MATH 264: Bayesian Project

Terry Situ, Caie Yan, Ryan Quigley

I. Executive Summary

II. Data Summary

Metadata extracted from IMDB

Movie	Genre	Year	Rating	Body Count	Hours	Kill Rate/Hr
Reservoir Dogs	crime, drama, thriller	92	R	11	1.65	6.67
Pulp Fiction	crime, drama	94	R	7	2.567	2.73
Jackie Brown	crime, thriller	97	R	4	2.567	1.56
Kill Bill Vol. 1	action, thriller	03	R	62	1.85	33.51
Kill Bill Vol. 2	action, crime drama, thriller	04	R	13	2.283	5.69
Death Proof	thriller	07	NR	6	1.883	3.19
Inglorious Bastards	adventure, drama, war	09	R	396	2.55	155.29
Django Unchained	drama, western	12	R	64	2.75	23.27
The Hateful Eight	crime, drama, mystery thriller, western	15	R	18	2.783	6.47

III. Models & Distributions

> "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

> "In practice, ignorance implies exchange- ability. Generally, the less we know about a problem, the more confidently we can make claims of exchangeability. (This is not, we hasten to add, a good reason to limit our knowl- edge of a problem before embarking on statistical analysis!)" - Gelman et. al BDA

Model summary:

				Posterior	99% HPD
Model	Prior	Likelihood	Posterior	Mode	Inteval
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]

Model	Prior	Likelihood	Posterior	Posterior Mode	99% HPD Inteval
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 with rate and exposure

According to Gelman et. al. (pg. 45), this model is NOT exchangeable in the y_i's but is exchangeable in the pairs $(x, y)_i$

Assumptions and Justification

Let $y(t)$ denote the number of events that have occurred during a time interval $[0, t]$

- 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 (want to relax this and replace with exchangeability)
- P4: $\lim_{h \to 0} \frac{\Pr(y(h)>1)}{h} = 0$
- P5: $0 < \Pr\{y(t) = 0\} < 1, \ \forall \ t > 0$

Under these conditions, there exists a positive number θ that produces the density below where $\theta =$ the true underlying kill rate per hour in Quentin Taratino movies:

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

Given the plot structure of movies we would not expect the kill rate to be constant within the movie; however, with the information available we cannot determine whether or not this assumption is violated within each movie.

Likelihood

A necessary and sufficient condition to get product of identical distributions is

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

For every n, where $s_n = y_1 + \ldots + y_n$. 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. Assuming this condition holds, the likelihood is:

$$p(y \mid \theta) \propto \theta^{\left(\sum_{i=1}^{n} y_i\right)} \exp\left(-\theta \sum_{i=1}^{n} t_i\right)$$

Conjugate Prior: Gamma

Since the conjugate prior distribution for the Poisson sampling distribution is the Gamma distribution, the prior distribution of θ will be of the form:

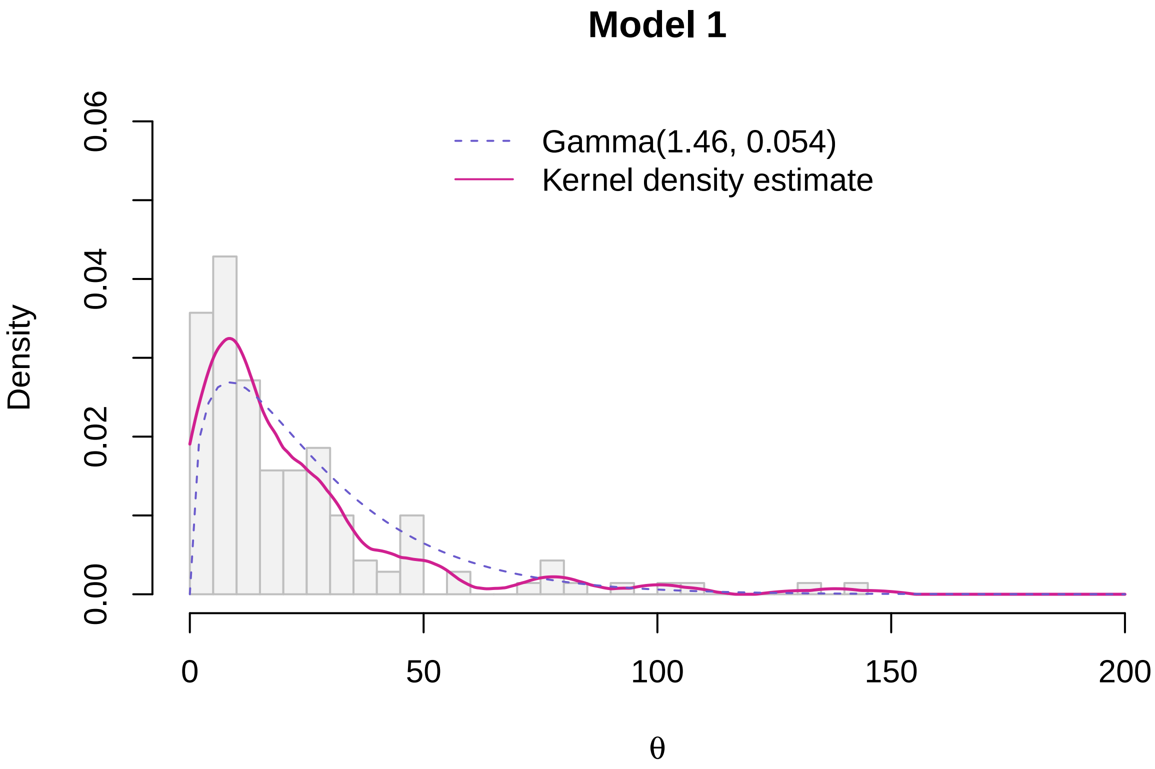
$$p(\theta) \propto \theta^{\alpha-1} e^{-\beta \cdot \theta}$$

