# Chapter 4

# Likelihoods

## 4.1 Blog blues

Suppose that visits to your newly launched blog occur sporadically. Imagine you are interested in the length of time between consecutive first-time visits to your homepage. You collect the time data for a random sample of 50 visits to your blog for a particular time period and day, and you decide to build a statistical model to fit the data.

**Problem 4.1.1.** What assumptions might you make about the first-time visits?

Assume visits occur continuously at a constant rate and independently of one another.

**Problem 4.1.2.** What might be an appropriate probability model for the time between visits?

If the number of visits is Poisson distributed (which it is if we have the above assumptions) then the time between the events is exponentially distributed.

**Problem 4.1.3.** Using your chosen probability distribution from the previous part, algebraically derive the maximum likelihood estimate (MLE) of the mean.

Assuming independence between time intervals we can write the overall likelihood as,

$$p(t_1, t_2, ..., t_T|\lambda) = \prod_{i=1}^{T} \lambda e^{-\lambda t_i} \tag{4.1}$$

We then take the log of this,

$$\log p(t_1, t_2, ..., t_T|\lambda) = \sum_{i=1}^{T} (\log \lambda - \lambda t_i) \tag{4.2}$$

$$= T\log \lambda - \lambda T\bar{t} \tag{4.3}$$

Now differentiating the above with respect to $\lambda$ and solving for the maximum,

$$\frac{\partial \log p}{\partial \lambda} = \frac{T}{\hat{\lambda}} - T\bar{t} = 0, \tag{4.4}$$

which has a solution $\hat{\lambda} = \frac{1}{\bar{t}}$, which makes sense intuitively: the mean event rate we estimate $\hat{\lambda}$ is the inverse of the average time interval between events.

**Problem 4.1.4.** You collect data from Google Analytics that contains the time (in minutes) between each visit for a sample of 50 randomly chosen visits to your blog. The data set is called `likelihood_blogVisits.csv`. Derive an estimate for the mean number of visits per minute.

Using the above estimator we estimate that $\hat{\lambda} \approx 1.63$ visits per minute.

**Problem 4.1.5.** Graph the log-likelihood near the MLE. Why do we not plot the likelihood?

Figure 4.1 shows this plot. We don't plot the likelihood as it is far too small for moderately-sized datasets.

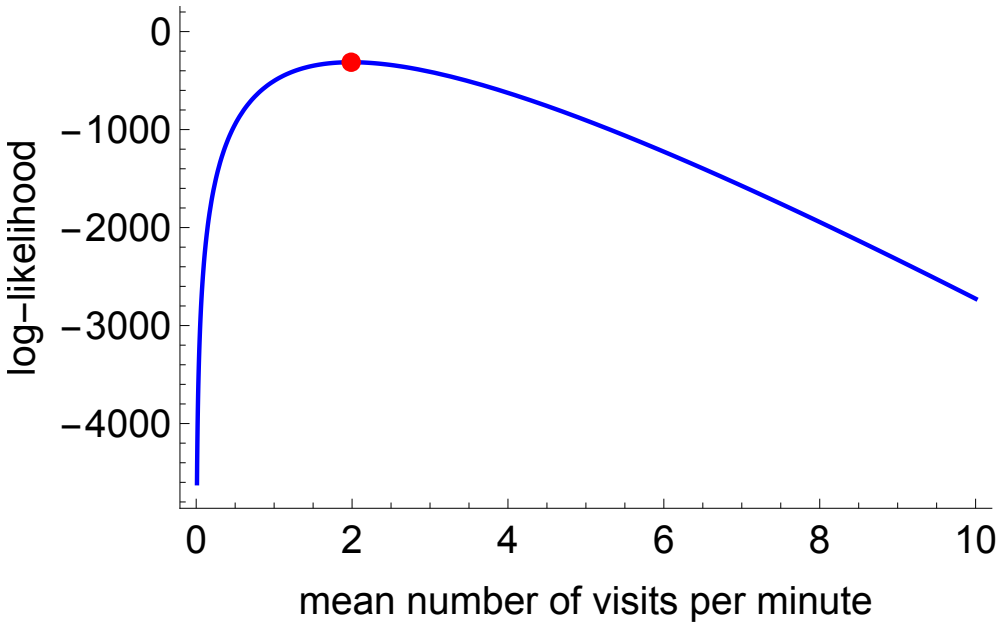


Figure 4.1: A plot of the log likelihood for the exponential model for blog visits. The maximum likelihood estimate is shown in red.

