One parameter models cont'd; Loss functions and Bayes risk

Dr. Olanrewaju Michael Akande

Jan 22, 2020

Announcements

- Add/drop today
- HW1 due tomorrow
- Take the participation quiz for today on Sakai

Outline

- Loss functions and Bayes risk
- Frequentist vs Bayesian intervals
- Poisson-Gamma model
  - Recap of the distributions
  - Conjugacy
  - Example
  - Posterior prediction
  - Other parameterizations

Loss functions and Bayes risk

Bayes estimate

- As we've seen by now, having posterior distributions instead of one-number summaries is great for capturing uncertainty.
- That said, it is still very appealing to have simple summaries, especially when dealing with clients or collaborators from other fields, who desire one.
- Can we obtain a single estimate of θ based on the posterior? Sure!
- Bayes estimate is the value $\hat{\theta}$, that minimizes the Bayes risk.
- Bayes risk is defined as the expected loss averaged over the posterior distribution.
- Put differently, a Bayes estimate $\hat{\theta}$ has the lowest posterior expected loss.
- That's fine, but what does expected loss mean?
- Frequentist risk also exists but we won't go into that here.

Loss functions

- A loss function $L(\theta, \delta(y))$ is a function of a parameter θ, where $\delta(y)$ is some decision about θ, based on just the data y.

- For example, $\delta(y) = \bar{y}$ can be the decision to use the sample mean to estimate θ, the true population mean.

- $L(\theta, \delta(y))$ determines the penalty for making the decision $\delta(y)$, if θ is the true parameter; $L(\theta, \delta(y))$ characterizes the price paid for errors.

- A common choice for example, when dealing with point estimation, is the squared error loss, which has

$$L(\theta, \delta(y)) = (\theta - \delta(y))^2.$$

- Bayes risk is thus

$$\rho(\theta, \delta) = \mathbb{E}\left[\, L(\theta, \delta(y)) \middle|\, y\right] = \int L(\theta, \delta(y))\, p(\theta|y)\, d\theta,$$

and we proceed to find the value $\hat{\theta}$, that is, the decision $\delta(y)$, that minimizes the Bayes risk.

Bayes estimator under squared error loss

- Turns out that, under squared error loss, the decision $\delta(y)$ that minimizes the posterior risk is the posterior mean.
- Proof: Let $L(\theta, \delta(y)) = (\theta - \delta(y))^2$. Then,

$$\begin{aligned} \rho(\theta, \delta) &= \int L(\theta, \delta(y))\, p(\theta|y)\, d\theta. \\ &= \int (\theta - \delta(y))^2\, p(\theta|y)\, d\theta. \end{aligned}$$

- Expand, then take the partial derivative of $\rho(\theta, \delta)$ with respect to $\delta(y)$.
- To be continued on the board!
- Easy to see then that $\delta(y) = \mathbb{E}[\theta|x]$ is the minimizer.
- Well that's great! The posterior mean is often very easy to calculate in most cases. In the beta-binomial case, we have

$$\hat{\theta} = \frac{a + y}{a + b + n}.$$

What about other loss functions?

- Clearly, squared error is only one possible loss function. An alternative is absolute loss, which has

$$L(\theta, \delta(y)) = |\theta - \delta(y)|.$$

- Absolute loss places less of a penalty on large deviations & the resulting Bayes estimate is **posterior median**.
- Median is actually relatively easy to estimate.
- Recall that for a continuous random variable Y with cdf F, the median of the distribution is the value z, which satisfies

$$F(z) = \Pr(Y \leq z) = \frac{1}{2} = \Pr(Y \geq z) = 1 - F(z).$$

- As long as we know how to evaluate the CDF of the distribution we have, we can solve for z.
- Think R!

What about other loss functions?

- For the beta-binomial model, the CDF of the beta posterior can be written as

$$F(z) = \Pr(\theta \leq z|y) = \int_0^z \text{beta}(\theta; a+y, b+n-y) d\theta.$$

- Then, if $\hat{\theta}$ is the median, we have that $F(\hat{\theta}) = 0.5$.
- To solve for $\hat{\theta}$, apply the inverse CDF $\hat{\theta} = F^{-1}(0.5)$.
- In R, that's simply

```
qbeta(0.5,a+y,b+n-y)
```

- For other popular distributions, switch out the beta.

Loss functions and decisions

- Loss functions are not specific to estimation problems but are a critical part of decision making.

- For example, suppose you are deciding how much money to bet ($A) on Duke in the first UNC-Duke men's basketball game this year (next month).

- Suppose, if Duke
  - loses (y = 0), you lose the amount you bet ($A)
  - wins (y = 1), you gain B per $1 bet

- What is a good sampling distribution for y here?

- Then, the loss function can be characterized as

$$L(A, y) = A(1 - y) - y(BA),$$

  with your action being the amount bet A.

- When will your bet be "rational"?

How much to bet on Duke?

- y is an unknown state, but we can think of it as a new prediction y_{n+1} given that we have data from win-loss records $(y_{1:n})$ that can be converted into a Bayesian posterior,

  $$\theta \sim \text{beta}(a_n, b_n),$$

  with this posterior concentrated slightly to the left of 0.5, if we only use data on UNC-Duke games (UNC men lead Duke 139-112 all time).

