Body Count in Quentin Tarantino's Next Movie
Math 264: Bayesian Analysis Project

Eduardo Gonzalez, Elizabeth Bunt, Krishna Singh

November 29, 2018

Contents

1 Executive Summary

Quentin Tarantino is an American director and screenwriter well known for his use of violence and strong language in his movies. Due to the violence in his previous movies, the audience can expect to see violence in any of his future movies. In this report, we are going to estimate the amount of deaths in Quentin Tarantino's next movie. Our results estimate between 13 and 33 deaths in Tarantino's next two-hour movie, with the caveat that we believe the variance of this interval is underestimated.

2 Introduction

Based on the number of deaths in Quentin Tarantino's previous movies we estimate the number of deaths in his next movie using Bayesian analysis. A Poisson distribution with a conjugate prior will be used to model the data. The Poisson distribution is used to model the probability of a given number of occurrences within a fixed time interval. The outline of the process of the analysis is as follows: 1) describe and provide basic measure statistics of the data, 2) detail the model selection and interpretations of the model, and finally 3) discuss conclusions of the results.

3 Data Characteristics

In the analysis we use data from Tarantino's movies including name of the movie title, the year the movie was created, the length of the movie and the number of deaths in each of the movies, see table below. The data gathered for the project was provided by Dr. Bee Leng Lee of San Jose State University. For this analysis, the length of the movie and death count will be considered. Five Tarantino films have 20 or less dead per movie and one film has 68 dead, see Table 1. With a Poisson distribution we are not surprised to see a skew to the right, see Figure 1. We will consider Kill Bill Vol.1 with 63 deaths to be a possible outlier.

Year	Film	Length	Body Count
1992	Reservoir Dogs	1h 39min	10
1994	Pulp Fiction	2h 34min	7
1997	Jackie Brown	2h 34min	4
2003	Kill Bill: Volume 1	1h 51min	63
2004	Kill Bill: Volume 2	2h 17min	11
2009	Inglorious Bastards	2h 33min	48
2012	Django Unchained	2h 45min	47
2015	The Hateful Eight	3h 7min	19

Table 1: Quentin Tarantino's films have a body count ranging from 4 to 63 people

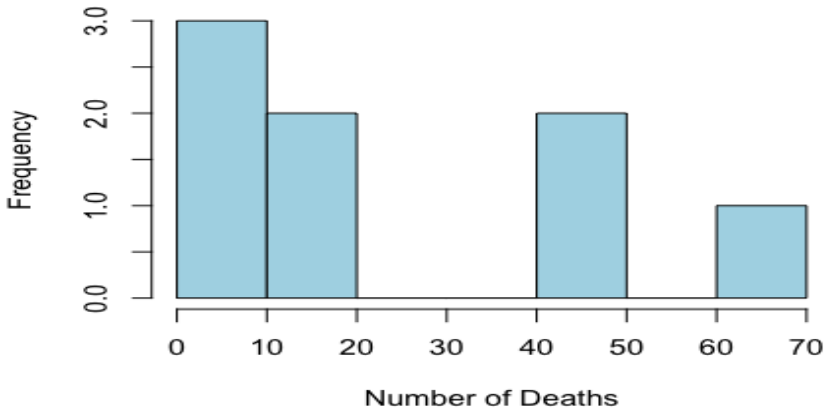

Figure 1: Five Tarantino films have a body count of 20 or less and one film, Kill Bill Vol.1, has a body count of 63.

4 Model Selections and Interpretations

4.1 Sampling Distribution

In this section we present our model for estimating deaths in an upcoming Tarantino movie, followed by an interpretation, parameter estimates, and justifications. Then we discuss an alternate model.