**Problem 4.1.6.** Estimate 95% confidence intervals around your estimate of the mean visit rate.

The correct way to do this is to calculate the Information matrix (here the observed and expected information matrices are the same),

$$\mathcal{I}(\lambda) = -\mathbb{E}\left(\frac{\partial^2 \log p}{\partial \lambda^2}\right) \tag{4.5}$$

$$= \frac{T}{\lambda^2} \tag{4.6}$$

To convert this to a confidence interval we note that asymptotically (because of the Cramer-Rao lower bound),

$$\sqrt{T}\left(\hat{\lambda} - \lambda\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \mathcal{I}(\hat{\lambda})^{-\frac{1}{2}}\right), \tag{4.7}$$

which means that for large samples we can approximate,

$$\hat{\lambda} \approx \mathcal{N}\left(\lambda, \frac{1}{\sqrt{T}}\mathcal{I}(\hat{\lambda})^{-\frac{1}{2}}\right), \tag{4.8}$$

where we have that $\frac{1}{\sqrt{T}}\mathcal{I}(\hat{\lambda})^{-\frac{1}{2}} = \frac{\hat{\lambda}}{T} = \frac{1.63}{50} \approx 0.0325$. We therefore construct 95% intervals using the $z = 1.96$ critical value from a standard normal,

$$1.56247 \leq \lambda \leq 1.68996, \tag{4.9}$$

which is a pretty tight interval!

**Problem 4.1.7.** What does this interval mean?

If we repeatedly sampled from the population (an infinite number of times) and for each sample constructed the 95% confidence interval, then 95% of those hypothetical samples would contain the true parameter value.

**Problem 4.1.8.** Using your maximum likelihood estimate, what is the probability you will wait: (a) 1 minute or more, (b) 5 minutes or more, (c) half an hour or more before your next visit?

We answer this question using the survival function of the exponential distribution,

$$Pr(t > a|\lambda) = e^{-\lambda a}. \tag{4.10}$$

So answering each part in turn,

1. $Pr(t > 1|\hat{\lambda}) \approx 0.20$.

2. $Pr(t > 5|\hat{\lambda}) \approx 0.0003$.
3. $Pr(t > 30|\hat{\lambda}) \approx 0.000$.

**Problem 4.1.9.** Evaluate your model.

One way to evaluate your model is to compare real data with that which you simulate from the exponential model at the maximum likelihood estimate of the rate parameter (see Figure 4.2). When we do this we see that the exponential model is clearly unable to generate the upper extremes that we see in the real data.

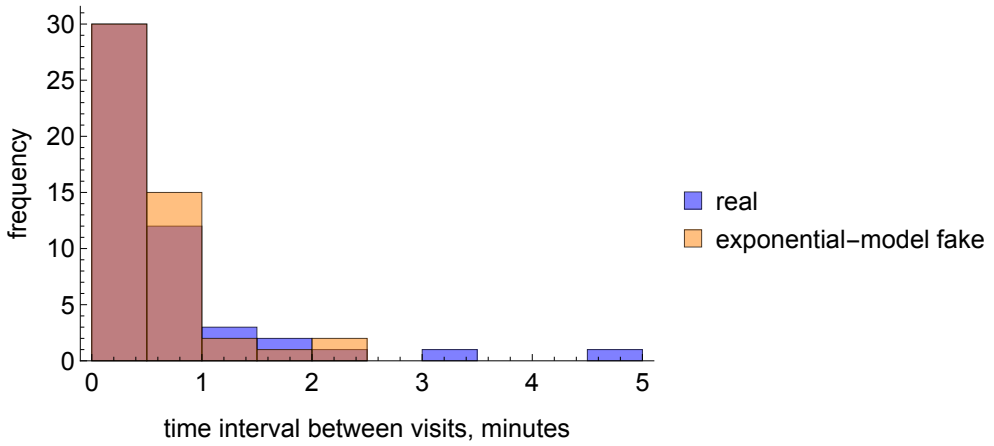


Figure 4.2: Real and fake data simulated from the exponential model using the maximum likelihood estimates of the parameter.