- Actually, it might make more sense to focus on more recent head-to-head data and not the all time record.

- In fact, we might want to build a model that predicts the outcome of the game using historical data & predictors (current team rankings, injuries, etc).

- However, to keep it simple for this illustration, go with the posterior above.

How much to bet on Duke?

- The Bayes risk for action A is then the expectation of the loss function,

$$\rho(A) = \mathbb{E}\left[\, L(A,y) \middle|\, y_{1:n}\right].$$

- To calculate this as a function of A and find the optimal A, we need to marginalize over the **posterior predictive distribution** for y.

- Why are we using the posterior predictive distribution here instead of the posterior distribution?

- Recall from the last class that

$$p(y_{n+1}|y_{1:n}) = \frac{a_n^{y_{n+1}} b_n^{1-y_{n+1}}}{a_n + b_n}; \quad y_{n+1} = 0, 1.$$

- Specifically, that the posterior predictive distribution here is $\text{Bernoulli}(\hat{\theta})$, with

$$\hat{\theta} = \frac{a_n}{a_n + b_n}$$

- By the way, what do a_n and b_n represent?

How much to bet on Duke?

- With the loss function $L(A, y) = A(1 - y) - y(BA)$, and using the notation y_{n+1} instead of y (to make it obvious the game has not been played), the Bayes risk (expected loss) for bet A is

$$\begin{aligned} \rho(A) &= \mathbb{E}\left[\left. L(A, y_{n+1}) \right| y_{1:n}\right] = \mathbb{E}\left[\left. A(1 - y_{n+1}) - y_{n+1}(BA) \right| y_{1:n}\right] \\ &= A\,\mathbb{E}\left[\left. 1 - y_{n+1} \right| y_{1:n}\right] - (BA)\,\mathbb{E}\left[\left. y_{n+1} \right| y_{1:n}\right] \\ &= A\ \left(1 - \mathbb{E}\left[\left. y_{n+1} \right| y_{1:n}\right]\right) - (BA)\,\mathbb{E}\left[\left. y_{n+1} \right| y_{1:n}\right] \\ &= A\ \left(1 - \mathbb{E}\left[\left. y_{n+1} \right| y_{1:n}\right](1 + B)\right). \end{aligned}$$

- Hence, your bet is rational as long as

$$\mathbb{E}\left[\left. y_{n+1} \right| y_{1:n}\right](1 + B) > 1.$$

- Clearly, there is no limit to the amount you should bet if this condition is satisfied (the loss function is clearly too simple).

- Loss function needs to be carefully chosen to lead to a good decision - finite resources, diminishing returns, limits on donations, etc.

- Want more on loss functions, expected loss/utility, or decision problems in general? Consider taking a course on decision theory (STA623?).

Frequentist vs Bayesian intervals

Frequentist confidence intervals

- Recall that a frequentist confidence interval $[l(y); u(y)]$ has 95% frequentist coverage for a population parameter θ if, before we collect the data,

$$\Pr[l(y) < \theta < u(y) | \theta] = 0.95.$$

- This means that 95% of the time, our constructed interval will cover the true parameter, and 5% of the time it won't.
- In any given sample, you don't know whether you're in the lucky 95% or the unlucky 5%.
- You just know that either the interval covers the parameter, or it doesn't (useful, but not too helpful clearly). There is NOT a 95% chance your interval covers the true parameter once you have collected the data.
- Asking about the definition of a confidence interval is tricky, even for those who know what they're doing.

Bayesian intervals

- An interval $[l(y); u(y)]$ has 95% Bayesian coverage for θ if

$$\Pr[l(y) < \theta < u(y) | Y = y] = 0.95.$$

- This describes our information about where θ lies *after* we observe the data.
- Fantastic!
- This is actually the interpretation people want to give to the frequentist confidence interval.
- Bayesian interval estimates are often generally called credible intervals.

Bayesian quantile-based interval

- The easiest way to obtain a Bayesian interval estimate is to use posterior quantiles.