Since our analysis is dealing with the number of occurrences within a given time interval, a Poisson sampling distribution is a reasonable choice to model the data. Thus,

$$y_i \sim Poisson(x_i\theta)$$

where y_i represents the death count for movie i, x_i represents the length in hours of movie i and $\theta > 0$ represents death per hour, which is unknown. This model can be used under certain assumptions (Appendix A). "This model is not exchangeable in the y_i's but is exchangeable in the pairs $(x, y)_i$" (Gelman et al., 2014, page 45). [1-2]

4.2 Prior Distribution for θ

Since the poisson distribution is a one-parameter exponential distribution, a conjugate prior distribution exists. The conjugate prior of a Poisson distribution is a Gamma distribution with parameters $\alpha > 0$ and $\beta > 0$. The parameters of the conjugate prior are estimated using movies that are similar in style to Quentin Tarantino's movies. The prior data was gathered by searching "directors similar to Tarantino"[4]. The following table provides information of the movies used to estimate the parameters of the prior distribution.

Film	Length	Body Count	Body Count / Hour
Battle Royale	1h 53min	42	22.3
The Boondock Saints	1hr 50min	34	18.5
City of Fire	1hr 45min	10	5.7
Get Shorty	1hr 45min	4	2.3
Lucky Number Slevin	1hr 50min	19	10.4
The Way of the Gun	1hr 59min	20	10.1

Table 2: Films similar in style to Tarantino [5 - 10]

From the movies in a similar style to Tarantino, the body count per movie is between 4 and 42 and the mean body count per hour is 11.55 with variance 57.41. Since the expected value and the variance of a Gamma distribution with parameters α and β is $\frac{\alpha}{\beta}$ and $\frac{\alpha}{\beta^2}$ respectively, we are able to estimate α as 2.32 and β as 0.2.

4.3 Posterior Distribution for θ

Since we are using a conjugate prior for the analysis, the posterior distribution will also be a Gamma distribution with parameters $\alpha + \sum_{i=1}^{8} y_i$ and $\beta + \sum_{i=1}^{8} x_i$, where α and β are the parameters of the prior distribution.

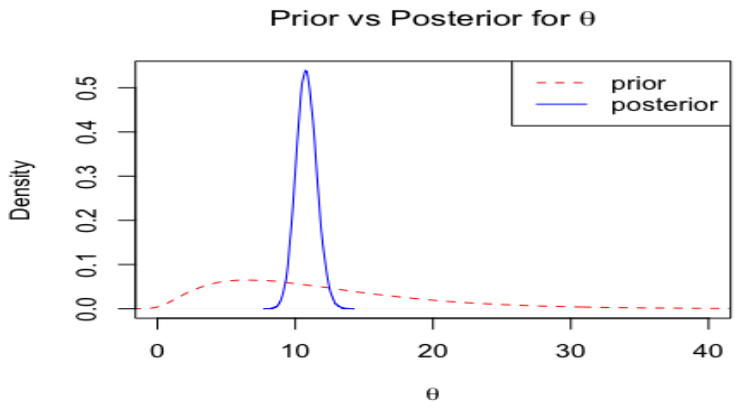

Figure 2: pdf of prior and posterior distribution. The variance is reduced in the posterior distribution for θ.

The plot above compares the prior and posterior distribution of θ. The variance shrinks from the prior to posterior, indicating our beliefs in theta are stronger after seeing the data. Thus our data is reinforcing our prior belief. It is reassuring to see that the data has a strong influence on the posterior distribution for θ. The 95% highest probability density

interval for θ, [9.46, 12.30], summarizes what we see in the plot of the prior and posterior distributions.

4.4 Posterior Predictive Distribution and Prediction

To predict future body counts in Quentin Tarantino's next movie $\tilde{y}$, the distribution of $\tilde{y}$ is necessary. The posterior predictive distribution is calculated by integrating the joint distribution of $\tilde{y}$ and θ given the data over the parameter space Θ,

$$p(\tilde{y}|y) = \int_{\Theta} p(\tilde{y}, \theta|y)d\theta = \int_{\Theta} p(\tilde{y}|\theta)p(\theta|y)d\theta$$

The posterior predictive distribution for a Poisson sampling distribution with parameter $\theta > 0$ and a Gamma posterior distribution with parameters $\alpha + \sum y_i > 0$ and $\beta + \sum x_i > 0$ is a Negative Binomial distribution (Appendix B).

$$\tilde{y} \sim NB(\alpha + \sum_{i=1}^{8} y_i, \frac{1}{\tilde{x}}[\beta + \sum_{i}^{8} x_i])$$

where $\tilde{x} > 0$ is the exposure of $\tilde{y}$.