Historical Movie Database Sample

In order to estimate the parameters of the conjugate prior distribution for each model, we acquired a dataset compiled by Randal Olson from the site http://www.moviebodycounts.com/. See references for links to original dataset. The original data is filtered to make it more relevant to our analysis. 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. One final note: the original historical movie dataset does not contain any movies with zero deaths. For our purposed, 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:

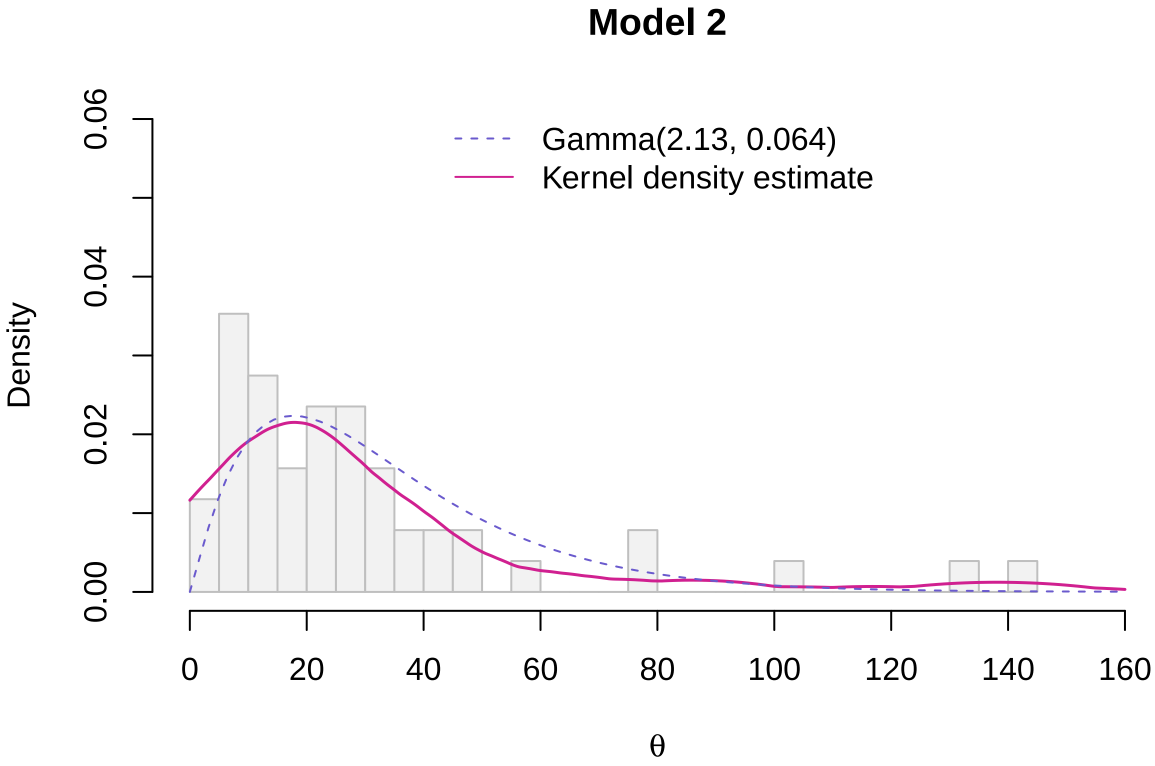
Model	N	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	140	0.49	6.28	14.11	22.65	27.77	143.9
2	51	1.52	10.81	21.94	29.59	34.00	143.9
3	19	9.80	43.50	58.94	106.80	167.60	307.7

Model 1:



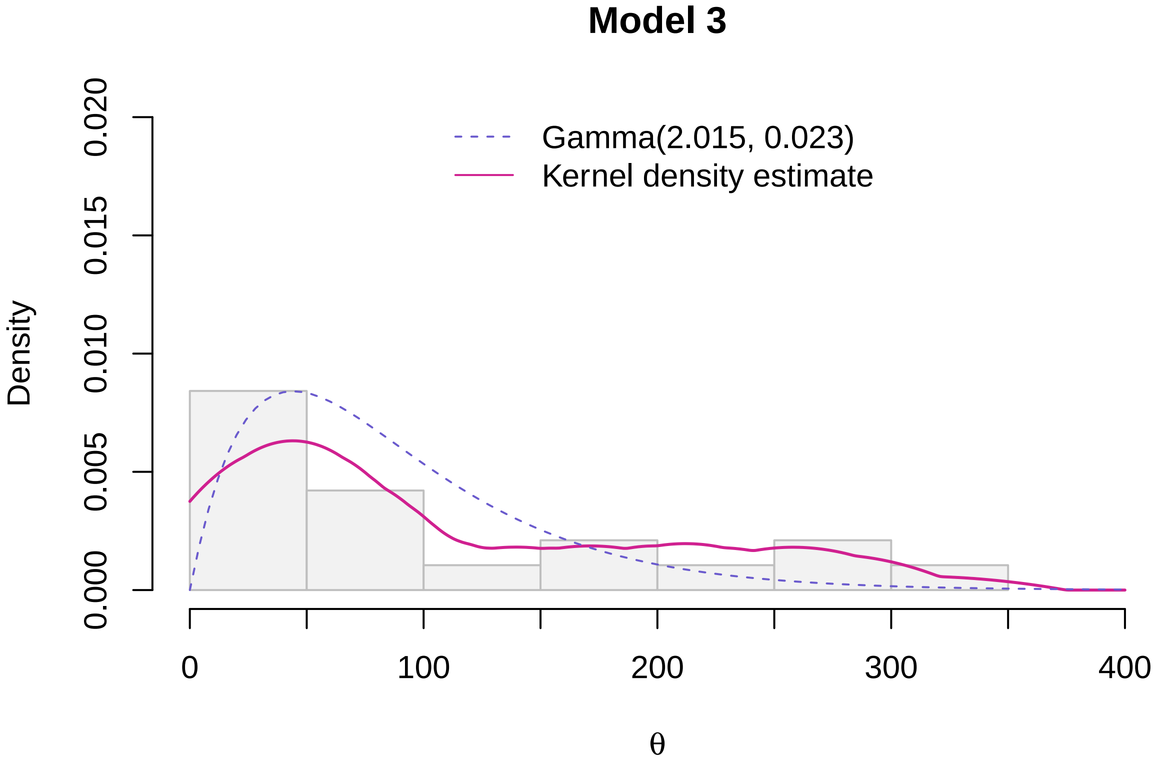
The parameters for the Gamma distribution were determined by solving the system of equations generated by setting the mode equal to 8.61 and by observing that approximately 99.9% of the data is in the range (0,150).

Model 2:



The parameters for the Gamma distribution were determined by solving the system of equations generated by setting the mode equal to 17.85 and by observing that approximately 99.9% of the data is in the range (0,150).

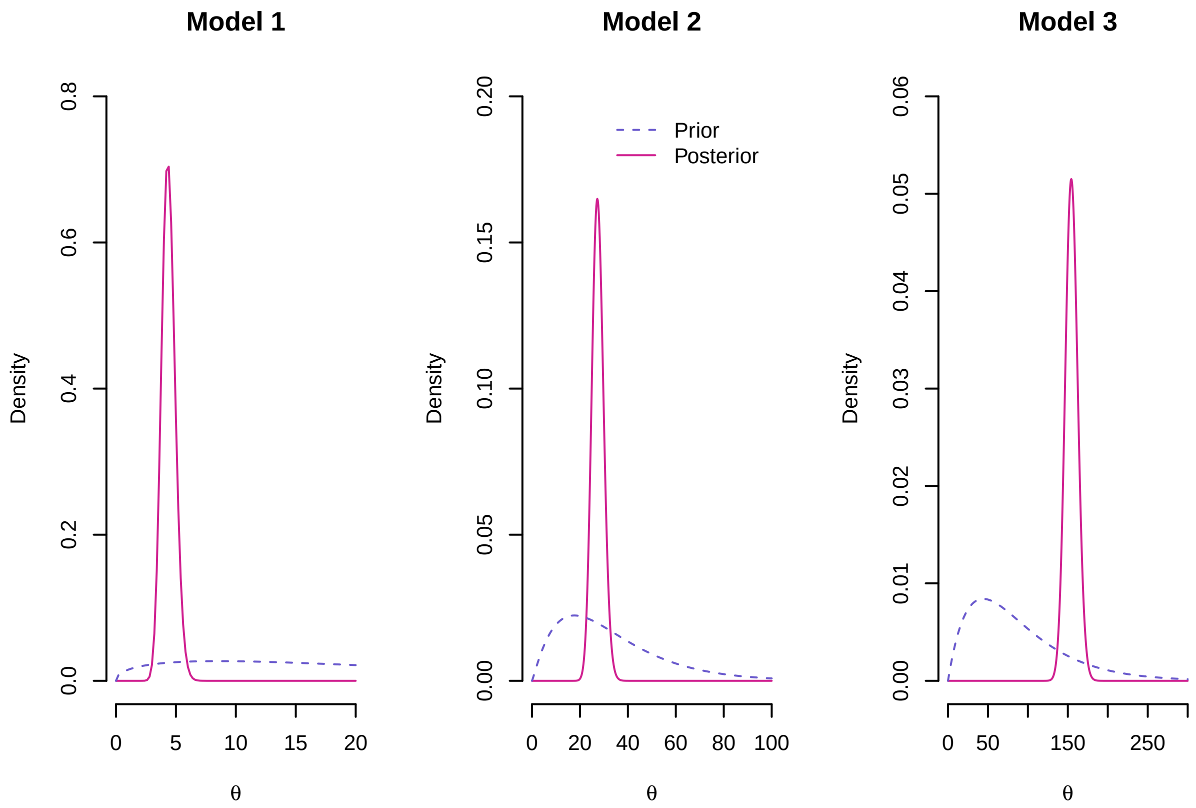
Model 3:



The parameters for the Gamma distribution were determined by solving the system of equations generated by setting the mode equal to 43.83 and by observing that approximately 99.9% of the data is in the range (0,400).