**Problem 4.1.10.** Can you think of a better model to use? What assumptions are relaxed in this model?

The Poisson is to the exponential what the negative binomial is to the generalised Pareto type 2 distribution, see:

http://stats.stackexchange.com/questions/37814/poisson-is-to-exponential-as-gamma-poisson-is-to-what

The events for the negative binomial distribution are no longer constrained to be independent.

**Problem 4.1.11.** Estimate the parameters of your new model, and hence estimate the mean number of website visits per minute.

The PDF for this model is of the form,

$$p(t|\alpha, \beta) = \frac{\alpha \left(\frac{\beta+t}{\beta}\right)^{-\alpha-1}}{\beta} \tag{4.11}$$

Using maximum likelihood estimators of the parameters we find that $(\hat{\alpha}, \hat{\beta}) = (0.984052, 2.52871)$. The mean of this distribution is given by,

$$\mathbb{E}(t|\alpha, \beta) = \frac{\beta}{\alpha - 1} \approx 0.64, \tag{4.12}$$

and so the mean number of website visits per minute is about 1.55.

**Problem 4.1.12.** Use your new model to estimate the probability that you will wait: (a) 1 minute or more, (b) 5 minutes or more, (c) half an hour or more before your next visit.

Using the survival function,

$$Pr(t > a|\alpha, \beta) = \left(\frac{a}{\beta} + 1\right)^{-\alpha}, \tag{4.13}$$

We now see (slightly) less extreme probabilities for waiting times,

1. $Pr(t > 1|\hat{\lambda}) \approx 0.17$.
2. $Pr(t > 5|\hat{\lambda}) \approx 0.01$.
3. $Pr(t > 30|\hat{\lambda}) \approx 0.0002$.

Whilst the question doesn't ask it, as before we can generate data from the model assuming the MLEs (see Figure 4.3), which now look in much better accordance with the real data.

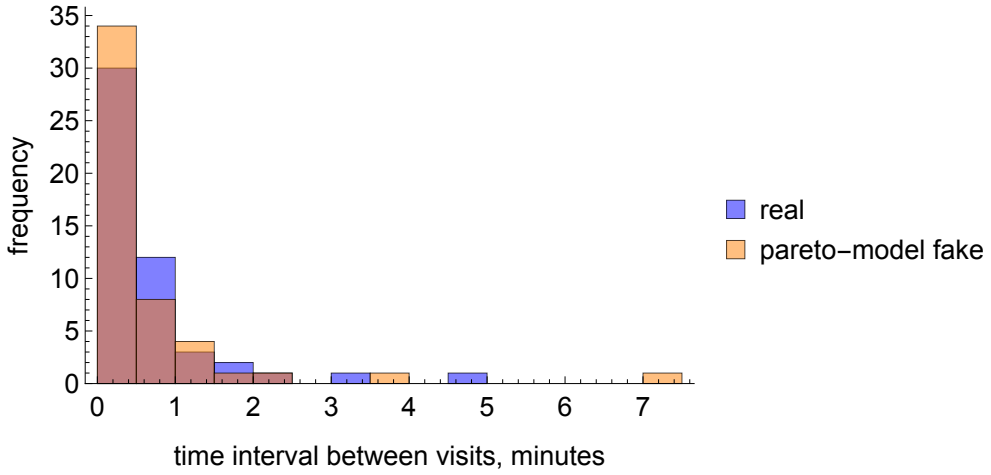


Figure 4.3: Real and fake data simulated from the pareto model of website visits using the maximum likelihood estimates of the parameter.

## 4.2 Violent crime counts in New York counties

In data file `likelihood_NewYorkCrimeUnemployment.csv` is a data set of the population, violent crime count and unemployment across New York counties in 2014 (openly available from the New York Criminal Justice website).

**Problem 4.2.1.** Graph the violent crime count against population size across all the counties. What type of relationship does this suggest?

A strong linear relationship (see Figure 4.4).

