Modeling Movie Violence:

A Statistical Look at Quentin Tarantino Filmography

Joey Fitch

Qiuyu Li

Christopher Rainey

San José State University

Math 264

December 1st, 2016

Contents

1 Introduction and Prior Beliefs

In this paper, we aim to model the kill count of Quentin Tarantino's movies and make predictions about future results. We proceed under the knowledge that Tarantino "has long been criticized for the amount of violence and the body count in his movies." Under this presumption, we investigated other directors who display tendency toward violent movies. We used a forum dedicated to movie enthusiasts[3] to gather a list of directors whom are considered to make violent movies or movies that have high body counts. Using this list, we were able to gather data on 8 other directors: Sam Peckinpah, John Woo, Peter Jackson, George Romero, Alexandra Aja, Jonathan Liebesman, Roland Emmerich, and Tony Scott. In the context of these directors, Tarantino's kill counts do not seem alarming, but Tarantino does display a capability for heavy violence in at least one movie: "Inglourious Basterds", which counted a surprising 396 on-screen deaths. Our goal is to analyze whether this massive kill rate will continue, based on the context of Tarantino's filmography and the filmography of similar directors. Reference Appendix A for more details on this data.

We did some background research to validate these directors as good choices. John Woo is known for his stylized action scenes. George Romero makes zombie movies, and – while zombies may or may not count as people – his movies display a good amount of violence. Roland Emmerich has made several movies with fairly high body counts, tending toward violent topics like alien invasions and the world ending. In all, these directors seemed like reasonable comparisons within our expectations of Quentin Tarantino.

1.1 Choice of Sampling Model

In order to analyze a kill rate in movies, the natural first guess for measuring count data over time is the Poisson distribution. The Poisson distribution is widely used in similar count data problems, but it has limitations: by necessity, the model should display equal mean and variance across the count data. However, after researching the kill counts in similar directors' movies, we saw that the mean and variance of the kill count differed wildly from one another, regardless of the director. The variance was consistently higher than the mean (see Table 1). This indicates a trend across directors, and we fairly assume that the same pattern holds for Tarantino. Below is a table illustrating some of the sampled directors.

Director	**Mean Kill Count**	**Kill Count Variance**
John Woo	123	9692
George Romero	77	7408
Roland Emmerich	269	124858

Kill count mean and variance for various directors.
See Appendix A for more detail.

Given this information, we can infer that a Poisson distribution is not a proper fit, as kill count variance differs drastically from the mean kill count for every choice of director. We proceed by considering a Poisson model with parameter λ, where λ varies across the population according to a Gamma(γ, α) distribution with shape parameter $\gamma > 0$ and scale parameter $\alpha > 0$. We can obtain the sampling distribution for an observation x (kill count of a movie) by marginalizing over λ:

$$p(x|\gamma,\alpha) = \int_0^\infty p(x|\lambda)\cdot p(\lambda|\gamma,\alpha)\,d\lambda = \int_0^\infty \frac{\alpha^\gamma}{\Gamma(\gamma)}\lambda^{\alpha-1}e^{-\alpha\lambda}\cdot\frac{e^{-\lambda}\lambda^x}{x!}\;d\lambda = \frac{\Gamma(\gamma+x)}{\Gamma(\gamma)x!}\;\alpha^\gamma(1+\alpha)^{\gamma+x}$$

Thereby, the marginal distribution for the kill count, x, of a movie gives way to the Negative Binomial sampling model, with parameters γ and α resulting from the over-dispersion of variance. We can look at the model as a generalized Poisson distribution, in that the Negative Binomial offers the same support and the same interpretation of the mean as the Poisson distribution but with a scaled variance. In fact, if we reparameterize according to $p = \alpha/(\alpha+1)$, the marginal PMF can be written:

$$p(x|\gamma,p) = \frac{\Gamma(\gamma+x)}{\Gamma(\gamma)x!}(p)^{\gamma}(1-p)^{x}$$

with $Mean(x) = \frac{\gamma(1-p)}{p}$ and $Var(x) = \frac{\mu}{p}$, where $\gamma > 0$ and $0 < p < 1$. As shown in this formula for variance, the parameter p can be interpreted as the inflation of variance, relative to the mean.