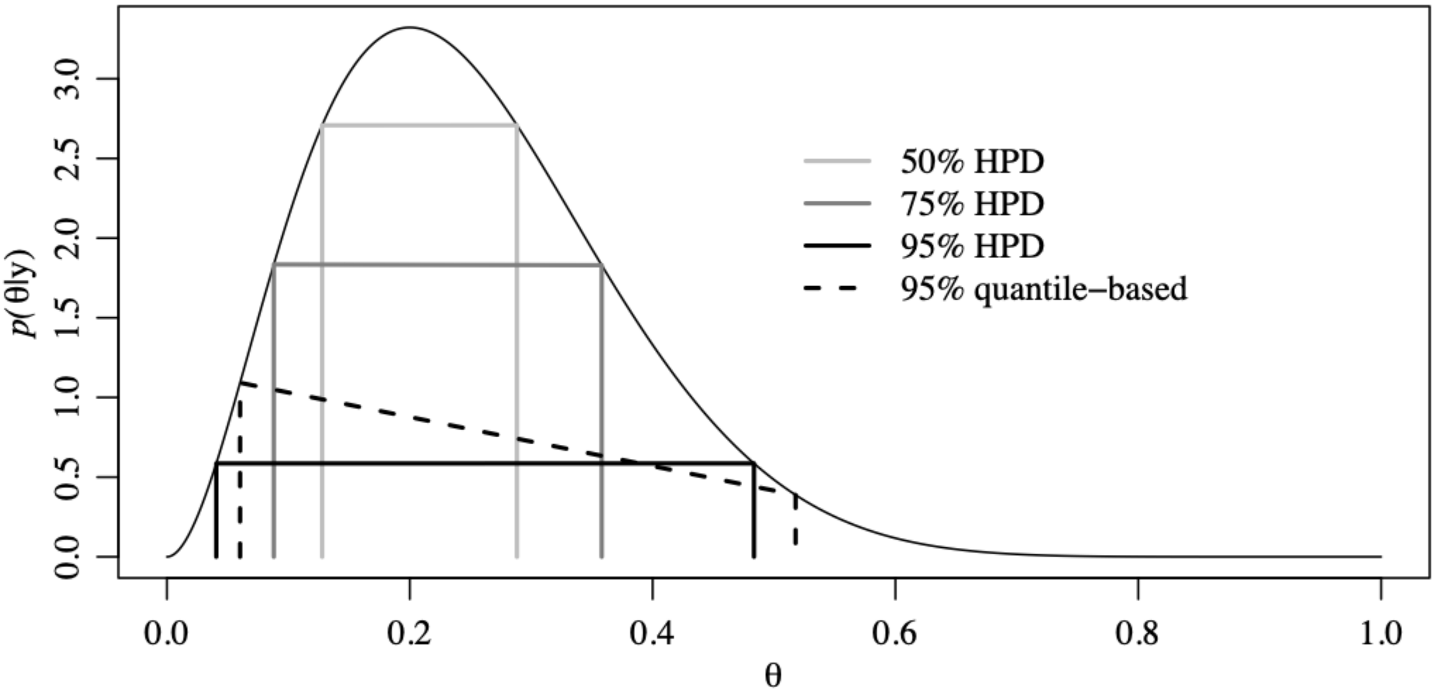
- Easy since we either know the posterior densities exactly or can sample from the distributions.

- To make a $100 \times (1-\alpha)$ quantile-based credible interval, find numbers (quantiles) $\theta_{\alpha/2} < \theta_{1-\alpha/2}$ such that

  1. $\Pr(\theta < \theta_{\alpha/2} | Y = y) = \frac{\alpha}{2}$; and
  2. $\Pr(\theta > \theta_{1-\alpha/2} | Y = y) = \frac{\alpha}{2}$.

- This is an equal-tailed interval. Often when researchers refer to a credible interval, this is what they mean.

Equal-tailed quantile-based interval



- This is Figure 3.6 from the Hoff book. Focus on the quantile-based credible interval for now.

- Note that there are values of θ outside the quantile-based credible interval, with higher density than some values inside the interval. This suggests that we can do better with interval estimation.

HPD REGION

- A $100 \times (1-\alpha)$ highest posterior density (HPD) region is a subset $s(y)$ of the parameter space Θ such that

  1. $\Pr(\theta \in s(y) | Y = y) = 1 - \alpha$; and

  2. If $\theta_a \in s(y)$ and $\theta_b \notin s(y)$, then $\Pr(\theta_a | Y = y) > \Pr(\theta_b | Y = y)$.

- $\Rightarrow$ **All** points in a HPD region have higher posterior density than points outside the region.