The figure below shows the probability density function of $\tilde{y}$ for a given movie length. As the length of the movie increases, there is more variability in predicting the number of deaths in a movie. This is reiterated in the 95% HPD intervals for $\tilde{y}$ as seen in the table below.

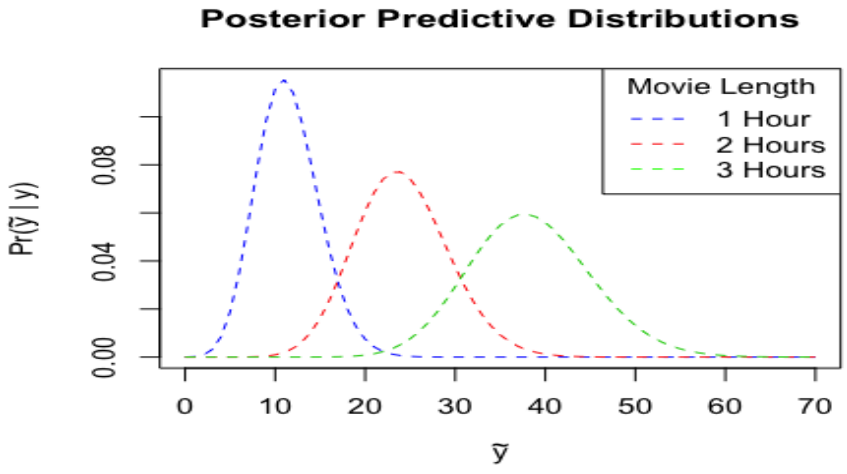

Figure 3: Three estimates for the number of deaths (x-axis) in a future Tarantino movie: an estimate for a one-hour movie, a two-hour movie, and a three-hour movie.

Movie Length	95% HPD Interval
1 Hour	[5, 18]
2 Hours	[13, 33]
3 Hours	[24, 50]

Table 3: The 95% probability of the true death count in Tarantino's next two-hour movie is between 13 and 33. (Appendix C)

4.5 Model Checking

To check if our sampling model is appropriate, aspects of the sampling model will be assessed using test quantities. The aspects that will be assessed will be the variance-to-mean ratio, the maximum, and the minimum of body count of the data.

For the Poisson distribution, the variance-to-mean ratio should be 1. However, in our case the observed variance-to-mean ratio is around 20, which tells us that the Poisson distribution is not capturing our data effectively. Using Monte Carlo methods to generate samples from the assumed Poisson model, a distribution is generated for the variance-to-mean ratio.This is shown in Figure 4 along with a red line for our observed quantity which clearly shows it is not expected from our chosen sampling distribution.