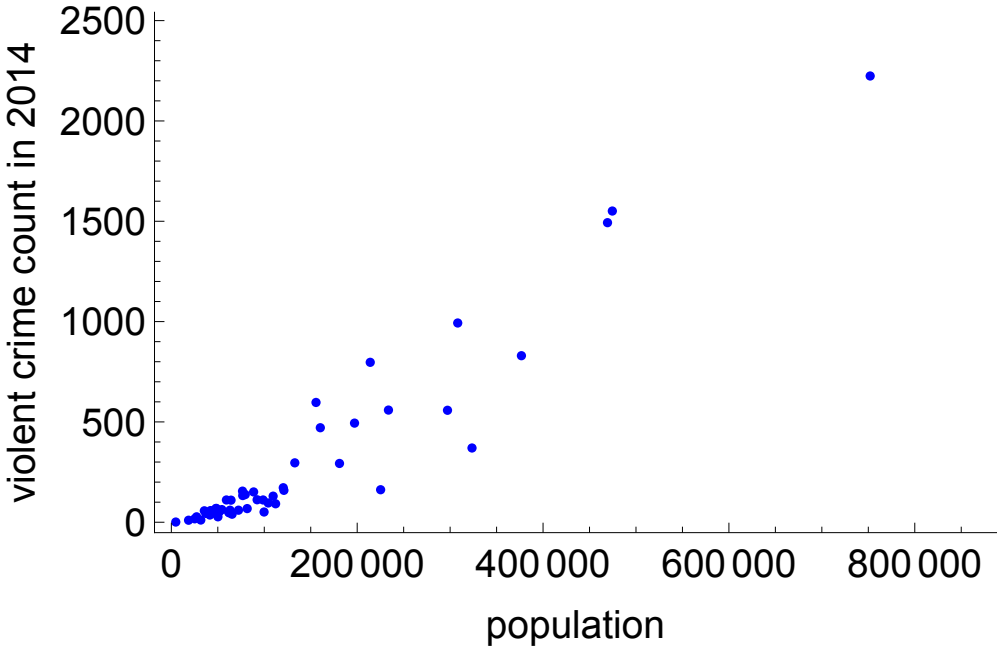


Figure 4.4: Population versus violent crime counts for New York counties in 2014.

**Problem 4.2.2.** A simple model here might be to assume that the crime count in a particular county is related to the population size by a Poisson model:

$$crime_i \sim Poisson(n_i\theta), \tag{4.14}$$

where $crime_i$ and $n_i$ are the crime count and population in county $i$. Write down an expression for the likelihood.

Assuming conditional independence between the pairs of observations $(crime_i, n_i)$,

$$p\left(\{crime_1, n_1\}, ..., \{crime_N, n_N\}|\theta\right) = \prod_{i=1}^{N} \frac{(n_i\theta)^{crime_i} e^{-n_i\theta}}{crime_i!}. \tag{4.15}$$

**Problem 4.2.3.** Find the maximum likelihood estimators of the parameters.

It's easiest to start by working with the log-likelihood,

$$\log p = \sum_{i=1}^{N} [crime_i \log(n_i\theta) - n_i\theta - \log crime_i!] \tag{4.16}$$

$$= -N\bar{n}\theta + \sum_{i=1}^{N} crime_i \log(n_i\theta) + const. \tag{4.17}$$

We then differentiate this expression,

$$\frac{\partial \log p}{\partial \theta} = -N\bar{n} + \frac{1}{\hat{\theta}} N\overline{crime} = 0, \tag{4.18}$$

which we then rearrange for $\hat{\theta} = \frac{\overline{crime}}{\bar{n}}$ which is the sample average violent crime count per capita. So in our case we have that,

$$\hat{\theta} = \frac{318417}{1165.69} \approx 0.004, \tag{4.19}$$

that is, a prevalence of about 4/1000.

**Problem 4.2.4.** By generating fake data, assess this model.

One way to compare the real and fake data across all counties is to look at the per capita crime rates (see Figure 4.5). We see that the variation in the real data is much greater than that we have in the fake, and hence our model is quite weak.

**Problem 4.2.5.** What are the assumptions of this model? And do you think that these hold in this case?

This model assumes that one instance of violent crime occurs independently of another, and that the underlying rate of violent crime is the same across all counties. Given that we found our model is deficient we suspect that one or both of the above are probably violated. This makes intuitive sense since violent crimes may often be linked to one another (violating independence) and the counties probably differ in societal ways meaning that some are more/less predisposed to crime.