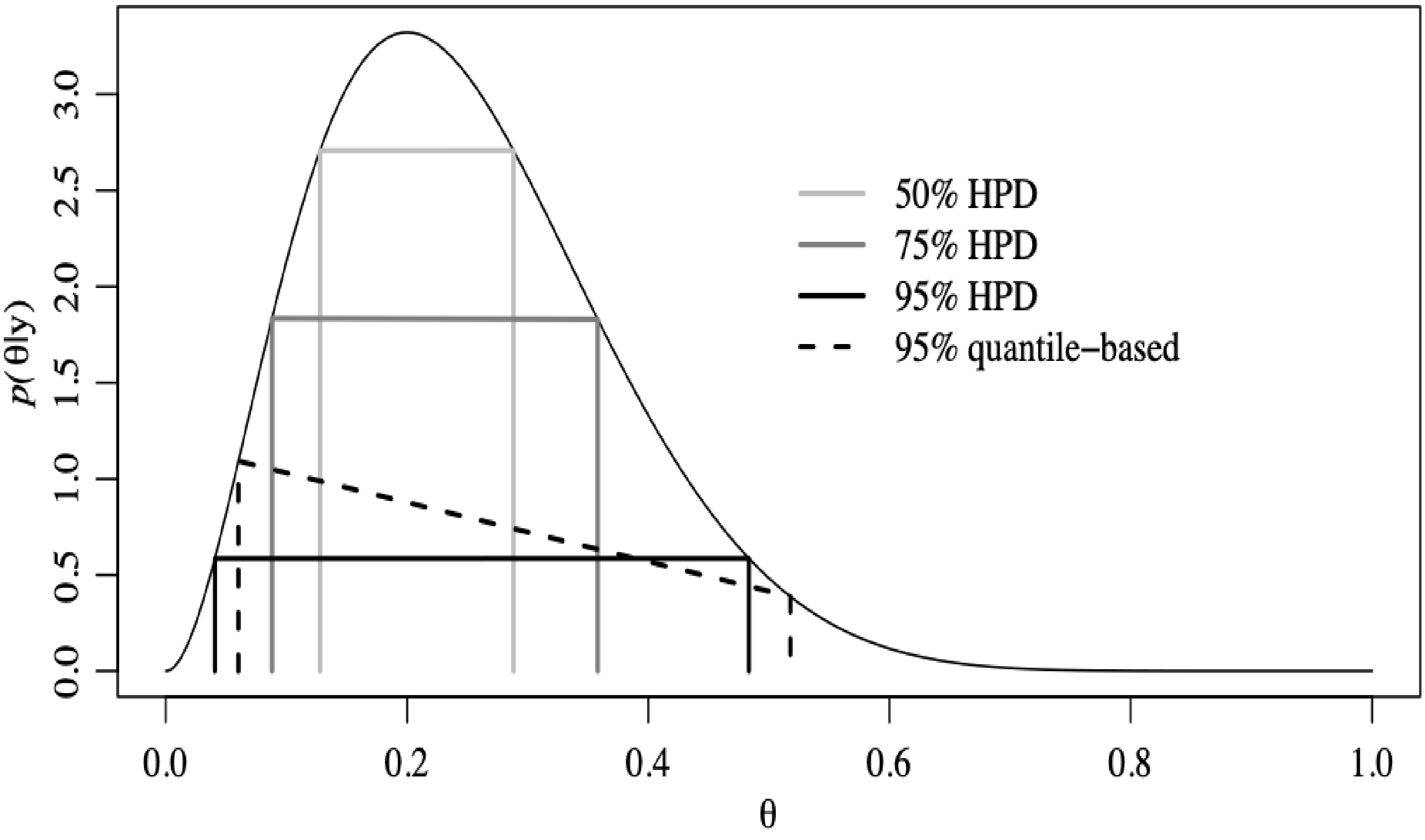
  Note this region is not necessarily a single interval (e.g., in the case of a multimodal posterior).

- The basic idea is to gradually move a horizontal line down across the density, including in the HPD region all values of θ with a density above the horizontal line.

- Stop moving the line down when the posterior probability of the values of θ in the region reaches $1 - \alpha$.

HPD REGION

Hoff Figure 3.6 shows how to construct an HPD region.



Poisson-gamma model

Poisson distribution recap

- $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathrm{Po}(\theta)$ denotes that each Y_i is a Poisson random variable.
- The Poisson distribution is commonly used to model count data consisting of the number of events in a given time interval.
- Some examples: # children, # lifetime romantic partners, # songs on iPhone, # tumors on mouse, etc.
- The Poisson distribution is parameterized by θ and the pmf is given by

$$\Pr[Y_i = y_i | \theta] = \frac{\theta^{y}_{i} e^{-\theta}}{y_i!}; \quad y_i = 0, 1, 2, \ldots; \quad \theta > 0.$$

where

$$\mathbb{E}[Y_i] = \mathbb{V}[Y_i] = \theta.$$

- What is the joint likelihood? What is the best guess (MLE) for the Poisson parameter? What is the sufficient statistic for the Poisson parameter?

Gamma density recap

- The gamma density will be useful as a prior for parameters that are strictly positive.

- If $\theta \sim \mathrm{Ga}(a, b)$, we have the pdf

  $$f(\theta) = \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta}.$$

  where a is known as the shape parameter and b, the rate parameter.

- Another parameterization uses the scale parameter $\phi = 1/b$ instead of b.

- Some properties:
  - $\mathbb{E}[\theta] = \dfrac{a}{b}$
  - $\mathbb{V}[\theta] = \dfrac{a}{b^2}$
  - $\mathrm{Mode}[\theta] = \dfrac{a-1}{b} \;\; \text{for} \;\; a \geq 1$

Gamma density

- If our prior guess of the expected count is μ & we have a prior "scale" ϕ, we can let

