

MATH 264: Bayesian Project

Terry Situ, Caie Yan, Ryan Quigley

December 1, 2016

I. Executive Summary

Quentin Tarantino’s moves are famous (or notorious) for violent death and massacre. In this project, we use Bayesian methods to analyze and predict the body count in Quentin Tarantino’s next movie: not a simple task for such an eccentric and unpredictable filmmaker. Even though he has a reputation for violence, the amount and type vary widely from movie to movie. With this in mind, we realized that his collection of movies did not all belong to a single population; there is a natural separation within the collection which suggests that the data be separated into three different groups. As a result, we created three separate models, one for each group. Given the nature of the data, we chose the Poisson model for our sampling distributions. Using an external dataset containing the body counts and metadata of about 200 movies, we explored the possible factors, such a genre, rating, and release year, that could be associated with body count of a movie. Among all the factors, the association of genre with body count is most notably. This allowed us to develop informed conjugate prior relevant to the data within each models. Our results show that for the same movie duration, if Quentin Tarantino’s next movie is about a war, the body count could be as many as 38 times a non-war related movie.

II. Data Summary

Tarantino Filmography (as writer and director)

The table below provides the background information of nine movies directed and written by Tarantino. The movies included in each model are indicated in the Model column. Metadata extracted from IMDB (<http://www.imdb.com/>); body count extracted from Vanity Fair infographic (Beggs & Handy, 2013)

Model	Movie	Genre	Year	Rating	Body Count	Hours	Kill Rate/Hr
1	Reservoir Dogs	crime, drama, thriller	92	R	11	1.65	6.67
1	Pulp Fiction	crime, drama	94	R	7	2.57	2.73
1	Jackie Brown	crime, thriller	97	R	4	2.57	1.56
2	Kill Bill Vol. 1	action, thriller	03	R	62	1.85	33.51
1	Kill Bill Vol. 2	action, crime drama, thriller	04	R	13	2.28	5.69
1	Death Proof	thriller	07	NR	6	1.88	3.19
3	Inglorious Basterds	adventure, drama, war	09	R	396	2.55	155.29
2	Django Unchained	drama, western	12	R	64	2.75	23.27
1	The Hateful Eight	crime, drama, mystery thriller, western	15	R	18	2.78	6.47

From the article where the infographic was originally published:

“A few numbers are approximated due to the impossibility of counting precisely how many ninjas are decapitated in **Kill Bill Vol. 1**, how many Nazis are in the theater when it gets set afire in **Inglorious Basterds**, and how many people fall in the never-ending shoot-out scene at the end of **Django Unchained**.”

- Vanity Fair

The body count numbers in these three movies are unusually large and require approximation for one main reason: each contains a single scene of unimaginable (to most people except Tarantino) violence that dramatically increases the death toll. In light of this information, we find it impossible to judge exchangeability for the entire nine movie filmography of Tarantino as both writer and director; however, we are not attempting to use this as evidence for throwing out the data points as outliers. Instead, given our limited knowledge of alternative Bayesian models, we have chosen to group the data into three distinct sets, and we have fit a model for each group. All movies not described in the quote above are grouped into model 1; the other three movies (in bold in the quote) are further separated into two groups based on the binary genre classification war/non-war. This justification was based on

analysis of the historical movie sample, which indicated that body counts and kill rates per hour were dramatically higher in the war genre than all other genres that Tarantino movies fall under. The movies included in each model are indicated in the *Model* column of the Tarantino filmography table above.

III. Models

The following table summarizes the model components, and provides two posterior summary statistics for θ (Note: $\Gamma(\alpha, \beta) = \text{Gamma}(\alpha, \beta)$).

Model	Prior	Likelihood	Posterior	Posterior Mode	99% HPD Interval
1	$\Gamma(1.46, 0.053)$	$\propto \theta^{59} e^{-13.73\theta}$	$\Gamma(60.46, 13.79)$	4.31	[3.03, 5.92]
2	$\Gamma(2.13, 0.064)$	$\propto \theta^{126} e^{-4.6\theta}$	$\Gamma(128.13, 4.66)$	25.26	[21.49, 33.97]
3	$\Gamma(2.015, 0.023)$	$\propto \theta^{396} e^{-2.55\theta}$	$\Gamma(398.015, 2.57)$	154. 3	[135.20, 175.12]

Sampling Distribution: Poisson (Rate and Exposure)

The Poisson model parameterized in terms of rate and exposure is,

$$p(y \mid \theta) = \frac{1}{y!} (\theta t)^y e^{-(\theta t)} \cdot 1_{\{0,1,2,\dots\}}(y)$$

For our analysis, θ is the unknown kill rate per hour. According to Gelman et al. (2003, pg. 45), this model is not exchangeable in the y_i 's but is exchangeable in the pairs $(t, y)_i$. Due to the genre similarity within each model, we do not have any reservations about judging exchangeability for the data points within each model.

Assumptions and Justification

There are a number of assumptions associated with the Poisson model that typically need to be met in order to justify using the model. In the case of observations that represent deaths per movie, at least two of the assumptions appear to be violated. First, imagine splitting a movie into two non-overlapping intervals of equal length such that they cover the entire length of the movie. Based on the general plot structure of movies, the probability that you would have a high number of kills in the first interval would be much lower than the probability of having a high number of kills in the second half, which would likely contain the climax and most of the action. Similarly, a low number of kills would be more likely in the first half than the second. Thus, the probability of n kills in the first interval does not equal the probability of n kills in the second interval.