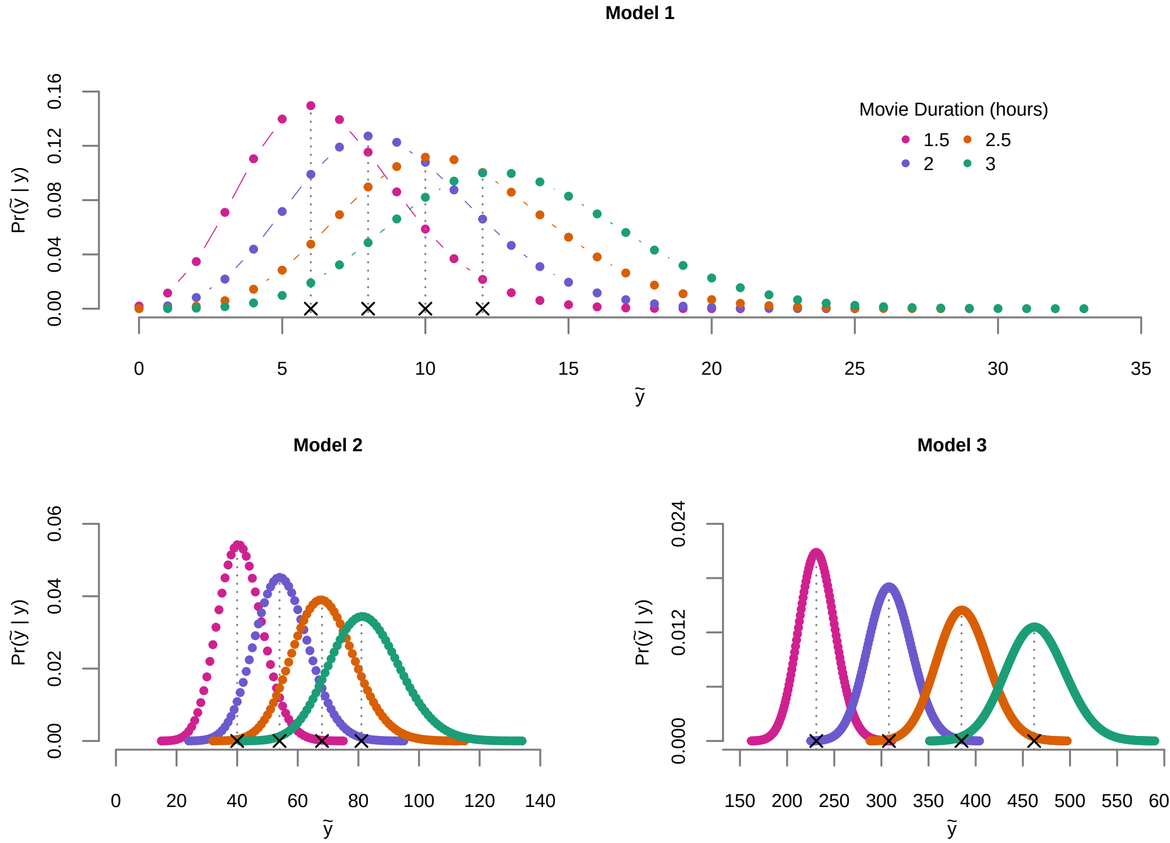
Posterior Distribution

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, pg. 45)



Posterior Predictive Distribution

The posterior predictive distribution for a single additional observation is a negative binomial distribution of the form NB $\left(\alpha + \sum_{i=1}^{n} y_i \ , \ \frac{1}{\tilde{t}}\left[\beta + \sum_{i=1}^{n} t_i\right]\right)$ (Gelman et.al., pg. 44-45). See appendix for derivation.