1.2 Prior Beliefs About Movie Kill Counts

Using our prior knowledge of directors similar to Quentin Tarantino, we expect high degrees of belief for small values of p, such that the variance will be much larger than the mean. If p is indeed small, than the quantity $\frac{(1-p)}{p}$ will be large, and thus we also expect small values of γ such that μ is kept to a reasonable value. We choose hyperprior distributions to express these beliefs and match the support space of each parameter:

$$p \sim \text{Beta}(\alpha,\beta) \qquad\qquad \gamma \sim \text{Gamma}(\delta_1,\delta_2)$$

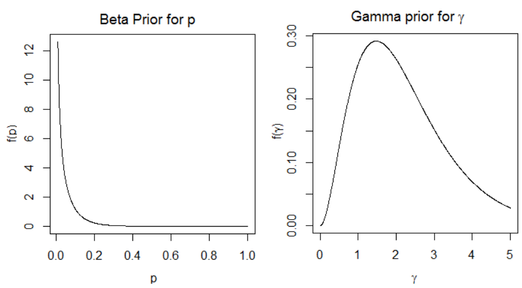
We capture these beliefs by using the prior data (gathered from other directors) to calculate estimates for the hyperparameters. Specifically, we assume that the kill count for each director follows a Negative Binomial distribution. We then posit that the parameters for each of the distributions are identically distributed, according to $p_d \sim \text{Beta}(\alpha,\beta)$ and $\gamma_d \sim \text{Gamma}(\delta_1,\delta_2)$. Given that time order and labeling are unimportant in this context, it is reasonable to assume exchangeability between the parameters. With this in mind, we estimate p for each director's data set X_d (except Tarantino) by setting $\widehat{p}_d = \frac{Mean(X_d)}{Var(X_d)}$, and then calculating $\widehat{\gamma}_d = \frac{\widehat{p}_d}{(1-\widehat{p}_d)}Mean(X_d)$. Therefore, each of the 9 sampled directors provide an estimate of p and γ, giving us 8 estimates for each.

With our set of 8 p estimates and 8 γ estimates, we can approximate the mean and variance for each parameter, which together lead to estimates of the hyperparameters:

Distr. Mean	Distr. Variance	Sample Measures	Hyperparameter Estimates
$\text{Mean}(p) = \frac{\alpha}{\alpha+\beta}$	$\text{Var}(p) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	$\bar{p} = 0.035, S_p^2 = 0.0025$	$\alpha = 0.43, \beta = 11.91$
$\text{Mean}(\gamma) = \frac{\delta_2}{\delta_1}$	$\text{Var}(\gamma) = \frac{\delta_2}{\delta_1^2}$	$\bar{r} = 1.47, S_r^2 = 1.08$	$\delta_1 = 1.36, \delta_2 = 2.01$

These hyperparameter estimates do have measurable consequences on our model. For example, note that the Effective Sample Size of a Beta prior on p is equal to $\alpha+\beta = 12.34$. Given that our actual data set contains only 8 observations, this allows the prior belief to have a significant effect on the posterior model. We accept this result, as we consider the prior knowledge credible and we acknowledge the small size of our own data set.

Acknowledging these concerns, we construct the following prior distributions for p and γ:



Notice how our prior beliefs about p and γ are captured by these distributions:

· Our degree of belief is highest for small values of p, increasing monotonically as $p \to 0$.

· Our degree of belief in γ is centered around small values; specifically, our highest belief lies at $\gamma = 1.47$.

2 Quentin Tarantino Data & Posterior Beliefs

We are given a data set generated by Vanity Fair magazine, in which they list all recorded kills in Quentin Tarantino's first 7 movies, including the method of death. The method of death is considered irrelevant for our purposes, so we ignore it in our model. While it is hard to count total body count in movies like Inglourious Basterds, due to a lot of chaos in some scenes, we assumed that the Vanity Fair dataset was as correct as humanly possible. Moreover, despite the outlier appearance of Inglourious Basterds relative to Tarantino's other films, we purposefully do *not* disregard this data. We have no reason to consider this behavior abnormal, even if relatively uncommon. We adapt this large variation into our model, in the hope that our model can account for such an outlier and describe Tarantino's habits accordingly.

We frame our model by using x_i (i=1,2,3, ...) to describe the total number of kills that have occurred in the i^{th} Quentin Tarantino movie, where x_i can take on values {0,1,2, ...}. The x_i are considered exchangeable; we have no reason to believe that Tarantino will drastically change his tendencies as a director between films over time. Reordering the movies in the sequence of filming would not have any known effect on their joint distribution. Essentially, our ignorance of any further nuances to Tarantino's filmography allows us to assume exchangeability between movies.