Now imagine taking two non-overlapping but back-to-back intervals during the climax of a movie. If people start getting killed in the first interval, the chances of continued killing in the next interval will be significantly increased. Thus, we cannot reasonably assume non-overlapping are mutually independent.

Additionally, the assumption that the simultaneous occurrence of two or more events is impossible seems tenuous. For example, movie explosions seem to imply that multiple people die instantaneously and simultaneously. We admit an argument could be made for differences in nanoseconds, but it would be nearly impossible to measure and verify. Nevertheless, the overall validity of the assumptions is questionable enough to preclude the use of the Poisson model. Therefore, we turn to the General Representation Theorem, with the additional necessary and sufficient condition developed by Freedman, in order to justify that the observations are conditionally independent given θ .

For model 1, if we imagine dropping 59 balls into 6 boxes each with probability 1/6, the result (11, 7, 4, 13, 6, 18) seems completely reasonable. Similarly for model 2, (62, 64) is a likely result of dropping 126 balls into 2 bins with 1/2 probability each. Model 3 does not require this justification because it only has one data point. To provide a more concrete justification, we simulated 10,000 draws from the multinomial distribution for each model and plotted

the results in the section *Multinomial Condition Check* of the appendix. The plots further support our conclusion that the condition is met. Therefore, the likelihood for each model is of the form $p(y | \theta) \propto \theta^{\left(\sum_{i=1}^n y_i\right)} \exp\left(-\theta \sum_{i=1}^n t_i\right)$

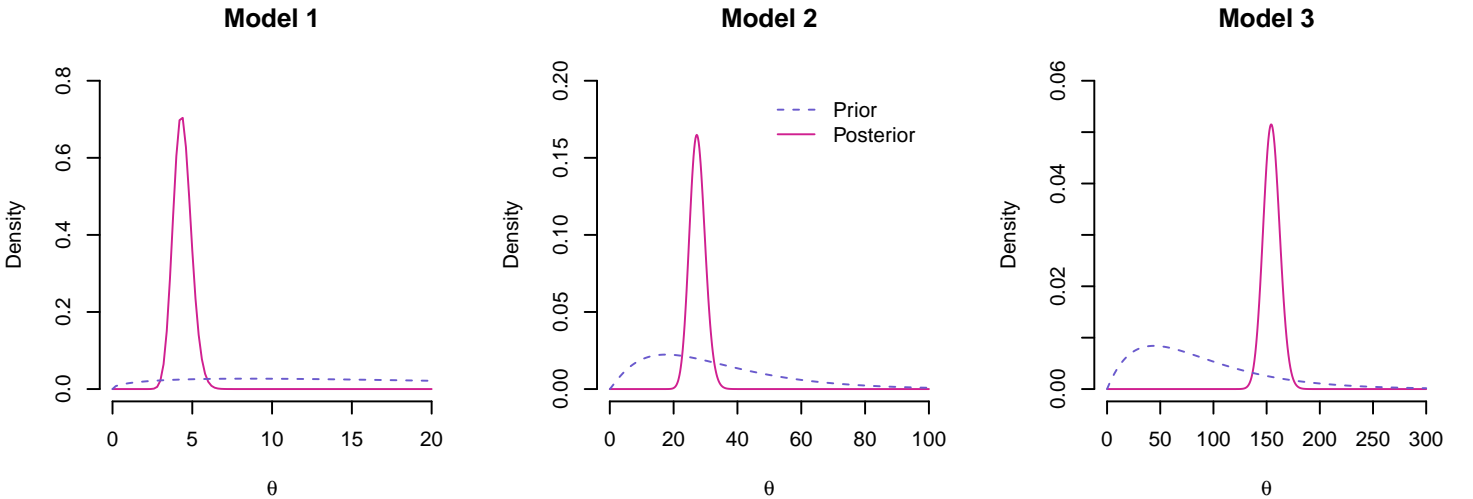
Conjugate Prior: Gamma

We chose to use conjugate priors for our models because data on body counts per movie was readily available thanks to the dedication of movie buffs on the internet. Since the conjugate prior distribution for the Poisson sampling distribution is the Gamma distribution, the prior distribution of θ will be of the form $p(\theta) \propto \theta^{\alpha-1} e^{-\beta \cdot \theta}$

In order to estimate the parameters of the conjugate prior distribution for each model in the next section, we acquired a dataset compiled by Randal Olson from the site <http://www.moviebodycounts.com/>. See references for links to original dataset, and the *Model Derivations* section of the appendix for a detailed summary of the historical data used. The parameters of the conjugate prior distributions for each model were calculated by solving analytically a pair of equations for α and β (Lee, 2016). The equations were determined by (1) setting the mode formula for the gamma distribution equal to the mode of the kernel density estimate, and by (2) observing the interval that approximately 99.9% of the historical data subset was contained in. For modes of the kernel density estimates are 8.61, 17.85, and 43.83, respectively. The intervals that appeared to contain 99.9% of the historical data were (0,150), (0,150), and (0,400), respectively.

Posterior: Gamma

In general, the posterior distribution is $\text{Gamma}\left(\alpha + \sum_{i=1}^n y_i, \beta + \sum_{i=1}^n t_i\right)$ (Gelman et al., 2003, pg. 45). The parameterization of each posterior distribution is summarized in the table at the top of this section, and plots of the prior and posterior are included below for additional clarity. Please take note of the different scale of the vertical axis in each plot. For a closer look at the prior distributions for each model see the *Model Derivations* section of the appendix.