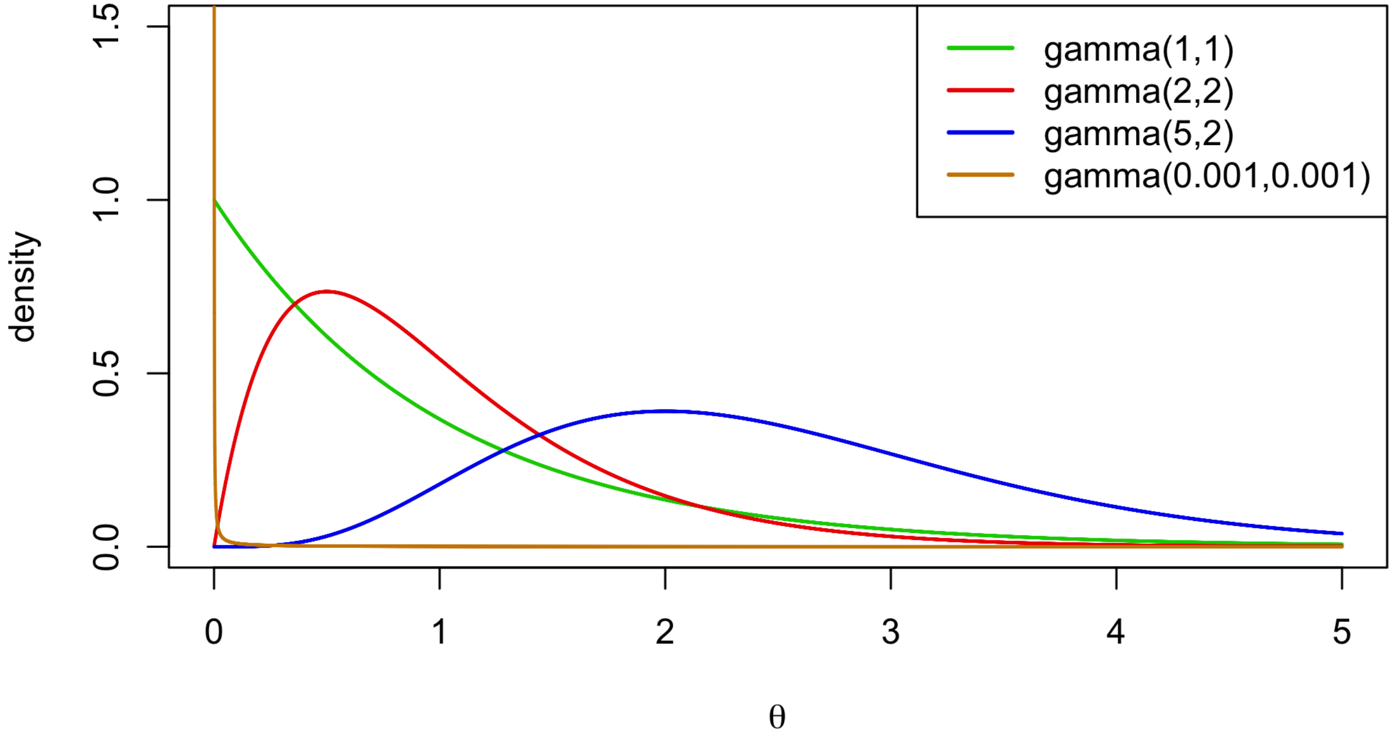
  $$\mathbb{E}[\theta] = \mu = \frac{a}{b}; \ \ \mathbb{V}[\theta] = \mu\phi = \frac{a}{b^2},$$

  and solve for a, b. We can play the same game if we have a prior variance or standard deviation.

- More properties:
  - If $\theta_1, \ldots, \theta_p \overset{ind}{\sim} \text{Ga}(a_i, b)$, then $\sum_i \theta_i \sim \text{Ga}(\sum_i a_i, b)$.
  - If $\theta \sim \text{Ga}(a, b)$, then for any $c > 0$, $c\theta \sim \text{Ga}(a, b/c)$.
  - If $\theta \sim \text{Ga}(a, b)$, then $1/\theta$ has an Inverse-Gamma distribution.

  We'll take advantage of these soon!

Example gamma distributions



R has the option to specify either the rate or scale parameter so always make sure to specify correctly when using "dgamma","rgamma",etc!.

Poisson-Gamma model

- Generally, it turns out that if

  - $f(y_i;\theta) : y_1, \ldots, y_n \overset{iid}{\sim} \text{Po}(\theta)$, and
  - $\pi(\theta) : \theta \sim \text{Ga}(a, b)$,

  then the posterior distribution is also a gamma distribution.

- Can we derive the posterior distribution and its parameters? Of course....let's do some work on the board!

- Updating a gamma prior with a Poisson likelihood leads to a gamma posterior - we once again have conjugacy!

- Specifically, we have.

$$\pi(\theta|\{y_i\}) : \theta|\{y_i\} \sim \text{Ga}(a + \sum y_i, b + n).$$

- What is the posterior mean? How about the posterior variance?

Hoff example: birth rates

- Survey data on educational attainment and number of children of 155 forty-year-old women during the 1990's.

- These women were in their 20s during the 1970s, a period of historically low fertility rates in the US.

- **Goal**: compare birth rate θ_1 for women with bachelor's degrees to the rate θ_2 for women without.

- **Data**:
  - 111 women without a bachelor's degree had 217 children: $(\bar{y}_1 = 1.95)$
  - 44 women with bachelor's degrees had 66 children: $(\bar{y}_2 = 1.50)$

- Based on the data alone, looks like θ_1 should be greater than θ_2. But...how sure are we?