We cannot assume the movie kill counts are independent, because information about one movie would affect our belief about other movies' kill counts. E.g., if I know that Quentin Tarantino's first movie has 300 kills, I would allow more belief to the possibility that his second movie also has a high kill count. However, if we condition on the parameters p and γ, then knowledge of the first movie would not effect our belief about the second movie kill count because the parameters would already be "known" in this case. Our degree of belief about the second movie would rely solely on that knowledge of the parameters, uninfluenced by the first observation, and hence these movies must be conditionally independent. We therefore justify conditional independence of our data, given parameters p and γ. In the next section, this allows us to construct our likelihood as the product of individual PDF's.

2.1 Constructing the Posterior

Using the Negative Binomial sampling model, we construct the posterior distribution by multiplying the Likelihood and the joint prior:

$$\begin{aligned} p(\gamma, p|x_i, \alpha, \beta, \delta_1, \delta_2) &\propto \frac{\prod_{i=1}^{8} \Gamma(\gamma + x_i)}{\Gamma(\gamma)^8} p^{8\gamma}(1-p)^{\sum_{i=1}^{8} x_i} \cdot \gamma^{\delta_2 - 1} e^{-\delta_1 \gamma} p^{\alpha-1}(1-p)^{\beta-1} \\ &\propto \frac{\prod_{i=1}^{8} \Gamma(\gamma + x_i)}{\Gamma(\gamma)^8} \gamma^{\delta_2-1} e^{-\delta_1 r} p^{8\gamma+\alpha-1}(1-p)^{\sum_{i=1}^{8} x_i + \beta - 1} \end{aligned}$$

For an analysis of this posterior distribution, see Appendix B. However, this distribution is not of immediate interest; rather, we proceed to postulate about the count data of movie kill rates.

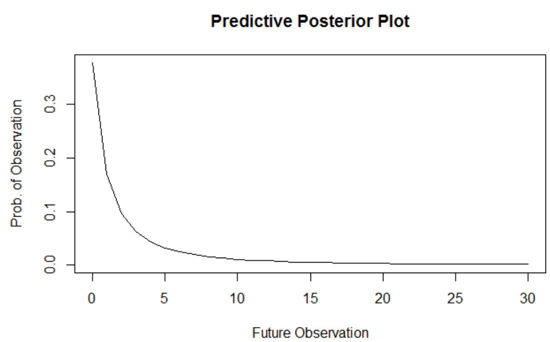
3 Posterior Predictive Distribution

To construct the Posterior Predictive Distribution for a future observation $\widetilde{y}$, we integrate the sampling model over our posterior probabilities for γ and p:

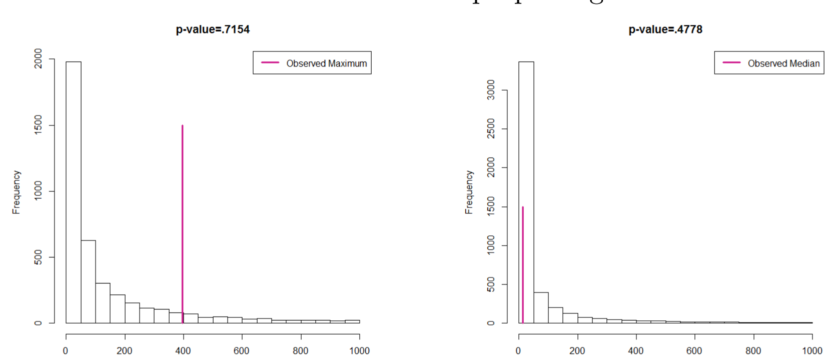
$$p(\widetilde{y}|x) \propto \int_\gamma \int_p \gamma^{\delta_1 - 1} e^{-\delta_1 \gamma} p^{\widetilde{y}} (1-p)^\gamma \frac{\Gamma(\widetilde{y}+\gamma)\Gamma(8\gamma+\alpha) \prod_{i=1}^{8} \Gamma(\gamma+x_i)}{\Gamma(\gamma)^9 \Gamma(8\gamma + \sum_{i=1}^{8} x_i + \alpha + \beta)} dp\, d\gamma$$

$$\propto \int_\gamma \gamma^{\delta_1 - 1} e^{-\delta_1 \gamma} \frac{\Gamma(\widetilde{y}+\gamma)\Gamma(8\gamma+\alpha) \prod_{i=1}^{8} \Gamma(\gamma+x_i)}{\Gamma(\gamma)^9 \Gamma(8\gamma + \sum_{i=1}^{8} x_i + \alpha + \beta)} d\gamma$$

This integral has no closed form, but can be evaluated over all values of $\widetilde{y}$ as a discrete pmf over $\mathbb{N} = \mathbb{Z}^+ \cup \{0\}$. This figure shows a linear interpolation of the PMF $p(\widetilde{y}|x)$:



We can check the validity of our model by conducting various "p-value" tests of the Predictive distribution against Test Quantities of our data. Specifically, we analyze the maximum value of hypothetical data sets (sampled under our Posterior Predictive distribution), in order to determine whether the outrageously high kill count of "Inglourious Basterds" can be accounted for by our model. If our model does not predict such outliers, then it is likely not a good fit for the data. We also compare the median value of hypothetical data sets with our given median of the Tarantino data set, to see if our distribution is centered in the proper region of count data.