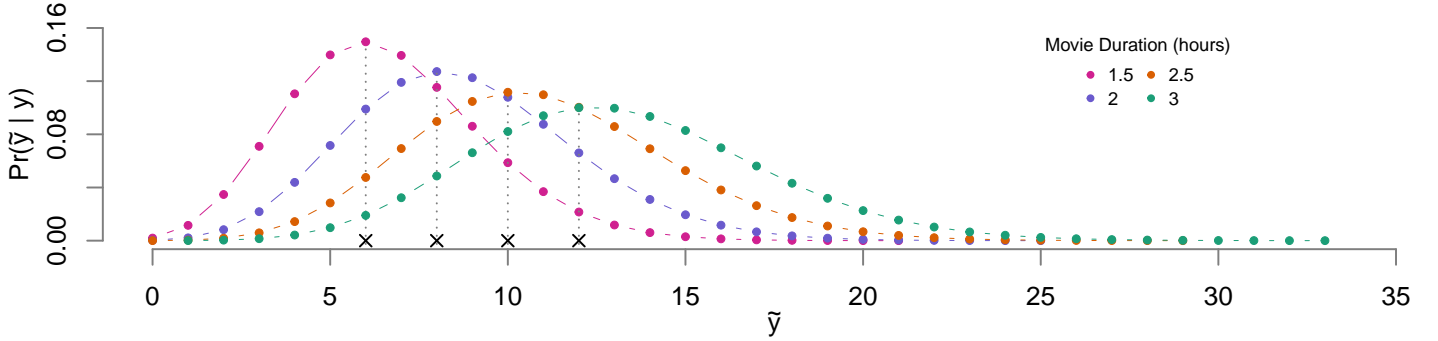
Posterior Predictive: Negative Binomial

The posterior predictive distribution for a single additional observation is a negative binomial distribution of the form $\text{NB}\left(\alpha + \sum_{i=1}^n y_i, \frac{1}{t} \left[\beta + \sum_{i=1}^n t_i\right]\right)$ (Gelman et al., 2003, pg. 44-45). See the *Model Derivations* section of the appendix for details. Taking the average movie duration for each model, the most likely body count values broken down by model in Quentin Tarantino's next movie are summarized in the following table.

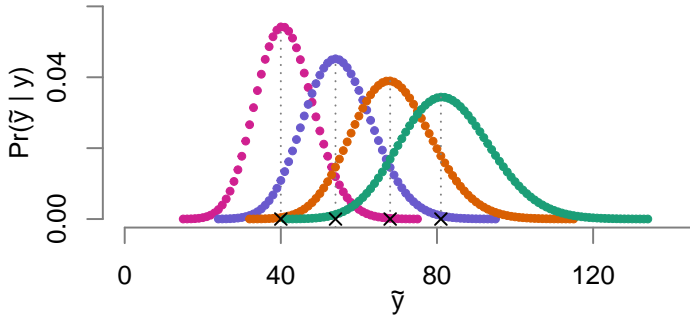
Model	Movie Duration (hours)	Body Count	$\Pr(\tilde{y} y)$
1	2.3	9	0.117
2	2.3	62	0.041
3	2.5	385	0.014

The plots show the full posterior predictive distribution of each model for a sample of easily interpretable movie durations. For each model, the spread of the distribution increases with movie duration meaning our predictions for the next movie are less precise.