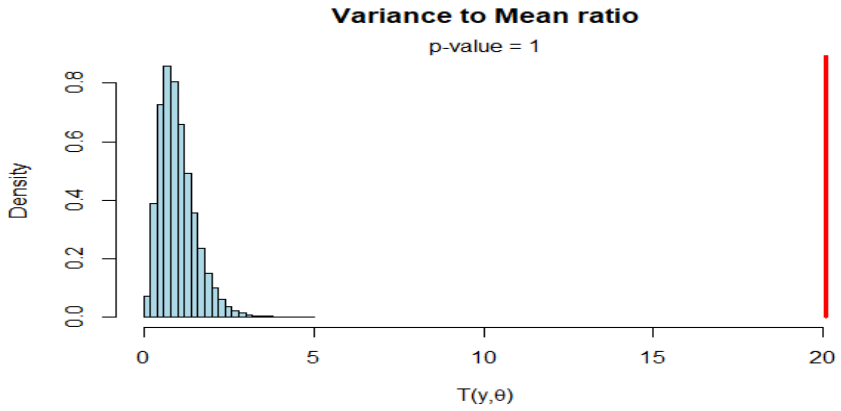

Figure 4: Distribution of simulated variance-to-mean ratio and observed value

We follow this process for generating distributions for the maximum and minimum body count. When we compare the observed maximum value to the simulated maximum values, the p-value is 1. This indicates that the observed pattern from the data is very unlikely to occur from the replicated data under the assumed Poisson distribution, see Figure 5. When we compare the observed minimum with the simulated minimum, the p-value is 0.951, which similarly indicates the minimum observed value is unlikely to occur from the replicated data, see Figure 5.

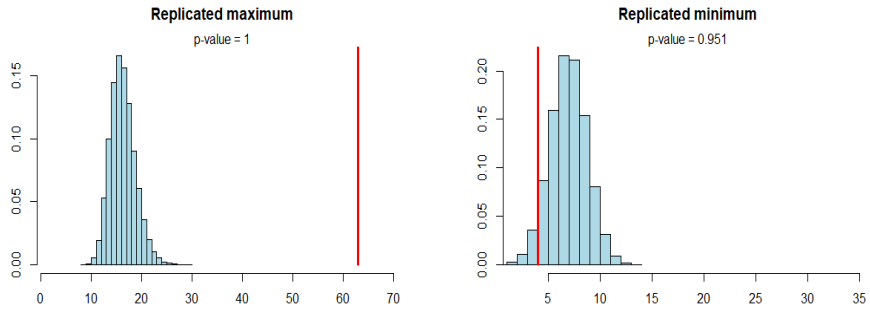

Figure 5: Distribution of simulated maximum and minimum with observed value

5 Conclusion

Our goal was to estimate the number of deaths in the next Tarantino movie. We estimate a new two-hour movie by Tarantino will have between 13 and 33 deaths with 95 percent accuracy. When modeling the distribution of the total count of occurrences (for example, number of deaths), the Poisson distribution seems a likely approach. Thus this led us to consider the number of deaths (per movie) to be modeled by a Poisson distribution, with the rate parameter being deaths per hour. A feature of the Poisson distribution is that the mean and variance are equal. After considering the data however, the variance (of deaths per movie) is not similar to the mean (number of deaths per hour); the variance is much larger than the mean. Because of this over-dispersion, we would like to investigate further the use of a Negative Binomial sampling distribution instead of using the Poisson sampling distribution. The Negative Binomial distribution provides a more robust model for this situation, allowing "the mean and variance to be fitted separately, with variance at least as great as the mean" (Gelman et al., 2014 page 437-438)

6 Appendix

6.1 Appendix A

Poisson Model Assumptions [11]:

- In general, let $y(t)$ denote the number of events that have occurred during a time interval $[0, t]$.
- Assumptions:

  P1. $y(0) = 0$.

  P2. For all $n \geq 0$, and for any two time intervals, I_1 and I_2, of equal length, $\Pr(n \text{ events in } I_1) = \Pr(n \text{ events in } I_2)$.

  P3. Events that occur in nonoverlapping time intervals are mutually independent.

  P4. $\lim_{h \to 0} \frac{\Pr\{y(h) > 1\}}{h} = 0$.

- Assumptions (cont'd):

  P5. $0 < \Pr\{y(t) = 0\} < 1$ for all $t > 0$.

- Under conditions P1–P5, there exists $\theta > 0$ such that

$$\Pr\{y(t) = y\} = \begin{cases} \frac{(\theta t)^y e^{-\theta t}}{y!} & \text{for } y = 0, 1, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

6.2 Appendix B

Let a' $= \alpha + \sum y_i > 0$, b' $= \beta + \sum x_i > 0$ and assume $\tilde{x} > 0$.