- **Priors**: $\theta_1, \theta_2 \sim \text{Ga}(2, 1)$ (not much prior information; equivalent to 1 prior woman with 2 children). Posterior means will be close to the MLEs.

Hoff example: birth rates

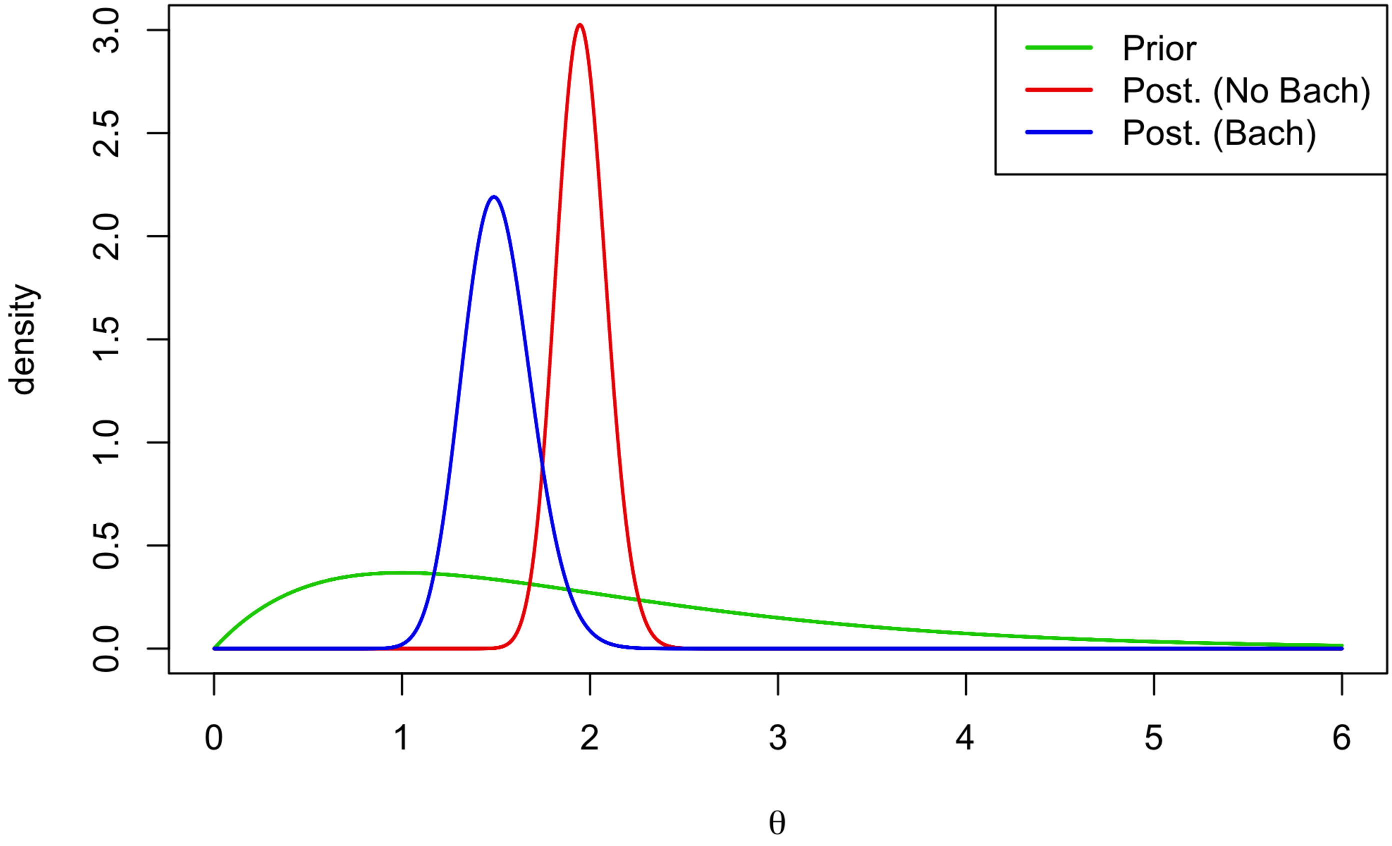
- Then,
  - $\theta_1 | \{n_1 = 111, \sum y_{i,1} = 217\} \sim \text{Ga}(2 + 217, 1 + 111) = \text{Ga}(219, 112)$.
  - $\theta_1 | \{n_1 = 44, \sum y_{i,1} = 66\} \sim \text{Ga}(2 + 66, 1 + 44) = \text{Ga}(68, 45)$.
- We can then use R to calculate posterior means, modes, and 95% credible intervals for θ_1 and θ_2.

```r
a=2; b=1; #prior
n1=111; sumy1=217; n2=44; sumy2=66 #data
(a+sumy1)/(b+n1); (a+sumy2)/(b+n2); #post means
qgamma(c(0.025, 0.975),a+sumy1,b+n1) #95\% ci 1
qgamma(c(0.025, 0.975),a+sumy2,b+n2) #95\% ci 2
```

- Posterior means: $\mathbb{E}[\theta_1 | \{y_{i,1}\}] = 1.955$ and $\mathbb{E}[\theta_2 | \{y_{i,2}\}] = 1.511$.
- 95% credible intervals
  - θ_1: [1.71, 2.22].
  - θ_2: [1.17, 1.89].