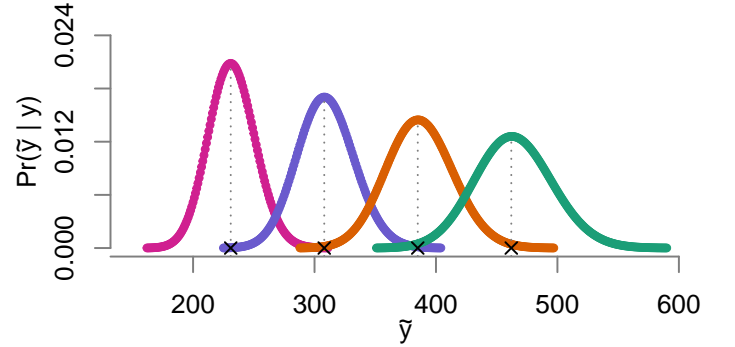
Model 1



Model 2

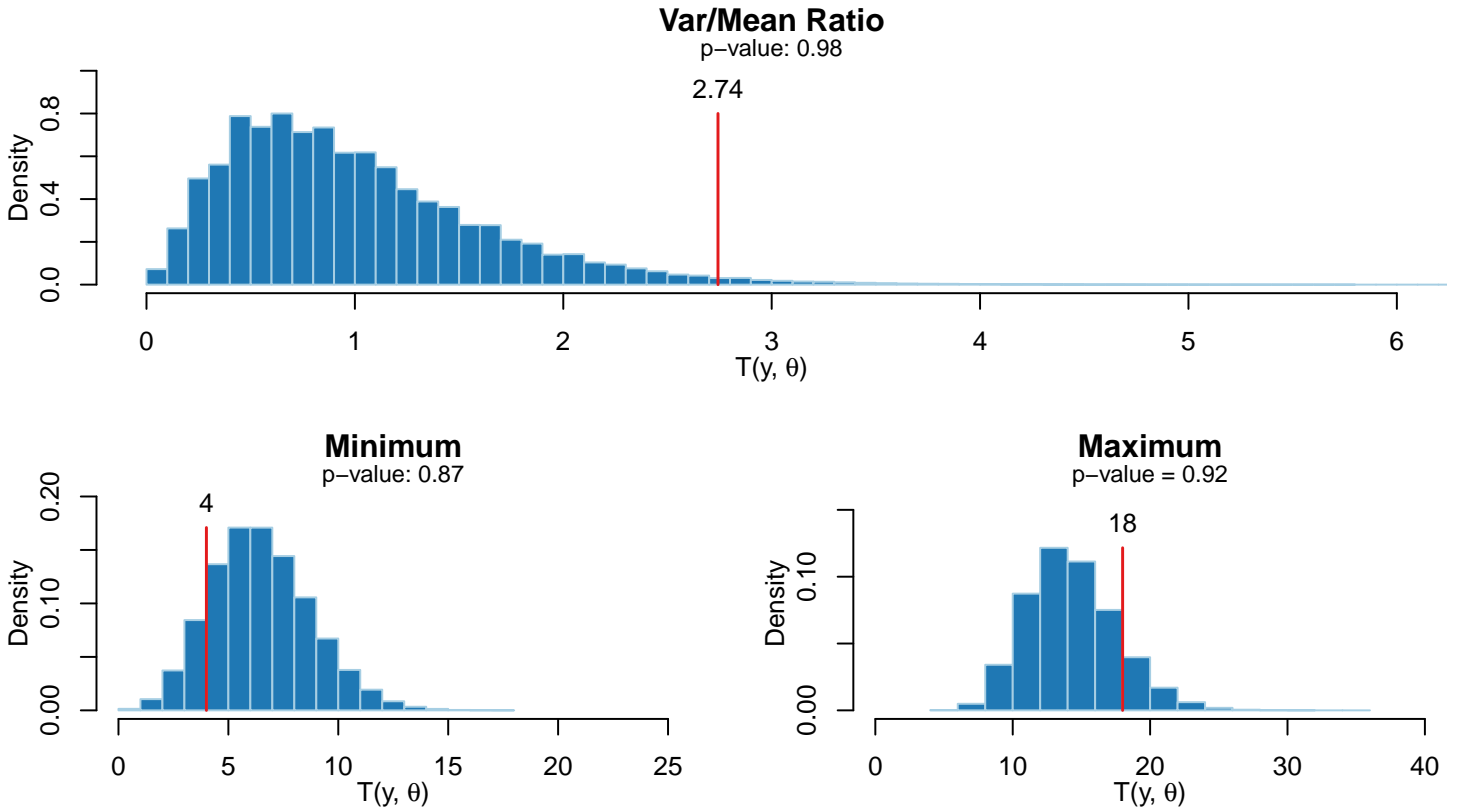


Model 3



Posterior Predictive Checking

The following plots summarize the results of three test quantities for model 1: the minimum, maximum, and the ratio of the sample variance to the sample mean. Note, only model 1 has enough data points to calculate meaningful test quantities, so posterior predictive checking was not performed for models 2 and 3. The red vertical line in the plots indicates the test quantity value for the observed data; the blue histogram represents the test quantity evaluated for 500,000 replications of simulated data. See the *Model Checking* section of the appendix for details on the simulation procedure and the code used to generate the replications. Observed individually, the plots for minimum and maximum do not suggest strong discrepancies between the model and the observed data, but interpreted together they do raise some concern about the spread of the data. Too much spread in the data suggests overdispersion, which in turn may mean the Poisson model is not appropriate. An interesting feature of the Poisson sampling distribution is that the expectation and variance are equal. Thus, we would expect the ratio of sample variance to sample mean to be around one if the data is truly from the Poisson distribution. The plot of variance-mean ratio shows a significant discrepancy between the observed value and the simulated values. Thus, our model is not capturing the variance in our data well. A model that allows the variance to be fit separately from the mean, such as negative binomial or normal, would likely provide a better fit to our data.



Conclusions

The approach we took for this analysis is unorthodox: it is uncommon to break a dataset into distinct groups and fit a model to each individual group. However the dataset is peculiar and it demands special treatment, perhaps using more advanced Bayesian modelling methods unknown to us. For example, we know the negative binomial sampling distribution can be used as a robust alternative to the Poisson distribution to help account for the overdispersion of the data. Since we have presented an argument for grouping the data, it seems reasonable to assume that a hierarchical model could fit the data particularly well.

If readers are uncomfortable with the data grouping decision, they may be more accepting of model 1 if viewed as a model for the entire data set where three large outliers have been removed. The parameters of the distributions and the predicted values will be the same, but the interpretation will be slightly different. That being said, we believe our justification for breaking up the dataset is sound given the knowledge we have about the data. Observing only the body count numbers of Tarantino movies, it is clear that the majority of the data is clustered around 10 deaths; that is, the majority of his movies do not have epic scenes of death where the bodies fall so fast that it is difficult to count. Intuitively, it seems likely that his next movie will not be dominated by an epic death scene. Thus, our prediction from model 1 of 9 deaths in the next movie seems reasonable.

But! Tarantino has been quoted saying that he only plans to make 10 movies in total (Robinson, 2013). If we combine the *Kill Bills* into one movie, which most Tarantino fanatics do, that puts us at 8 movies produced so far. Could he end his saga with another two volume film where he pulls out all the stops and introduces more violence and death than ever before? Quite possibly. Thus, it's nice to know we have a model to address this possibility, and model 2 tells us that 62 deaths is the most likely value in this case. And finally, if he attempts another war themed blood bath, then model 3 tells us 382 deaths is most likely what we are going to get so we better prepare ourselves for a gore fest.

IV. Appendix

Table of Posterior Predictions for Next Movie

For a sample of potential movie durations, the most likely body count values broken down by model in Quentin Tarantino's next movie are summarized in the following table.

