STA 360/601: Bayesian and Modern Statistics

Lecture 3:
Count data, Gamma-Poisson model, & Posterior summaries

Jeff Miller

Department of Statistical Science, Duke University

Wednesday, September 3, 2014

Count data

Suppose our data is counts $y_i \in \mathcal{Y} = \{0, 1, 2, \dots\}$ for $i = 1, \dots, n$.

- e.g., # friends on facebook, # website hits per minute, # points scored in a game, # neuron spikes in a given interval, # photons hitting a CCD pixel.
- Often, a natural choice of likelihood is Poisson:

$$L(y; \theta) = \prod_{i=1}^{n} \frac{\exp(-\theta)\theta^{y_i}}{y_i!},$$

assuming conditional independence of the counts given θ.

Siméon Denis Poisson (1781 – 1840)

Poisson distribution

$Y \sim \text{Poisson}(\theta)$ (or $\text{Pois}(\theta)$), where $\theta > 0$, means

$$\Pr(Y = y \mid \theta) = \frac{\theta^y}{y!} e^{-\theta}.$$

Notes:

- <u>Mean = Variance:</u> $E(Y|\theta) = V(Y|\theta) = \theta$.
- <u>Sum of Poissons is Poisson:</u>
  If $Y_i \overset{ind}{\sim} \text{Pois}(\theta_i)$ for $i = 1, ..., n$, then $\sum Y_i \sim \text{Pois}(\sum \theta_i)$.
- Limit of $\text{Binomial}(n, p_n)$ with $p_n = \theta/n$ as $n \to \infty$ is Poisson:

$$\binom{n}{y} (\theta/n)^y (1 - \theta/n)^{n-y} \longrightarrow \frac{\theta^y}{y!} e^{-\theta}$$

  (special case of the "law of small numbers").

Fake real-world example

- You are planning to start a pizza delivery business.
- It's essential to know how many orders you will get.
- A priori, you think your average number of orders/hour will be around 15–25 (in the evening), but you're not really sure.
- To get some data, you stakeout a comparable pizza delivery business over a few evenings, and record how many deliveries they make each hour. Over $n = 6$ hours, you observe

$$y_{1:n} = (16, 10, 22, 14, 19, 18).$$

- More data would be nice but you've already spent 6 hours ... you can use your prior knowledge to help make inferences.
- You're happy with a Poisson likelihood, but to do a Bayesian analysis, you also need a prior on θ, the mean # pizzas/hour.

Sufficient statistics for Poisson

- The likelihood simplifies:

$$L(y;\theta) = \prod_{i=1}^{n} \frac{\theta^{y_i} \exp(-\theta)}{y_i!} = C(y)\, \theta^{\sum_{i=1}^{n} y_i} \exp(-n\theta).$$

- $S(y) = \sum y_i$ is a *sufficient statistic*: as a function of θ the likelihood depends only on $S(y)$, up to a constant of proportionality $C(y)$.
- Intuitive interpretation: $S(y)$ contains all the information about θ present in the data. "$Y \perp \theta | S(Y)$"
- Practical upshot: We don't need to store the individual counts $y_1, \ldots, y_n$ — just keep the sum (and n).
- As a function of θ, $L(y;\theta) \propto \theta^{S(y)} \exp(-n\theta)$. The Gamma distribution gives us a conjugate prior.

Gamma distribution

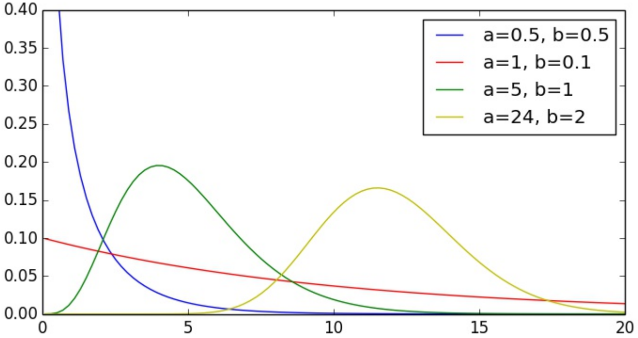
$\theta \sim \text{Ga}(a, b)$ (where $a, b > 0$) means the pdf of θ is

$$\text{Ga}(\theta; a, b) = \frac{b^a}{\Gamma(a)} \theta^{a-1} \exp(-b\theta).$$

Notes:

- $a =$ "shape", $b =$ "rate".
- <u>Achtung!</u> Alternate parametrizations are in common use.
- $\text{E}(\theta) = a/b, \quad \text{V}(\theta) = a/b^2$
- To obtain a given prior mean and std. dev. $\mu > 0$ and $\sigma > 0$, we can solve for a, b s.t. $\mu = a/b$ and $\sigma^2 = a/b^2$.
- <u>Sum of Gammas:</u>
  If $\theta_i \overset{ind}{\sim} \text{Ga}(a_i, b)$ for $i = 1, \ldots, n$, then $\sum \theta_i \sim \text{Ga}(\sum a_i, b)$.
- <u>Scaling:</u>
  If $\theta \sim \text{Ga}(a, b)$ and $c > 0$, then $c\theta \sim \text{Ga}(a, b/c)$.

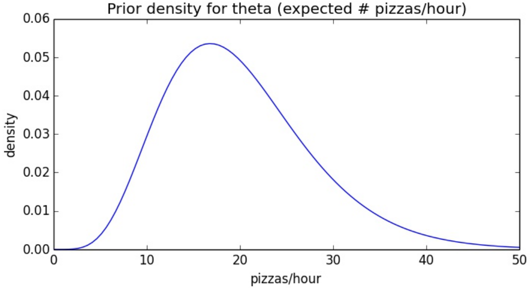
Some Gamma densities for various a, b



Pizza prior

Back to pizza . . .

- You need to put a prior on θ (mean # pizzas sold per hour).
- For convenience, you choose a Gamma prior.
- Based on your (somewhat uncertain) prior belief, you choose $\mu = 20$ and $\sigma = 8$, thus $b = \mu/\sigma^2 = 0.3125 \approx 0.31$ and $a = b\mu = 6.25$.



Posterior of Gamma-Poisson model

$$\begin{aligned}
\pi(\theta|y) &\propto \text{likelihood} \times \text{prior} = L(y;\theta)\text{Ga}(\theta;a,b) \\
&\propto \theta^{S(y)} \exp(-n\theta)\, \theta^{a-1} \exp(-b\theta) \\
&\propto \theta^{a+S(y)-1} \exp\big(-\theta(b+n)\big) \\
&\propto \text{Ga}(\theta;\widehat{a},\widehat{b}),
\end{aligned}$$

where $\widehat{a} = a + S(y)$, $\widehat{b} = b + n$, and $S(y) = \sum_i y_i$.

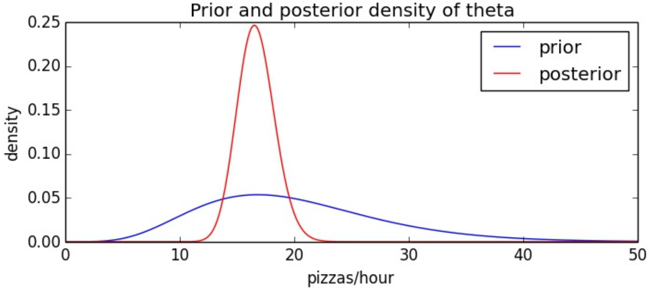
- Can roughly interpret b as prior "sample size".
- Posterior mean: (convex combo of prior & sample means)

$$\text{E}(\theta|y) = \frac{a + \sum y_i}{b+n} = \frac{b}{b+n} a/b + \frac{n}{b+n}\overline{y}.$$

- $\text{E}(\theta|y) \approx \overline{y} = \frac{1}{n}\sum y_i$ for large n.
- Posterior variance: $\text{V}(\theta|y) = (a + \sum y_i)/(b+n)^2 \approx \overline{y}/n$ for large n.

Pizza posterior

- Your angel investor just called, and he wants to know how many pizzas you expect to sell per hour, on average?
- And how certain are you about that?
- Posterior: $\text{Ga}(\theta; \widehat{a}, \widehat{b})$ with $\widehat{a} = a + S(y) = 105.25$, $\widehat{b} = b + n \approx 6.31$.



- Your investor never took Bayesian statistics, so you need to summarize this posterior.