Hoff example: birth rates

Prior and posteriors:



Hoff example: birth rates

- Posteriors indicate considerable evidence birth rates are higher among women without bachelor's degrees.
- Confirms what we observed.
- Using sampling we can quickly calculate $\text{Pr}(\theta_1 > \theta_2 | \text{data})$.

  ```
  mean(rgamma(10000,219,112)>rgamma(10000,68,45))
  ```

  We have $\text{Pr}(\theta_1 > \theta_2 | \text{data}) = 0.97$.
- Why/how does it work?
- Monte Carlo approximation coming soon!
- Clearly, that probability will change with different priors.

Posterior predictive distribution

- What is the posterior predictive distribution for the Poisson-gamma model?
- Let $a_n = a + \sum y_i$ and $b_n = b + n$.
- We have

$$\begin{aligned} f(y_{n+1}|y_{1:n}) &= \int f(y_{n+1}|\theta)\pi(\theta|y_{1:n})\, d\theta \\ &= \int \mathrm{Po}(y_{n+1};\theta)\mathrm{Ga}(\theta; a_n, b_n)\, d\theta \\ &= \ldots \\ &= \ldots \\ &= \frac{\Gamma(a_n + y_{n+1})}{\Gamma(a_n)\Gamma(y_{n+1}+1)} \left(\frac{b_n}{b_n+1}\right)^{a_n} \left(\frac{1}{b_n+1}\right)^{y_{n+1}} \end{aligned}$$

which is the negative binomial distribution, $\text{Neg-binomial}\left(a_n, \dfrac{1}{b_n+1}\right)$.

- The prior predictive distribution takes a similar form.

Negative binomial distribution

- Originally derived as the number of successes in a sequence of independent $\text{Bernoulli}(p)$ trials before r failures occur.

- The negative binomial distribution $\text{Neg-binomial}\,(r, p)$ is parameterized by r and p and the pmf is given by

$$\Pr[Y = y|r, p] = \binom{y + r - 1}{y}(1 - p)^r p^y; \quad y = 0, 1, 2, \ldots; \quad p \in [0, 1].$$

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

$$\Pr[Y = y|r, p] = \frac{\Gamma(y + r)}{\Gamma(y + 1)\Gamma(r)}(1 - p)^r p^y; \quad y = 0, 1, 2, \ldots; \quad p \in [0, 1].$$

- Some properties:
  - $\mathbb{E}[\theta] = \dfrac{pr}{1 - p}$
  - $\mathbb{V}[\theta] = \dfrac{pr}{(1 - p)^2}$

Posterior predictive distribution

- The negative binomial distribution is an over-dispersed generalization of the Poisson.

- What does over-dispersion mean?

- In marginalizing θ out of the Poisson likelihood, over a gamma distribution, we obtain a negative-binomial.

- For $(y_{n+1}|y_{1:n}) \sim \text{Neg-binomial}\left(a_n, \dfrac{1}{b_n+1}\right)$, we have
  - $\mathbb{E}[y_{n+1}|y_{1:n}] = \dfrac{a_n}{b_n} = \mathbb{E}[\theta|y_{1:n}] =$ posterior mean, and
  - $\mathbb{V}[y_{n+1}|y_{1:n}] = \dfrac{a_n(b_n+1)}{b_n^2} = \mathbb{E}[\theta|y_{1:n}]\left(\dfrac{b_n+1}{b_n}\right)$,

  so that variance is larger than the mean by an amount determined by b_n, which takes the over-dispersion into account.

Predictive uncertainty

- Note that as the sample size n increases, the posterior density for θ becomes more and more concentrated.

$$\mathbb{E}[\theta|y_{1:n}] = \frac{a_n}{b_n^2} = \frac{a + \sum_i y_i}{(b+n)^2} \approx \frac{\bar{y}}{n} \to 0.$$

- Also, recall that $\mathbb{V}[y_{n+1}|y_{1:n}] = \mathbb{E}[\theta]\left(\frac{b_n+1}{b_n}\right)$.