Model	Movie Duration (hours)	Body Count	$\Pr(\tilde{y} y)$
1	1.5	6	0.15
	2	8	0.127
	2.5	10	0.112
	3	12	0.1
2	1.5	40	0.054
	2	54	0.045
	2.5	68	0.039
	3	81	0.034
3	1.5	231	0.021
	2	308	0.017
	2.5	385	0.014
	3	462	0.013

Multinomial Condition Check

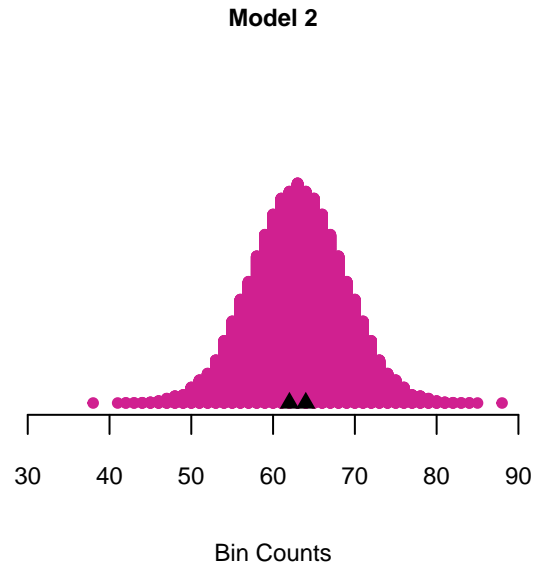
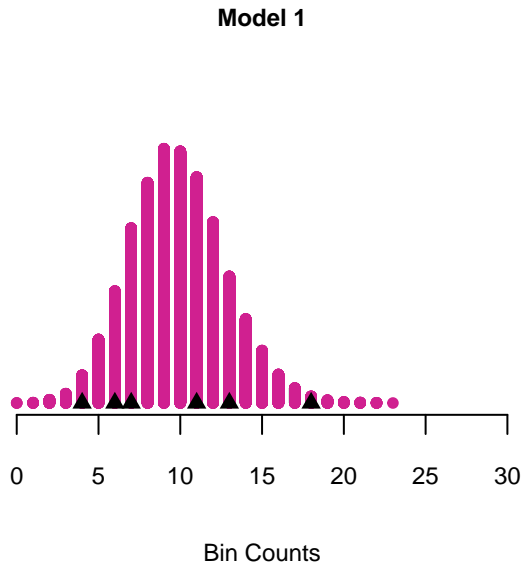
In order to justify that y_i, \dots, y_n are c.i.i.d poisson observations given θ , a necessary and sufficient condition is that

$$p(y_1, \dots, y_n | s_n) = \frac{s_n!}{y_1! \dots y_n!} \prod_{i=1}^n \left(\frac{1}{n}\right)^{y_i}$$

for every n , where $s_n = y_1 + \dots + y_n$.

Note, this condition is not reasonably justified if we group all the data together; however, once grouped into three models, the condition seems reasonable within each model. The left panel in the following plot displays 10,000 simulations of dropping 59 balls in 6 equally likely bins; the right panel, simulates dropping 126 balls in 2 equally likely bins. The black triangle indicate the actually data points used for that model.

```
mltnom.check.m1 <- rmultinom(n = 10000, prob = rep(1/6, 6), size = 59)
mltnom.check.m2 <- rmultinom(n = 10000, prob = rep(1/2, 2), size = 126)
```