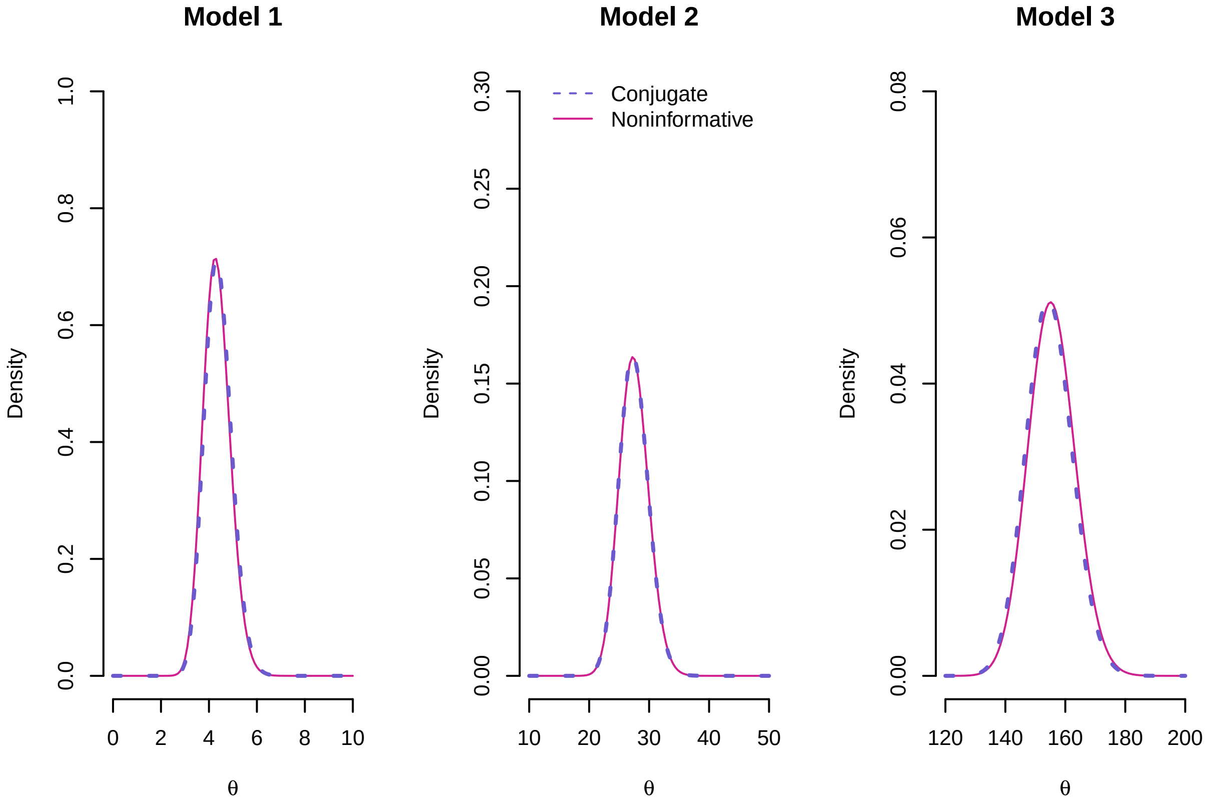
Thus, the most likely values for the body count in Quentin Taratino's next movie are:

Model	Movie Duration (hours)	Body Count	$\Pr(\tilde{y} \mid y)$
1	1.5	6	0.15
	2	8	0.127
	2.5	10	0.112
	3	12	0.1
2	1.5	6	0.15
	2	8	0.127
	2.5	10	0.112
	3	12	0.1
3	1.5	6	0.15
	2	8	0.127
	2.5	10	0.112
	3	12	0.1

IV. Model Checking

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, which results in an improper noninformative prior: $p(\theta) \propto \sqrt{\frac{1}{\theta}}$; however, this does not prevent us from finding a posterior distribution for θ [**WHY???**]. 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 neglible, so a choosing a non-informative prior instead of a conjugate prior would have little impact on posterior inference.

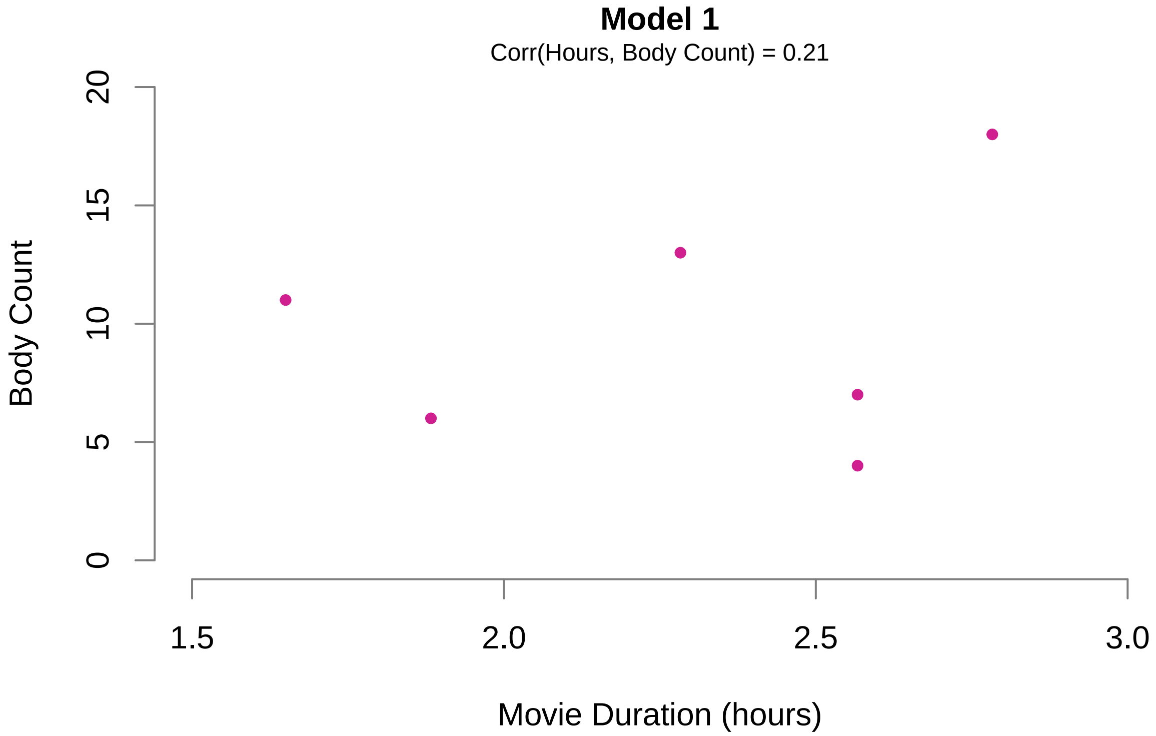


Posterior Predictive Checking

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:

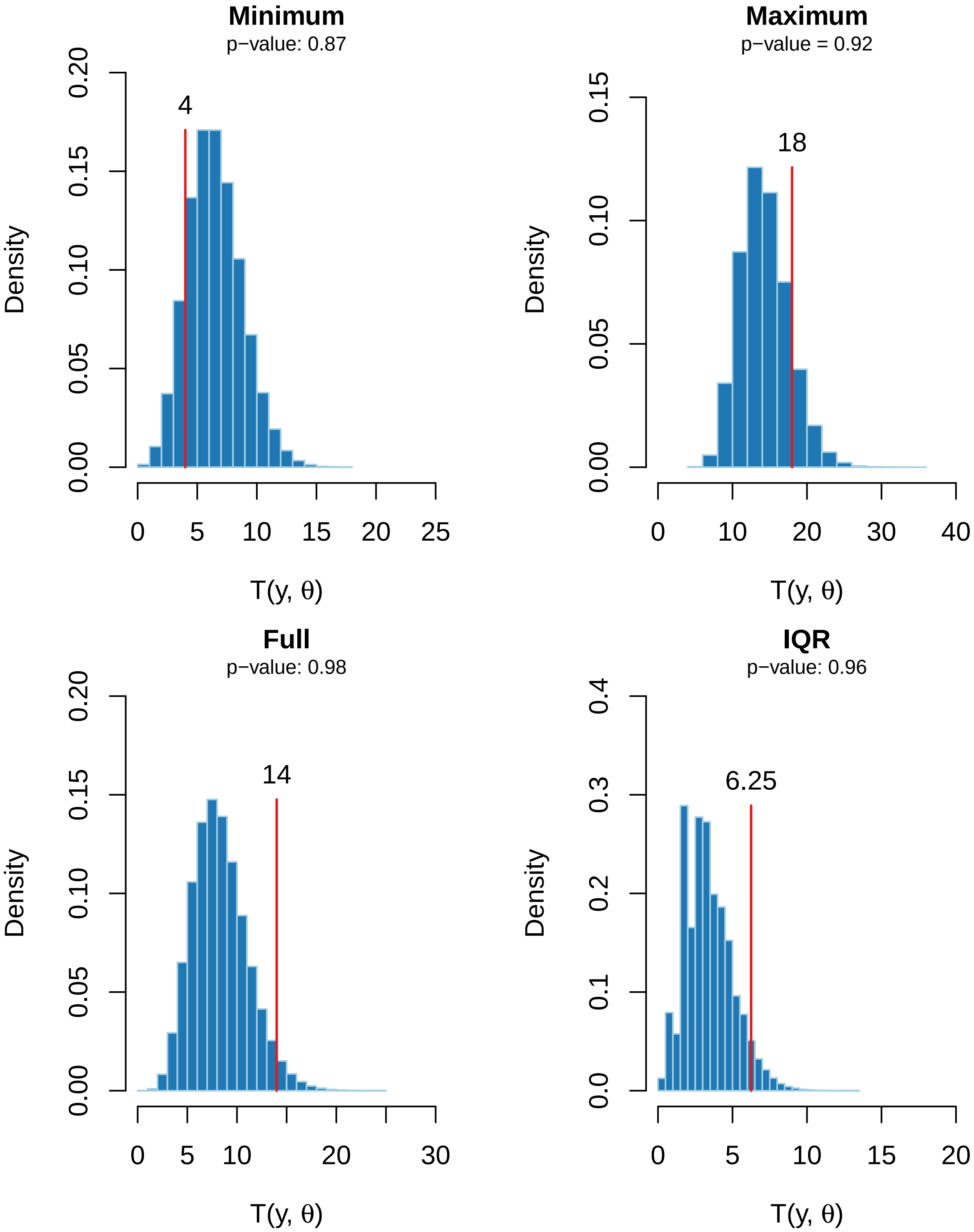
1. Draw $\theta^{(i)}$ from Gamma$\left(\alpha + \sum_{i=1}^{n} y_i, \beta + \sum_{i=1}^{n} t_i\right)$
2. Multiply $\theta^{(i)}$ by t^*
3. For each $(\theta^{(i)} \cdot t^*)$, simulate a draw $y^{rep(i)}(t^*)$ from Poisson$(\theta^{(i)} \cdot t^*)$
4. Calculate $T(y^{rep(i)}(t^*), \theta^{(i)})$ for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