$$\begin{aligned}
p(\tilde{y}|y) &= \int_0^\infty p(\tilde{y}|\theta) p(\theta|y) d\theta \\
&= \int_0^\infty \frac{(\tilde{x}\theta)^{\tilde{y}} e^{-\tilde{x}\theta}}{\tilde{y}!} \frac{(b')^{a'}}{\Gamma(a')} \theta^{a'-1} e^{-b'\theta} d\theta \\
&= \frac{\tilde{x}^{\tilde{y}} (b')^{a'}}{\Gamma(a')\Gamma(\tilde{y}+1)} \int_0^\infty \theta^{\tilde{y}+a'-1} e^{-(b'+\tilde{x})\theta} d\theta \\
&= \frac{\tilde{x}^{\tilde{y}} (b')^{a'}}{\Gamma(a')\Gamma(\tilde{y}+1)} \frac{\Gamma(\tilde{y}+a')}{(b'+\tilde{x})^{\tilde{y}+a'}} \\
&= \frac{\tilde{x}^{\tilde{y}} (b')^{a'}}{\Gamma(a')\Gamma(\tilde{y}+1)} \frac{\Gamma(\tilde{y}+a')}{(b'+\tilde{x})^{a'}(b'+\tilde{x})^{\tilde{y}}} \\
&= \frac{\Gamma(\tilde{y}+a')}{\Gamma(a')\Gamma(\tilde{y}+1)} (\frac{b'}{b'+\tilde{x}})^{a'} (\frac{\tilde{x}}{b'+\tilde{x}})^{\tilde{y}} \\
&= \frac{\Gamma(\tilde{y}+a')}{\Gamma(a')\Gamma(\tilde{y}+1)} (\frac{b'/\tilde{x}}{b'/\tilde{x}+1})^{a'} (\frac{1}{b'/\tilde{x}+1})^{\tilde{y}}
\end{aligned}$$

Therefore

$$\tilde{y} \sim NegativeBinomial(a', \frac{b'}{\tilde{x}})$$

$$\tilde{y} \sim NegativeBinomial(\alpha + \sum y_i, \frac{\beta + \sum x_i}{\tilde{x}})$$

6.3 Appendix C

Monte Carlo Method to generate simulated predictive values using R code

Draw θ from p($\theta|y$):
theta <- rgamma(1000, a+n*ybar)/ (b+n*xbar)
xtilde <- 2 #length in hours of future Tarantino movie

Draw $\tilde{y}$ from p($\tilde{y}|\theta, y$):
ytilde <- rnbinom(1000, size = a+n*ybar, prob=1- 1/((1/xtilde)*(b+n*xbar)))

7 Reference

[1] Gelman, Andrew, et al. Bayesian Data Analysis. CRC Press, 2014.

[2] "Poisson Distribution", https://www.umass.edu/wsp/resources/poisson/#paper, Accessed 21 Nov 2018.

[3] "Exponential family: conjugate priors", https://people.eecs.berkeley.edu/∼jordan/courses/260-spring10/other-readings/chapter9.pdf, Accessed 23 Nov 2018.

[4] "If You Love Quentin Tarantino Films, You Should Also Check Out These 15 Movies" https://www.scoopwhoop.com/Quentin-Tarantino-Films-Also-Check-Out-These-15-Movies/#.vt5hftcui, Accessed 29 Oct 2018.

[5] Battle Royale, http://homepage.tinet.ie/∼screamanthology/br/deaths.htm, Accessed on 29 Oct 2018.

[6] The Boondock Saints, http://www.allouttabubblegum.com/main/?p=4122, Accessed on 29 Oct 2018.

[7] Get Shorty, http://moviebodycounts.proboards.com/thread/1817/get-shorty-1995, Accessed on 29 Oct 2018.

[8] Lucky Number Slevin, http://www.moviebodycounts.com/Lucky_Number_Sleven.htm, Accessed on 29 Oct 2018.

[9] The Way of The Gun, http://www.allouttabubblegum.com/forum/viewtopic.php?f=16&t=7908&st
Accessed on 29 Oct 2018.

[10] City of Fire, http://www.allouttabubblegum.com/main/?p=11211, Accessed on 25 Nov 2018.

[11] Lecture-D Slides of Bayesian Analysis, taught by Dr. Bee Leng Lee at San Jose State University