Posterior Intervals

- Names: Credible intervals/sets, Bayesian confidence intervals/sets, Posterior intervals.
- Central intervals (equal tails):
  $[\ell(y), u(y)]$ is a $100(1-\alpha)\%$ central credible interval if

  $$\Pr(\theta < \ell(y)|y) = \alpha/2, \quad \Pr(\theta > u(y)|y) = \alpha/2.$$

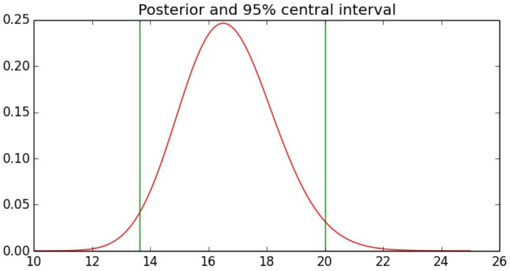
  E.g., for a 95% interval, choose $\alpha = 0.05$.
- Highest posterior density (HPD) set:
  A set $A(y)$ is a $100(1-\alpha)\%$ HPD set if

  $$\Pr(\theta \in A(y)|y) = 1 - \alpha$$

  and $\pi(\theta_1|y) \geq \pi(\theta_2|y)$ for any $\theta_1 \in A(y)$, $\theta_2 \notin A(y)$.

Pizza interval

In our example, $[13.6, 20.0]$ is the 95% central credible interval:



- You can tell your investor that <u>your belief</u> is that there is a 95% probability that a business like yours sells between 13.6 and 20 pizzas/hour, on average.
- (Note: This is a statement about the mean, not the # in any given hour.)

Bayesian vs. Frequentist intervals

- Bayesian confidence intervals (credible intervals) are different than frequentist confidence intervals.
- Bayesian:
  $\Pr(\ell(y) < \theta < u(y) \mid y) = 0.95$ for any y. (θ is random)
- Frequentist:
  $\Pr(\ell(Y) < \theta < u(Y) \mid \theta) = 0.95$ for any θ. (Y is random)
- If you had constructed a frequentist interval, you would tell your investor that 95% of the time, an analysis like yours would yield an interval containing the true value.
- Credible intervals do not always guarantee coverage in the frequentist sense — however, they do asymptotically (see Hoff, p. 41). Many frequentist methods also only guarantee coverage asymptotically.

Hiring — a decision problem

- Your investor is satisfied for now, but you have a new problem: How many delivery people should you have working each evening?
- Each deliverer costs you $c = 14$ dollars/hour.
- Each deliverer can handle a maximum of $m = 6$ orders/hour.
- If you have d deliverers, you can handle md orders/hour.
- Your business guarantees delivery within 30 minutes, so for each order in excess of md you lose \$20.
- Loss function (dollars/hour): $\mathcal{L}(d, y) = cd + 20\max(y - md, 0) = 14d + 20\max(y - 6d, 0)$.
- Bayes risk: $R(d) = \mathrm{E}(\mathcal{L}(d, y_{n+1})|y_{1:n})$.
- To compute this, you need the posterior predictive $y_{n+1}|y_{1:n}$.

Prediction

We need the posterior predictive pmf $f(y_{n+1}|y_{1:n})$. To simplify notation, write y for y_{n+1}.

$$
\begin{aligned}
f(y|y_{1:n}) &= \int \text{Pois}(y;\theta)\text{Ga}(\theta;\widehat{a},\widehat{b})d\theta \\
&= \frac{\widehat{b}^{\widehat{a}}}{y!\Gamma(\widehat{a})}\int_0^\infty \theta^{\widehat{a}+y-1}\exp\big(-\theta(\widehat{b}+1)\big)d\theta \\
&= \frac{\widehat{b}^{\widehat{a}}}{y!\Gamma(\widehat{a})}\frac{\Gamma(\widehat{a}+y)}{(\widehat{b}+1)^{\widehat{a}+y}} \\
&= \frac{\Gamma(\widehat{a}+y)}{\Gamma(y+1)\Gamma(\widehat{a})}\left(\frac{\widehat{b}}{\widehat{b}+1}\right)^{\widehat{a}}\left(\frac{1}{\widehat{b}+1}\right)^{y}.
\end{aligned}
$$

This is the negative-binomial dist, $\text{NegBinom}(\widehat{a}, 1/(\widehat{b}+1))$.