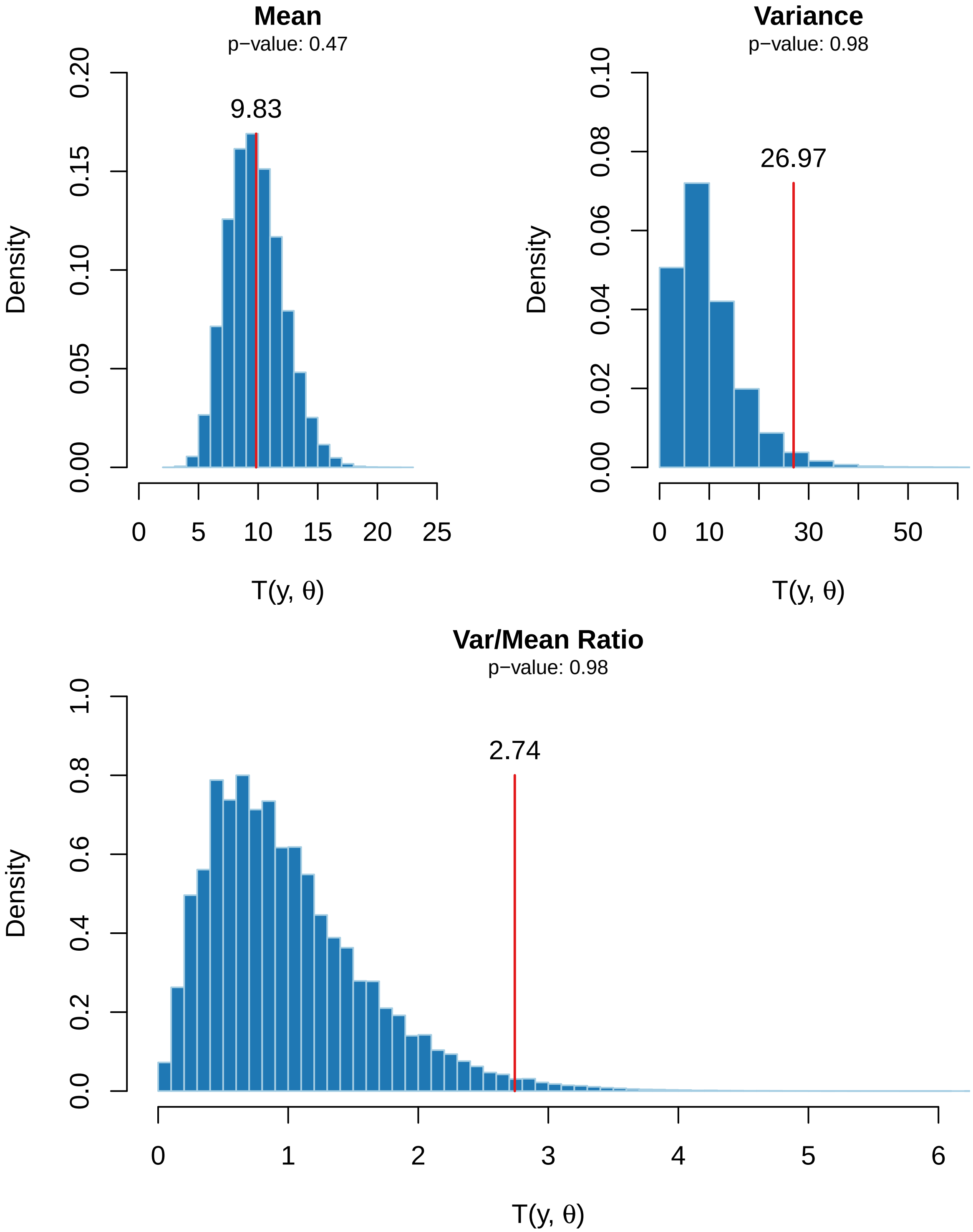
For model 1 mostly, this presents a problem when deciding what value of t^* to choose for each simulation. To address this issue, we note that there does not appear to be a strong relationship between body count and movie duration. This can be seen in the plot below, and is supported by a low correlation value of 0.21. As a result, we draw t^* randomly from $U(min\{t_1, ...t_n\}, max\{t_1, ...t_n\})$ in step 2 of the simulation procedure. In the case of model 1, this is the interval [1.65, 2.78]; for model 2, this is the interval [1.85, 2.75]. For model 3, we simply take $t^* = 2.55$ since the model only has a single data point.