- As we have less uncertainty about θ, the inflation factor

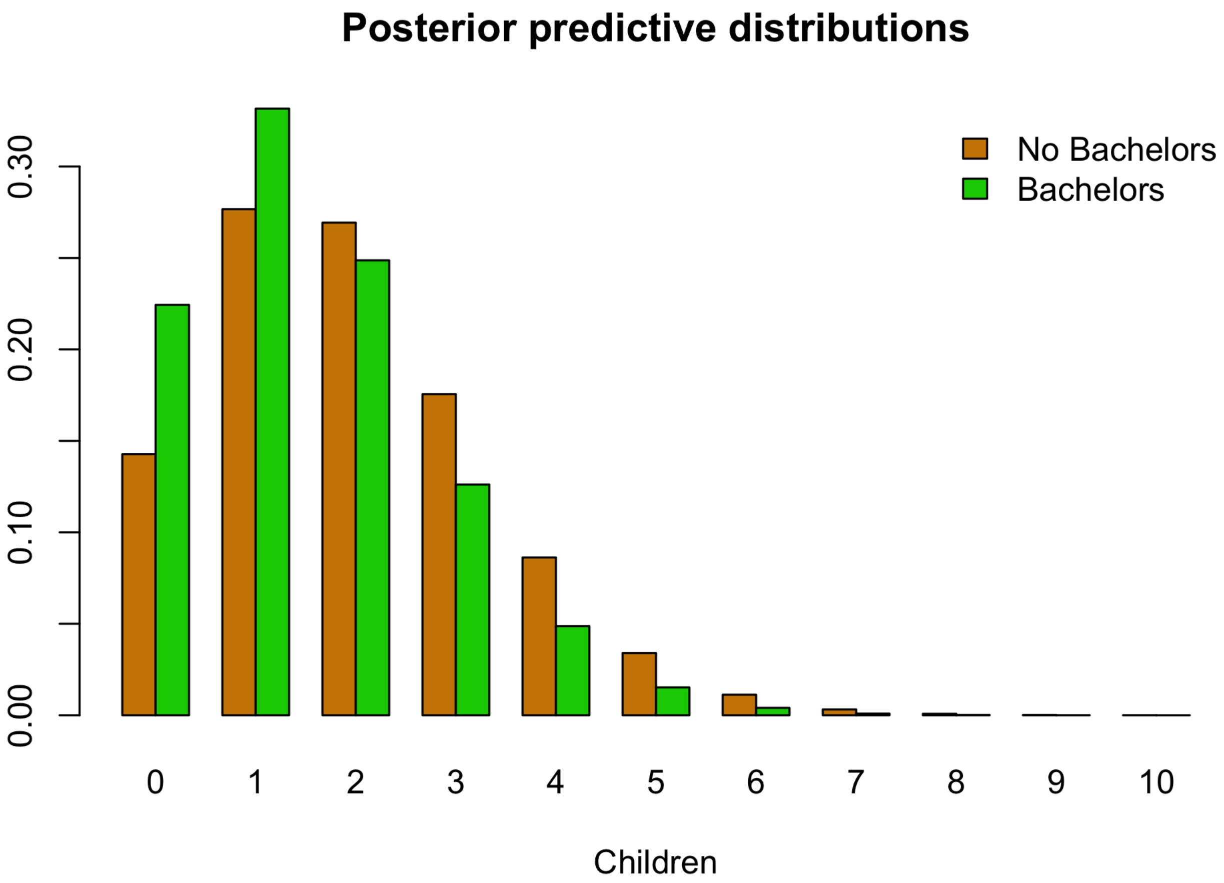
$$\frac{b_n+1}{b_n} = \frac{b+n+1}{b+n} \to 1$$

and the predictive density $f(y_{n+1}|y_{1:n}) \to \mathrm{Po}(\bar{y})$.

- Of course, in smaller samples, it is important to inflate our predictive intervals to account for uncertainty in θ.

Back to birth rates

- Let's compare the posterior predictive distributions for the two groups of women.



Back to birth rates

- Suppose we randomly sample two women, one with degree and one without. To what extent do we expect the one without the degree to have more kids than the other, e.g. $\tilde{y}_1 > \tilde{y}_2$?
- Using R, $\Pr(\tilde{y}_1 > \tilde{y}_2) \approx 0.48$ and $\Pr(\tilde{y}_1 > \tilde{y}_2) \approx 0.22$.

```r
set.seed(01222020)
a=2; b=1; #prior
n1=111; sumy1=217; n2=44; sumy2=66 #data
mean(rnbinom(100000,size=(a+sumy1),mu=(a+sumy1)/(b+n1)) >
rnbinom(10000,size=(a+sumy2),mu=(a+sumy2)/(b+n2)))
```

```
## [1] 0.48218
```

```r
mean(rnbinom(100000,size=(a+sumy1),mu=(a+sumy1)/(b+n1))==
rnbinom(10000,size=(a+sumy2),mu=(a+sumy2)/(b+n2)))
```

```
## [1] 0.21799
```

- Strong evidence of difference between two populations does not really imply the difference in predictions is large.

Poisson model in terms of rate

- In many applications, it is often convenient to parameterize the Poisson model a bit differently. One option takes the form

  $$y_1, \ldots, y_n \sim \mathrm{Po}(x_i \theta)$$

  where x_i represents an explanatory variable and θ is once again the population parameter of interest. The model is not exchangeable in the y_i 's but is exchangeable in the pairs $(x, y)_i$.

- In epidemiology, θ is often called the population "rate" and x_i is called the "exposure" of unit i.

- When dealing with mortality rates in different counties for example, x_i can be the population n_i in county i, with $\theta =$ the overall mortality rate.

- The gamma distribution is still conjugate for θ, with the resulting posterior taking the form

  $$\pi(\theta | \{x_i, y_i\}) : \theta | \{x_i, y_i\} \sim \mathrm{Ga}(a + \sum_i y_i, b + \sum_i x_i).$$

- We will look at an example in the next class.