Prediction (continued)

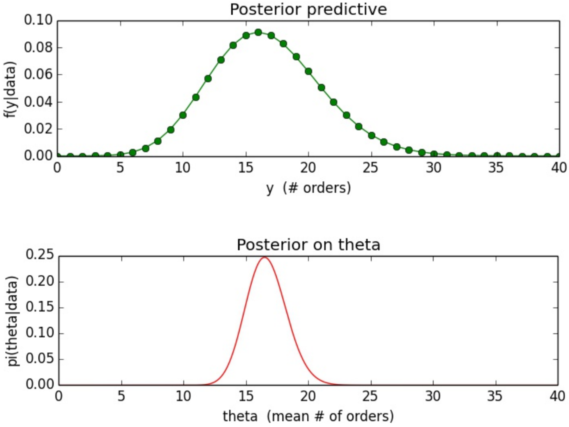
- In marginalizing θ out of the Poisson$(y; \theta)$ likelihood over a gamma distribution, we obtain a negative-binomial.
- The negative-binomial distribution also models count data, but has somewhat more flexibility, with two parameters, allowing control of mean and variance.
- For $(y|y_{1:n}) \sim \text{NegBinom}(\widehat{a}, 1/(\widehat{b}+1))$, we have

$$
\begin{aligned}
\mathsf{E}(y|y_{1:n}) &= \widehat{a}/\widehat{b} = \mathsf{E}(\theta|y_{1:n}) = \text{Posterior mean} \\
\mathsf{V}(y|y_{1:n}) &= \frac{\widehat{a}(\widehat{b}+1)}{\widehat{b}^2} = \mathsf{E}(\theta|y_{1:n})\left(\frac{\widehat{b}+1}{\widehat{b}}\right).
\end{aligned}
$$

So, the variance is larger than the mean by an amount determined by $\widehat{b}$.

Pizza prediction

The posterior predictive distribution of the number of pizza orders in a given hour is $\text{NegBinom}(y; \widehat{a}, 1/(\widehat{b}+1))$.



Predictive uncertainty

- Note that as the sample size n increases, the posterior density for θ becomes more and more concentrated: $\mathsf{V}(\theta|y_{1:n}) = \widehat{a}/\widehat{b}^2 = (a + \sum_i y_i)/(b+n)^2 \approx \overline{y}/n \to 0$.
- As we have less uncertainty about θ, the inflation factor $(\widehat{b}+1)/\widehat{b} \to 1$ and the predictive density $f(y|y_{1:n}) \to \text{Pois}(\overline{y})$.
- In smaller samples, though, using this approximation can lead one to underestimate predictive variance, since it's important to account for uncertainty in $\theta|y_{1:n}$ (not just in $y|\theta$).

More on the Negative Binomial

- Can be derived as the # successes in a sequence of Bernoulli(p) trials before r failures occur.
- This is denoted $Y \sim \text{NegBinom}(r, p)$ and the pmf is

$$\Pr(Y = k) = \binom{k + r - 1}{k} (1 - p)^r p^k.$$

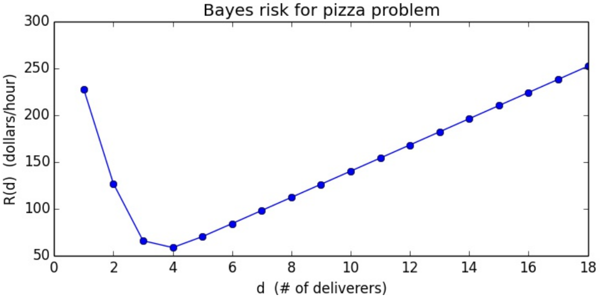
- Starting with this, the distribution can be extended to allow noninteger $r \in (0, \infty)$ as

$$\Pr(Y = k) = \frac{\Gamma(k + r)}{\Gamma(k + 1)\Gamma(r)} (1 - p)^r p^k,$$

  which is the form we obtained above as the predictive with $r = \widehat{a}$, $p = 1/(\widehat{b} + 1)$.

How many deliverers to have?

- Loss function: $\mathcal{L}(d, y) = 14d + 20 \max(y - 6d, 0)$.
- Bayes risk: $R(d) = \mathrm{E}(\mathcal{L}(d, y)|y_{1:n}) = \sum_y \mathcal{L}(d, y) f(y|y_{1:n})$.
- Looks nasty to compute analytically, but it's easy numerically.



- If too few, we often have to pay for > 30 minute deliveries.
- If too many, have to pay too much in wages, etc.

Homework exercise

- Suppose for subjects $1, \ldots, n$, we observe that y_i is the length of time it takes to perform a task.
- Assume $y_i \overset{iid}{\sim} \text{Exp}(\theta)$ given θ:

$$L(y; \theta) = \prod_{i=1}^{n} \theta \exp(-\theta y_i)$$

- Assume $\theta \sim \text{Ga}(a, b)$ a priori.
- Calculate the posterior distribution of θ.
- Calculate the posterior predictive distribution $f(y_{n+1} | y_{1:n})$.
- Describe how this could be used for prediction, including quantification of uncertainty.