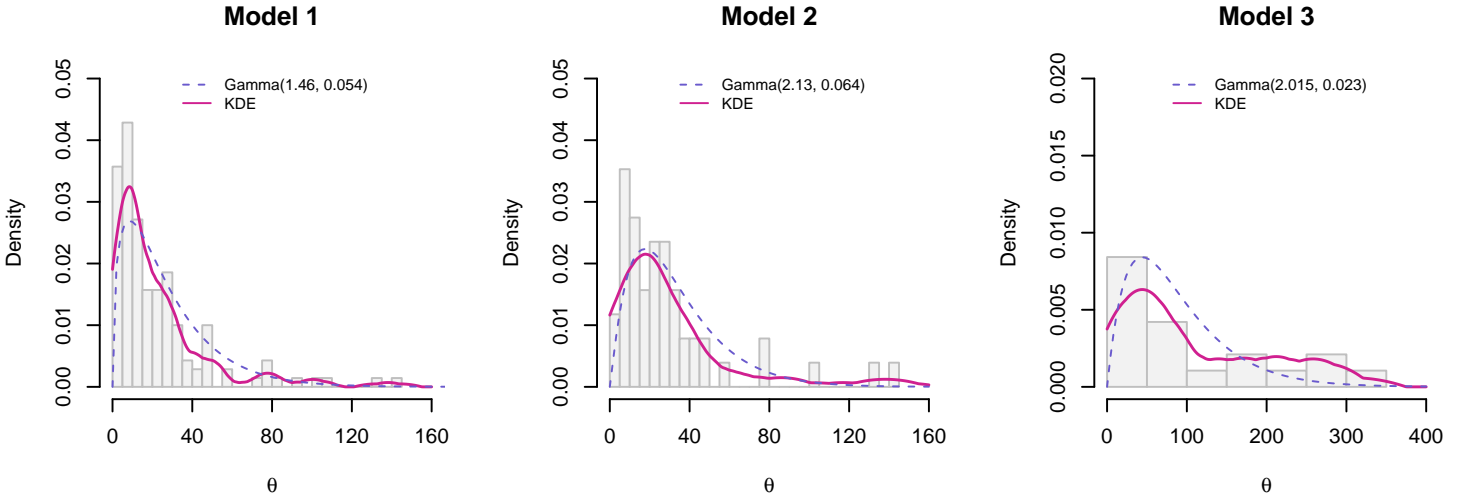


Model Derivations

Conjugate prior: gamma

The following is a detailed description of the filtering applied to the original historical movie sample. First all movies with Quentin Tarantino as director were removed. This excluded one movie not in the dataset described above where Tarantino was a director but not a writer: *Sin City*. Next, we restricted the range of years to exclude any movies released prior to 1989. This was done to avoid influence from movies in a time period with dramatically different social views about movie violence. We assumed movies released in the three years prior to the release of his first movie (*Reservoir Dogs*, 1992) would also be similar enough in nature. All of the Tarantino movies have an MPAA rating of R with the exception of *Death Proof*, which was unrated. As a result, we included movies with ratings R or Unrated. The filter conditions discussed so far apply to all three models, but the filtering based on genre is specific to each model. For both model 1 and model 2, the unique set of genres was determined for the data points in each model. For model 1, the set consists of crime, drama, thriller, action, western, mystery; for model 2, action, thriller, drama, western. Using these sets, a movie from the historical sample was included if one of two conditions was met: (1) all its genres matched the unique genre set, or (2) at least 3 of its genres matched the unique genre set. The number of genres listed for a movie can vary quite a bit, so these conditions help prevent the filtering from excluding too many movies. For Model 3, the genre filter condition is simply a check to see if the movie has the genre war. This condition is far less restrictive than those for models 1 and 2, but is necessary due to the small number of war movies with recorded body counts (we suspect this is due to the difficulty and tedium of recording body counts for war movies). The influence war has on body count will outweigh the influence of any other genre also listed, so we believe this less restrictive condition for model 3 is not problematic. One final note: the original historical movie dataset does not contain any movies with zero deaths. For our purposes, this ensures the kill rate per hour is greater than zero, which is appropriate for the support of the Gamma distribution. The following table summarizes the kill rate per hour of the subset of the historical data used for each model.

Model	N	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	140	0.49	6.28	14.11	22.65	27.77	143.9
2	51	1.52	10.81	21.94	29.59	34.00	143.9
3	19	9.80	43.50	58.94	106.80	167.60	307.7



The system of equations used to determine the parameters of the gamma distribution from the historical data is,

$$\text{Mode}(\theta) = \frac{\alpha - 1}{\beta} \quad (1)$$

$$0.999 = \int_0^u \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} \quad (2)$$

Posterior predictive distribution: negative binomial

Let $\alpha' = \alpha + \sum_{i=1}^n y_i$ and $\beta' = \beta + \sum_{i=1}^n t_i$

$$\begin{aligned} p(\tilde{y} \mid y) &= \int_0^\infty p(\tilde{y} \mid \theta) \cdot p(\theta \mid y) \, d\theta \\ &= \int_0^\infty \text{Poisson}(\tilde{y}(t) \mid \theta) \cdot \text{Gamma}\left(\alpha + \sum_{i=1}^n y_i, \beta + \sum_{i=1}^n t_i\right) \, d\theta \\ &= \frac{\beta'}{\Gamma(\alpha')} \tilde{t}^{\tilde{y}} \cdot 1_{\{0,1,\dots\}}(\tilde{y}) \int_0^\infty \theta^{(\alpha'+\tilde{y})-1} e^{-(\beta'+t)\theta} \, d\theta \\ &= \frac{\beta'}{\Gamma(\alpha')} \tilde{t}^{\tilde{y}} \cdot 1_{\{0,1,\dots\}}(\tilde{y}) \cdot \frac{\Gamma(\alpha' + \tilde{y})}{(\beta' + \tilde{t})^{\alpha'+\tilde{y}}} \\ &= \frac{\Gamma(\alpha' + \tilde{y})}{\tilde{y}! \Gamma(\alpha')} \left(\frac{\beta'}{\beta' + \tilde{t}}\right)^{\alpha'} \left(\frac{\tilde{t}}{\beta' + \tilde{t}}\right)^{\tilde{y}} 1_{\{0,1,\dots\}}(\tilde{y}) \\ &= \frac{\Gamma(\alpha' + \tilde{y})}{\tilde{y}! \Gamma(\alpha')} \left(\frac{\beta'/\tilde{t}}{\beta'/\tilde{t} + 1}\right)^{\alpha'} \left(\frac{1}{\beta'/\tilde{t} + 1}\right)^{\tilde{y}} 1_{\{0,1,\dots\}}(\tilde{y}) \\ &= \text{NB}\left(\alpha + \sum_{i=1}^n y_i, \frac{1}{\tilde{t}} \left[\beta + \sum_{i=1}^n t_i\right]\right) \end{aligned}$$