These p-values are suitably moderate, showing that our data falls perfectly in line with the expectations of our model. It is reasonable to believe that our distribution accurately models the given kill rate count data.

4 Conclusions

4.1 Summary of Predictions

From the plot of $p(\widetilde{y}|x)$, it is clear that the singularly most probable value of $\widetilde{y}$ is simply $\widetilde{y} = 0$. While this is true mathematically, it fails to describe the nuance of the predictive distribution. The tail of the distribution extends through all values $\widetilde{y} > 0$, and –in context– $x = 0$ is not actually a very probable observation, since no director we sampled had any observation lower than 2 and very few were even under 10.

To address this shortcoming, we consider the expected value of $\widetilde{y}$ instead. As a weighted probability, it takes the full support space into account and is better suited to capture the scope of the model. We calculate the infinite sum $\sum_{i=0}^{\infty} \widetilde{y} \cdot p(\widetilde{y}|y)$ and find that the Posterior Predictive Distribution has expected value equal to about 14. We consider this value to be our best single-value prediction.

In a similar motivation, we can calculate the HPD and see that 95% of our predictive belief is concentrated at $\widetilde{y}$ values between 1 and 31. So, while it appears that our model over-estimates the probability of low kill counts, it seems to allow reasonable weight to the higher values as well, and overall produces meaningful results.

Overall, after careful evaluation of our research and our statistical model, we present evidence that Quentin Tarantino is not alarmingly violent, at least relative to other directors within his realm of entertainment. We have strong belief (95%) that Tarantino's next film will not have a kill count exceeding 31 deaths, and will likely be in the vicinity of about 14 deaths overall.

4.2 Comparison to Tarantino's "Hateful Eight"

As it turns out, Quentin Tarantino has released a new film not included in our data set. In his movie "*Hateful Eight*", 16 on-screen kills are recorded. This is almost perfectly in line with our expectations, and serves to further validate the accuracy of our model.

5 Appendix A: Data

5.1 Prior Data

Director	Movie	Kill Count	Mean	Variance	$\widehat{p} = \frac{Mean}{Var}$	$\widehat{\gamma} = \frac{Mean \cdot \widehat{p}}{(1-\widehat{p})}$
Sam Peckinpah	The Wild Bunch (1969)	145	116.33	9832.33	0.0118	1.3929
	Straw Dogs (1971)	6				
	Major Dundee (1965)	198				
John Woo	Hard Boiled (1992)	307	123.44	9692.03	0.0127	1.5925
	Bullet in the Head (1990)	214				
	A Better Tomorrow II (1987)	199				
	The Killer (1989)	149				
	A Better Tomorrow (1986)	65				
	Once a Thief (1991)	59				
	Face/Off (1997)	56				
	Mission: Impossible II (2000)	37				
	Broken Arrow (1996)	25				
Peter Jackson	LotR: Return of the King (2003)	836	361.25	136788.92	0.00264	0.9565
	LotR: The Two Towers (2002)	468				
	LotR: Fellowship of the Ring (2001)	118				
	King Kong (2005)	23				
George Romero	Dawn of the Dead (1978)	175	76.67	7408.33	0.01034	0.8017
	Day of the Dead (1985)	40				
	Creepshow(1982)	15				
Alexandre Aja	The Hills Have Eyes (2006)	18	12.25	122.92	0.0996	1.3556
	Piranha 3D(2010)	25				
	P2(2007)	2				
	mirror(2008)	4				
Jonathan Liebesman	Wrath Of The Titans (2012)	283	147.67	18497.33	0.007983	1.1883
	Clash Of The Titans (2010)	149				
	The Texas Chainsaw Massacre: The Beginning (2006)	11				

Director	**Movie**	**Kill Count**	**Mean**	**Variance**	$\widehat{p} = \frac{Mean}{Var}$	$\widehat{\gamma} = \frac{Mean \cdot \widehat{p}}{(1-\widehat{p})}$
Roland Emmerich	The Patriot (2000)	123	268.60	124858	0.002151	0.579064
	Universal Soldier (1992)	47				
	The Day After Tomorrow (2004)	130				
	Godzilla (1998)	146				
	Independence Day (1996)	897				
Tony Scott	Domino (2005)	51	26.40	204.80	0.1289	3.9067
	The Last Boy Scout (1991)	26				
	True Romance (1993)	21				
	Man on Fire (2004)	19				
	Enemy of the State (1998)	15				