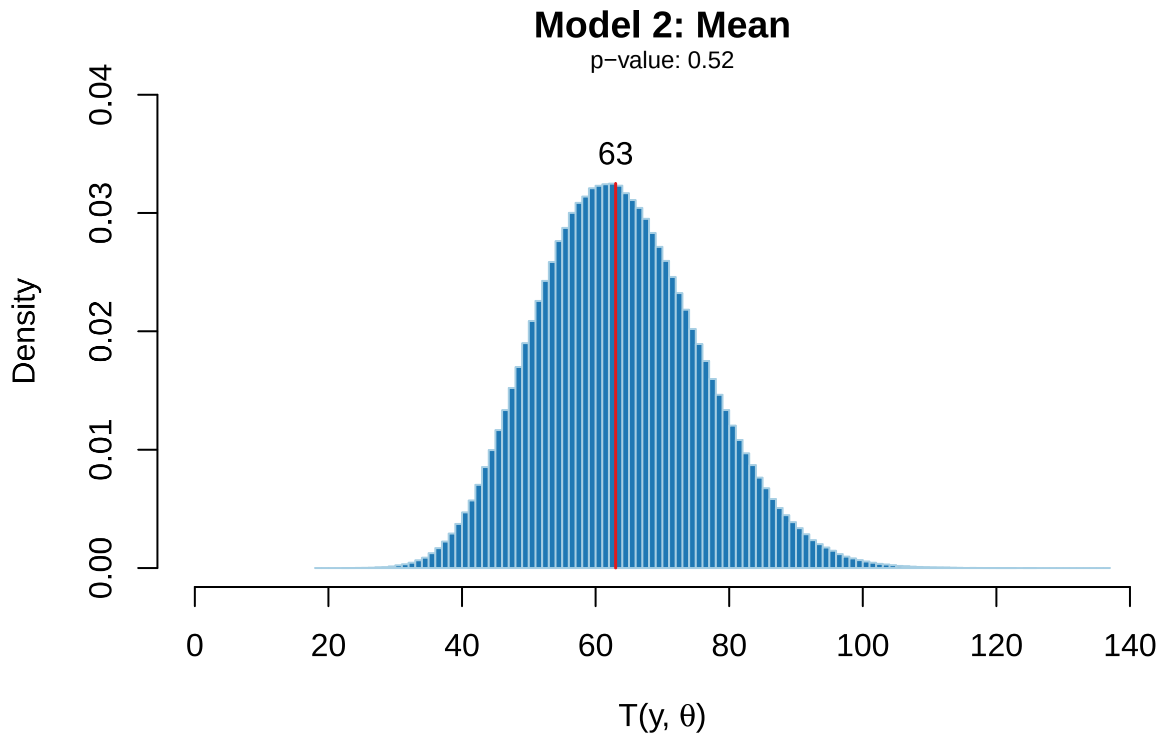


The following table summarizes the posterior predictive p-values from the seven test quantities calculated for model 1. For each, 500,000 replications were calculated. Column 3 indicates whether the p-value was calculated using the proportion of simulated values of $T(y^{rep}(t^*), \theta)$ greater than or equal to the observed value $T(y, \theta)$ or the proportion of simulated values less than or equal to the observed value. See the appendix for more details.

Test Quantity	P-Value	Calculation Method
Minimum	0.87	>=
Maximum	0.92	<=
Range	0.98	<=
Interquartile Range	0.96	<=
Mean	0.47	>=
Variance	0.98	<=
Var/Mean Ratio	0.98	<=


Minimum
p−value: 0.87
4
Density
0.00
0.05
0.10
0.15
0.20
0
5
10
15
20
25
T(y, θ)
Maximum
p−value = 0.92
18
Density
0.00
0.05
0.10
0.15
0
10
20
30
40
T(y, θ)
Full
p−value: 0.98
14
Density
0.00
0.05
0.10
0.15
0.20
0
5
10
20
30
T(y, θ)
IQR
p−value: 0.96
6.25
Density
0.0
0.1
0.2
0.3
0.4
0
5
10
15
20
T(y, θ)


Mean
p−value: 0.47
9.83
Density
0.00
0.05
0.10
0.15
0.20
0
5
10
15
20
25
T(y, θ)
Variance
p−value: 0.98
26.97
Density
0.00
0.02
0.04
0.06
0.08
0.10
0
10
30
50
T(y, θ)
Var/Mean Ratio
p−value: 0.98
2.74
Density
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
6
T(y, θ)



Alternative Models

- Hierachichal model
- Negative binomial for model 1 in order to handle overdispersion
- Poisson regression

V. Appendix

Model Derivations

Negative binomial posterior preditive distribution: [ADD STEPS!!!]

$$
\begin{aligned}
p(\tilde{y} \mid y) &= \int_0^\infty p(\tilde{y} \mid \theta) \cdot p(\theta \mid y) \; d\theta \\
&= \int_0^\infty \text{Poisson}(\tilde{y}(t) \mid \theta) \cdot \text{Gamma}\left(\alpha + \sum_{i=1}^n y_i, \beta + \sum_{i=1}^n t_i\right) \; d\theta \\
&= \text{NB}\left(\alpha + \sum_{i=1}^n y_i \; , \; \frac{1}{\tilde{t}}\left[\beta + \sum_{i=1}^n t_i\right]\right)
\end{aligned}
$$

Noninformative (Jeffreys) Prior:

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

Code: Posterior Predictive Checking

```
## 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

## Range: Full
obs.range <- diff(range(qt.m1$body.count))  # observed range
sim.itmd <- apply(yrep, 2, range)       # simulated range
sim.range <- apply(sim.itmd, 2, diff)
pval.range <- length(sim.range[sim.range <= obs.range]) / m

## Range: Interquartile
obs.IQRange <- IQR(qt.m1$body.count)
sim.IQRange <- apply(yrep, 2, IQR)
pval.IQR <- length(sim.IQRange[sim.IQRange <= obs.IQRange]) / m

## Mean
```

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

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

## Ratio: Sample Variance/ Sample Mean
obs.ratio <- obs.var / obs.mean
sim.ratio <- sim.var / sim.mean
pval.ratio <- length(sim.ratio[sim.ratio <= obs.ratio]) / m


## Model 2
## Simulation set-up
m <- 500000
n.obs.m2 <- length(qt.m2$body.count)
theta.sim.m2 <- rgamma(m, shape = alpha.post.m2, rate = beta.post.m2)
t.sim.m2 <- runif(m, min = min(qt.m2$hours), max = max(qt.m2$hours))
yrep.m2 <- round(mapply(rpois, n = n.obs, lambda = theta.sim.m2*t.sim.m2))
## Mean
obs.mean.m2 <- mean(qt.m2$body.count)
sim.mean.m2 <- apply(yrep.m2, 2, mean)
pval.mean.m2 <- length(sim.mean.m2[sim.mean.m2 <= obs.mean.m2]) / m
```

VI. References

IMDB. http://www.imdb.com/

http://www.moviebodycounts.com/

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)

Vanity Fair. http://www.vanityfair.com/hollywood/2013/02/quentin-tarantino-deaths-movies

Gelman et al.

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

The Tarantino Death Toll. https://vimeo.com/148832585