Let $\alpha = \alpha_{post},\ \beta = \beta_{post}$. Based on this posterior, the posterior predictive distribution $\pi(z|\underline{x})$ for an unseen count z is given by

$$\begin{aligned}
\pi(z|\underline{x}) &= \int_{p=0}^{1} \pi(p|\underline{x}) P(Z = z|p)\, dp \\
&= \int_{p=0}^{1} \frac{\tau(\alpha+\beta)}{\tau(\alpha)\tau(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \times \binom{z-1}{k-1} p^{k}(1-p)^{z-k}\, dp \\
&= \frac{\tau(\alpha+\beta)}{\tau(\alpha)\tau(\beta)} \binom{z-1}{k-1} \int_{p=0}^{1} p^{\alpha+k-1}(1-p)^{\beta+z-k-1}\, dp.
\end{aligned}$$

Note the integral can be written as a beta function, namely

$$\int_{p=0}^{1} p^{\alpha+k-1}(1-p)^{\beta+z-k-1}\, dp = \frac{\tau(\alpha+k)\tau(\beta+z-k)}{\tau(\alpha+\beta+z)},$$

resulting in the probability mass function of the Beta-Negative Binomial distribution

$$\pi(z|\underline{x}) = \binom{z-1}{k-1}\frac{B(\alpha+k,\beta+z-k)}{B(\alpha,\beta)}.$$

However, with large α and β, the Beta functions get astronomically small, making it impossible to calculate. To fix this, we replace the beta functions with log-beta functions, which then are transformed back by exponentiating. In other words,

$$\pi(Z=5|\underline{x}) = \exp\left\{\log\binom{z-1}{k-1} + \log B(\alpha+k,\beta+z-k) - \log B(\alpha,\beta)\right\}.$$

Given $z=5, k=5$, $\alpha=5000$, $\beta=18041.5$, the value of $\pi(Z=5|\underline{x})$ is given by

$$\pi(Z=5|\underline{x}) = \exp\{\log B(5005, 18041.5) - \log B(5000,18041.5)\}.$$

Using this form, the value of $\pi(Z=5|\underline{x})$ is calculated and stored in variable `Quantity3` = 0.00048, showing that smaller counts are predicted to be highly improbable.

3.2.4 Simulations

Using the result of derivation of the posterior predictive distribution, we run a simulation of 10000 realisations following a Negative binomial distribution $NegBin(5,p)$, where $p \sim Beta(\alpha,\beta)$ is sampled from the posterior distribution. To gauge the performance of the predictive distribution, we also analyze the observed data with it. Figure 2 plots the histograms for the observed data and the simulated data.

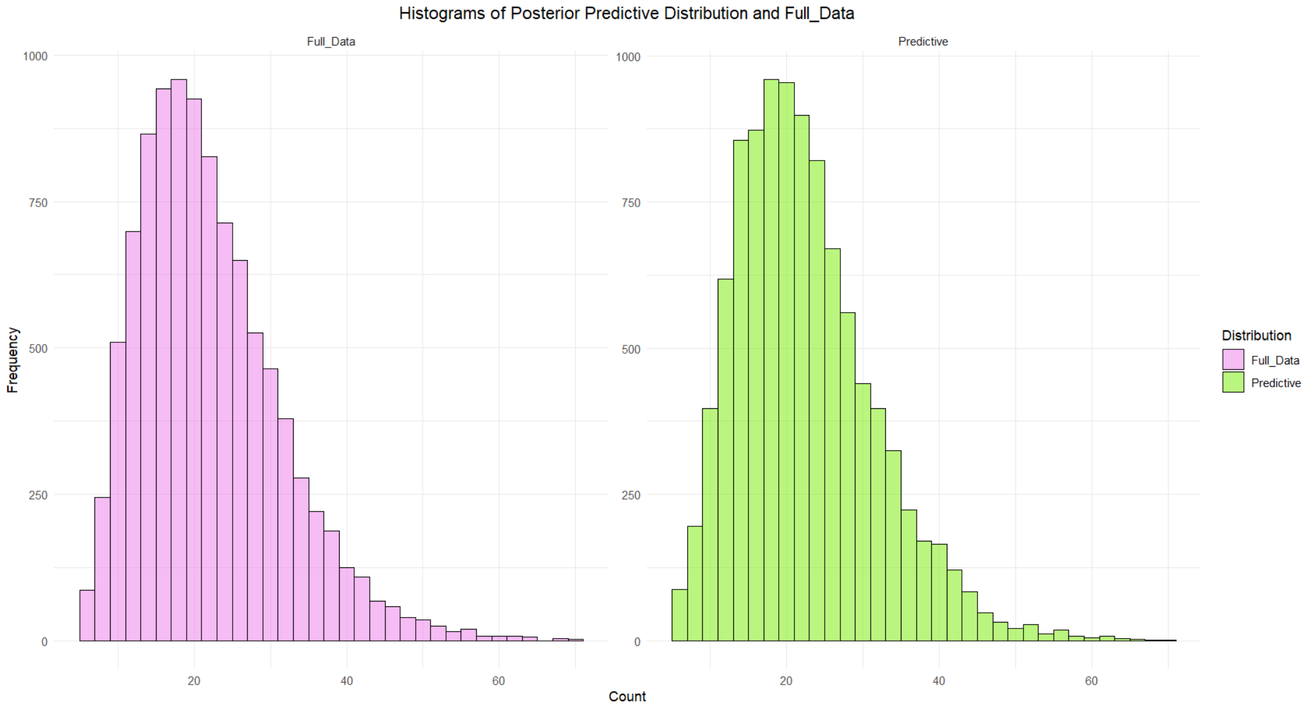

Figure 2: Histograms of the predictive distribution and `Full_Data`

The predictive distribution closely matches the overall shape and location of the observed data, showing that the negative binomial model provides a good fit. Note that both histograms show that, as stated earlier, the negative binomial distribution can capture the uncertainty of p and the overdispersion of the data, which is reflected by the tail of the distributions. This shows that future observations can be accurately predicted using this fitted model.

4. Conclusion

From this analysis, we have determined that, in biology and many other contexts, modelling count data via a negative binomial distribution stem from the Poisson-gamma model and is more effective than directly using the Poisson distribution due to its overdispersed nature. We have also seen that despite the Jeffrey's prior being improper and non-informative, the likelihood well enough provides the information needed to derive the posterior. The posterior mean was approximately equal to the MLE of p, leading to a rare case of agreement between the Bayesian and frequentist approaches to estimating p. By deriving the predictive distribution to be a Beta-Negative Binomial mixture, we have seen that it well matches the observed data, making it suitable for biologists to use this model in predicting the number of counts of a genetic mutation in an RNA sequence.

5. References

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. and Rubin, D.B. (2013). *Bayesian Data Analysis*. 3rd ed. Chapman and Hall/CRC. doi:https://doi.org/10.1201/b16018.

F79MA Statistical Models A, *Week 8: Bayesian statistical inference.*

Pishro-Nik, H. (n.d.). *Gamma Distribution | Gamma Function | Properties | PDF*. [online] www.probabilitycourse.com. Available at: https://www.probabilitycourse.com/chapter4/4_2_4_Gamma_distribution.php [Accessed 22 Nov. 2025].