5.2 Quentin Tarantino Data

Director	**Movie**	**Kill Count**	**Mean**	**Variance**	$\widehat{p} = \frac{Mean}{Var}$	$\widehat{\gamma} = \frac{Mean \cdot \widehat{p}}{(1-\widehat{p})}$
Quentin Tarantino	Reservoir Dogs (1992)	11	70.375	17932	0.00392	0.277
	Pulp Fiction (1994)	7				
	Jackie Brown (1997)	4				
	Kill Bill Vol. 1 (2003)	62				
	Kill Bill Vol. 2 (2004)	13				
	Death Proof (2007)	6				
	Inglourious Basterds (2009)	396				
	Django Unchained (20012)	64				

6 Appendix B: Analysis of Posterior Parameters γ and p

After analysis of our prior beliefs and given data, we constructed the posterior distribution for the parameters of our Negative Binomial sampling model:

$$p(\gamma, p|x_i, \alpha, \beta, \delta_1, \delta_2) \quad \propto \quad \frac{\prod_{i=1}^{8} \Gamma(\gamma+x_i)}{\Gamma(\gamma)^8} \gamma^{\delta_2-1} e^{-\delta_1 r} p^{8\gamma+\alpha-1} (1-p)^{\sum_{i=1}^{8} x_i+\beta-1}$$

6.1 Posterior Analysis of p

If we take the derivative of our posterior, with respect to p:

$$\frac{d}{dp} p(\gamma, p|x_i, \alpha, \beta, \delta_1, \delta_2) \quad \propto \quad \frac{\prod_{i=1}^{8} \Gamma(\gamma+x_i)}{\Gamma(\gamma)^8} \gamma^{\delta_2-1} e^{-\delta_1 r} \cdot p^{8\gamma+\alpha-1} (1-p)^{\sum_{i=1}^{8} x_i+\beta-1} \log(p)$$

Notice that all terms are strictly positive for $\gamma > 0$ and $p \in (0,1)$, with the exception of $\log(p)$ which is strictly negative on $p \in (0,1)$. Therefore, this expression is strictly negative for any value of p or γ. In other words – the posterior distribution is monotonically decreasing as p increases from 0 to 1. This means that we *always* assign stronger beliefs to smaller values of p, even as it approaches 0, without any singular mode.

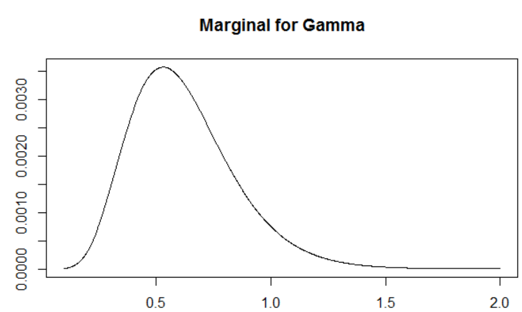
6.2 Posterior Analysis of γ

Our results for p are not particularly interesting, or even unexpected. The movie kill count data naturally tends toward tiny values of p, as you can see from the multiple director data sets. This leaves γ as the real parameter of interest, so we marginalize over p to produce the marginal posterior density function $p(\gamma, |p, x_i, \alpha, \beta, \delta_1, \delta_2)$.

$$p(\gamma|p, x_i, \alpha, \beta, \delta_1, \delta_2)$$

$$= \int_0^1 p(\gamma, p|x_i, \alpha, \beta, \delta_1, \delta_2) dp$$

$$\propto \quad r^{\delta_2-1} e^{-\delta_1 r} \frac{\prod_{i=1}^{8} \Gamma(\gamma+x_i) \Gamma(8\gamma+\alpha)}{\Gamma(\gamma)^8 \, \Gamma(8\gamma+\alpha+\beta+\sum_{i=1}^{8} x_i)}$$



Using this distribution, we find the posterior mode of γ to be 0.54.

7 References

1. Lio, Y L. 2009. “A Note on Bayesian Estimation for the Negative-Binomial Model.” Pliska -Studia Mathematica Bulgarica 19 (2009), 207-216

2. Cook, John D. 2009. “Notes on the Negative Binomial Distribution.” http://www.johndcook.com/negative binomial.pdf

3. Listal.com. 2008. “Your Top 10 Most Violent Directors of All Time.” [Forum] Available at: http://www.listal.com/topic/your-top-10-most-violent.