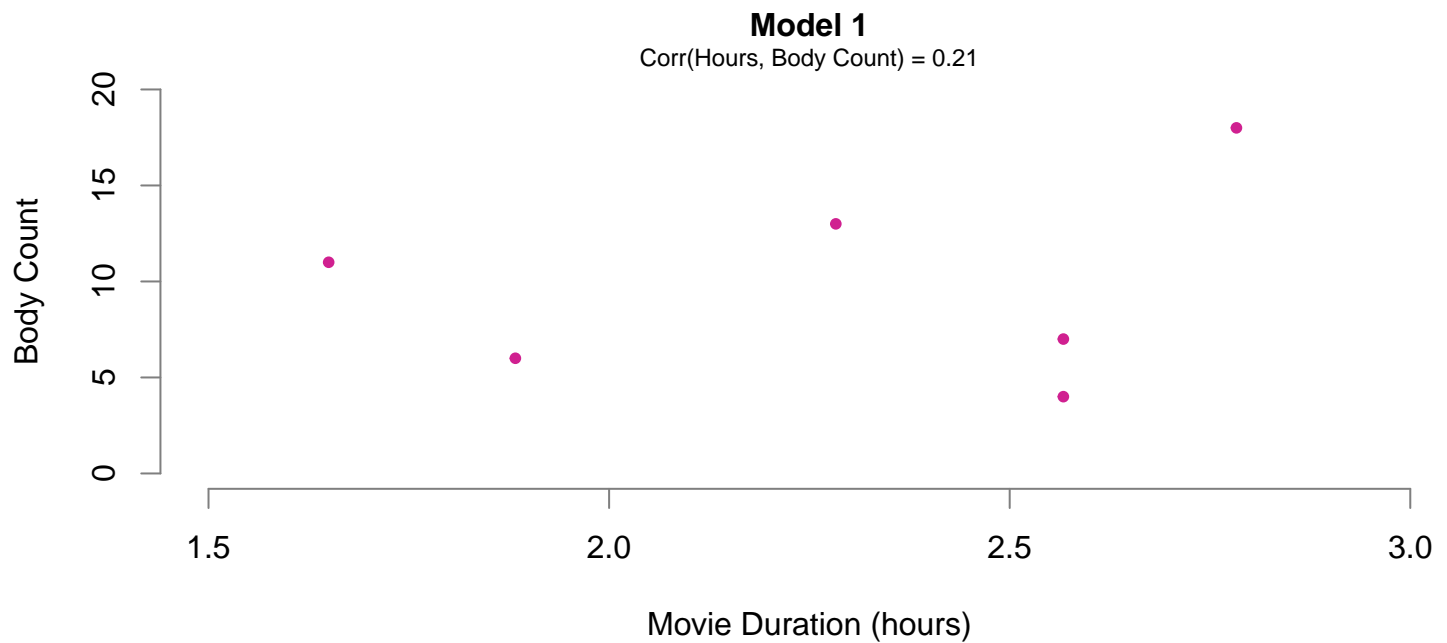
Model Checking

Posterior predictive checking: simulation procedure

Because our sampling distribution involves rate and exposure, we must appropriately account for the exposure, t , when approximating the distribution of $T(y^{rep}, \theta)$ using simulation. The simulation will be performed as follows:

1. Draw $\theta^{(i)}$ from $\text{Gamma}\left(\alpha + \sum_{i=1}^n y_i, \beta + \sum_{i=1}^n t_i\right)$
2. Multiply $\theta^{(i)}$ by t^*
3. For each $(\theta^{(i)} \cdot t^*)$, simulate a draw $y^{rep(i)}(t^*)$ from $\text{Poisson}(\theta^{(i)} \cdot t^*)$
4. Calculate $T(y^{rep(i)}(t^*), \theta^{(i)})$ for $i = 1, \dots, 500000$.

For model 1, this presents a problem when deciding what value of t^* to choose for each simulation. To address this issue, we note that there does not appear to be a strong relationship between body count and movie duration, which is supported by a low correlation value of 0.21 and the plot below. As a result, we draw t^* randomly from $U(\min\{t_1, \dots, t_n\}, \max\{t_1, \dots, t_n\})$ in step 2 of the simulation procedure. In the case of model 1, this is the interval $[1.65, 2.78]$.



Posterior predictive checking: simulation code

```
## Model 1
## Simulation set=up
set.seed(2016)
m <- 500000
n.obs <- length(qt.m1$body.count)
theta.sim <- rgamma(m, shape = alpha.post.m1, rate = beta.post.m1)
t.sim <- runif(m, min = min(qt.m1$hours), max = max(qt.m1$hours))
yrep <- round(mapply(rpois, n = n.obs, lambda = theta.sim*t.sim))

## Minimum
obs.min <- min(qt.m1$body.count)      # observed minimum
sim.min <- apply(yrep, 2, min)        # simulated minimum
pval.min <- length(sim.min[sim.min >= obs.min]) / m

## Maximum
obs.max <- max(qt.m1$body.count)      # observed maximum
sim.max <- apply(yrep, 2, max)        # simulated maximum
pval.max <- length(sim.max[sim.max <= obs.max]) / m

## Ratio: Sample Variance/ Sample Mean
obs.mean <- mean(qt.m1$body.count)    # observed mean
sim.mean <- apply(yrep, 2, mean)      # simulated mean
obs.var <- var(qt.m1$body.count)      # observed variance
sim.var <- apply(yrep, 2, var)        # simulated variance

obs.ratio <- obs.var / obs.mean
sim.ratio <- sim.var / sim.mean
pval.ratio <- length(sim.ratio[sim.ratio <= obs.ratio]) / m
```