**Problem 4.2.6.** Suggest an alternative model and estimate its parameters by maximum likelihood.

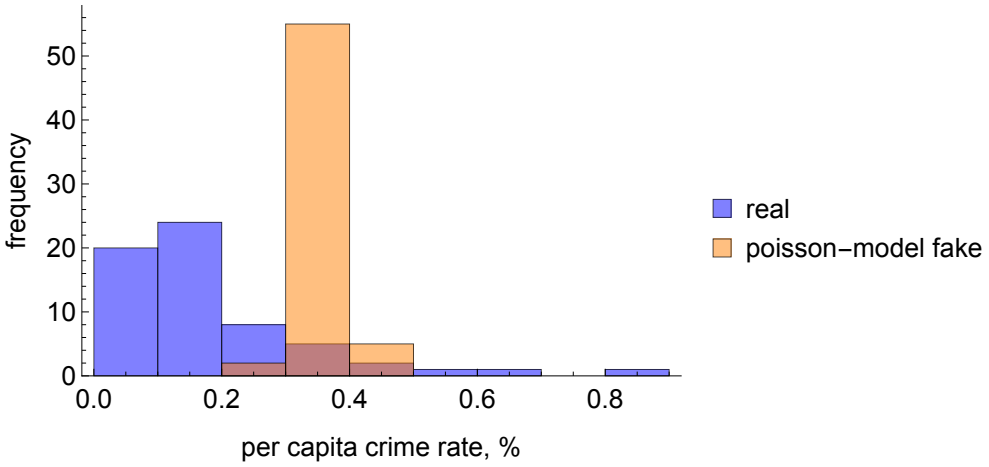


Figure 4.5: The per capita violent crime rate for New York counties in 2014 for the real and Poisson model generated datasets.

A better model that allows a degree of overdispersion in the data is the *negative binomial*, which has a pdf,

$$Pr(crime = x|\lambda,\kappa) = \left(\frac{k}{k+\lambda}\right)^{\frac{k\lambda}{(k+\lambda)\left(1-\frac{k}{k+\lambda}\right)}} \left(1-\frac{k}{k+\lambda}\right)^{x} \binom{x+\frac{k\lambda}{(k+\lambda)\left(1-\frac{k}{k+\lambda}\right)}-1}{\frac{k\lambda}{(k+\lambda)\left(1-\frac{k}{k+\lambda}\right)}-1}, \tag{4.20}$$

where $\lambda = \theta n_i$ is the mean and $\kappa \geq 0$ is the overdispersion parameter. Assuming conditional independence for each of the data points we can again estimate the model via maximum likelihood and obtain $(\hat{\theta}, \hat{\kappa}) = (0.00185, 2.2222)$.

**Problem 4.2.7.** Evaluate this new model.

Again we can compare model-simulated data with the actual and this time we find that there is much better correspondence (see Figure 4.6)

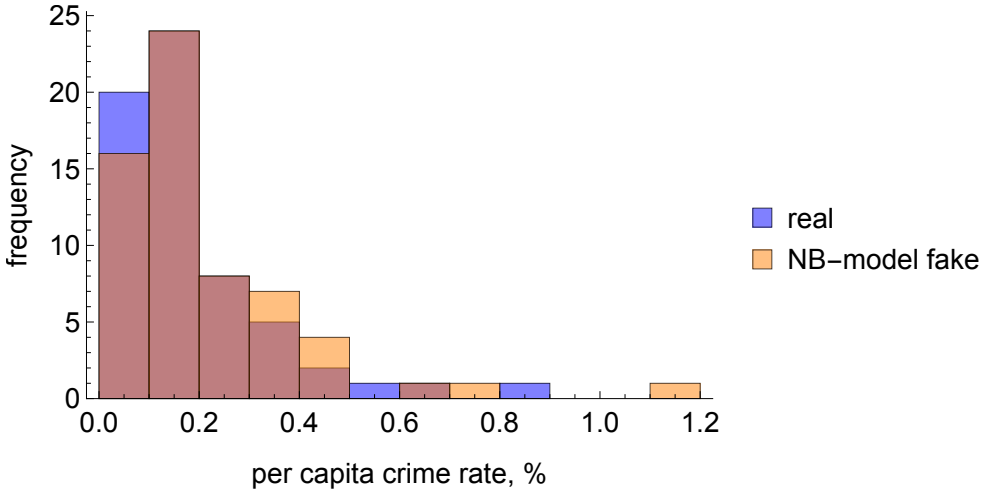


Figure 4.6: The per capita violent crime rate for New York counties in 2014 for the real and negative binomial model generated datasets.

# Bibliography