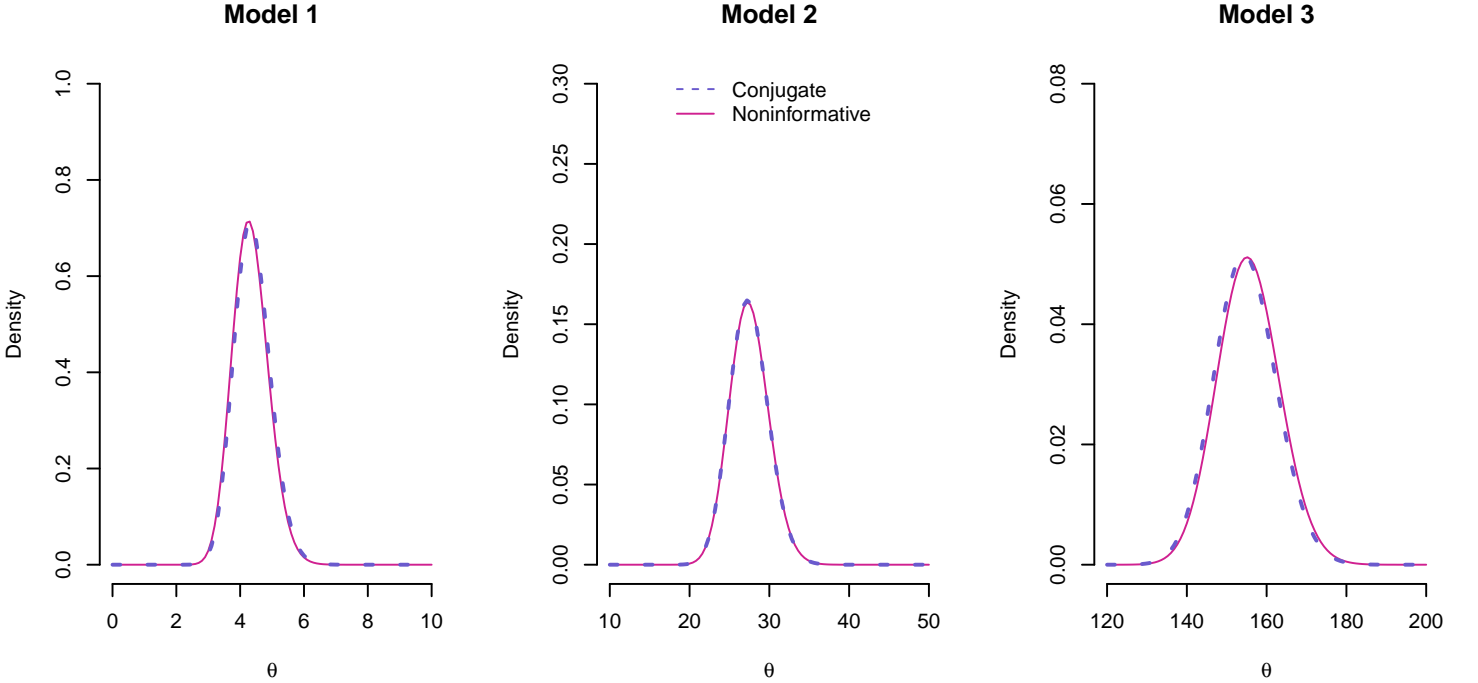
Sensitivity to Prior Distribution

In this section we consider the effects of the choice of prior distribution on posterior inference, particularly the

effects of a noninformative prior instead of the gamma conjugate prior. The Jeffreys prior for θ follows from the Fisher information,

$$\begin{aligned}
I_n(\theta) &= \mathbb{E} \left\{ \left(\frac{\partial}{\partial \theta} \log p(y \mid \theta) \right)^2 \mid \theta \right\} \\
&= \frac{1}{\theta^2} \mathbb{E} \left[\left(\sum_{i=1}^n y_i - \theta \sum_{i=1}^n t_i \right)^2 \mid \theta \right] \\
&= \frac{1}{\theta^2} \text{Var} \left(\sum_{i=1}^n y_i \mid \theta \right) && \text{assuming } y_i \text{ are } c.i.i.d \text{ given } \theta \\
&= \frac{1}{\theta} \cdot \sum_{i=1}^n t_i \\
\implies p(\theta) &\propto \sqrt{\frac{1}{\theta}}
\end{aligned}$$

This results in an improper noninformative prior; however, this does not prevent us from finding a posterior distribution for θ . The resulting posterior distribution for θ is $\text{Gamma}\left(0.5 + \sum_{i=1}^n y_i, \sum_{i=1}^n t_i\right)$. Below are the plots of the posterior distributions of θ as a result of using a conjugate prior and a non-informative prior. In all three models the differences are negligible, so a choosing a non-informative prior instead of a conjugate prior would have little impact on posterior inference.



V. References

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). Bayesian data analysis (Vol. 2). Boca Raton, FL, USA: Chapman & Hall/CRC.

Dr. Lee R code

Olson, Randy (2013): On-screen movie kill counts for hundreds of films. figshare. <https://dx.doi.org/10.6084/m9.figshare.889719.v1> Retrieved: 01 39, Nov 29, 2016 (GMT)

Beggs, A. & Handy, B. (2013, February 13). *Down for the count*. Retrieved from <http://www.vanityfair.com/hollywood/2013/02/quentin-tarantino-deaths-movies>

Robinson, J. (2016, July 12). *Does Quentin Tarantino Really Have Just Two More Movies in Him?* Retrieved from <http://www.vanityfair.com/hollywood/2016/07/quentin-tarantino-ten-movies-retirement>

Poisson distribution. (2016, November 27). In Wikipedia, The Free Encyclopedia. Retrieved 18:33, November 27, 2016, from https://en.wikipedia.org/w/index.php?title=Poisson_distribution&oldid=751763566

Gamma distribution. (2016, November 8). In Wikipedia, The Free Encyclopedia. Retrieved 19:13, November 8, 2016, from https://en.wikipedia.org/w/index.php?title=Gamma_distribution&oldid=748540178

Lee, K. B. (2016). *The Tarantino Death Toll* [Video]. Retrieved from <https://vimeo.com/148832585>