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 $\mathrm{Bernoulli}(\hat{\theta})$, 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

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

- Hence, your bet is rational as long as

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

- 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.

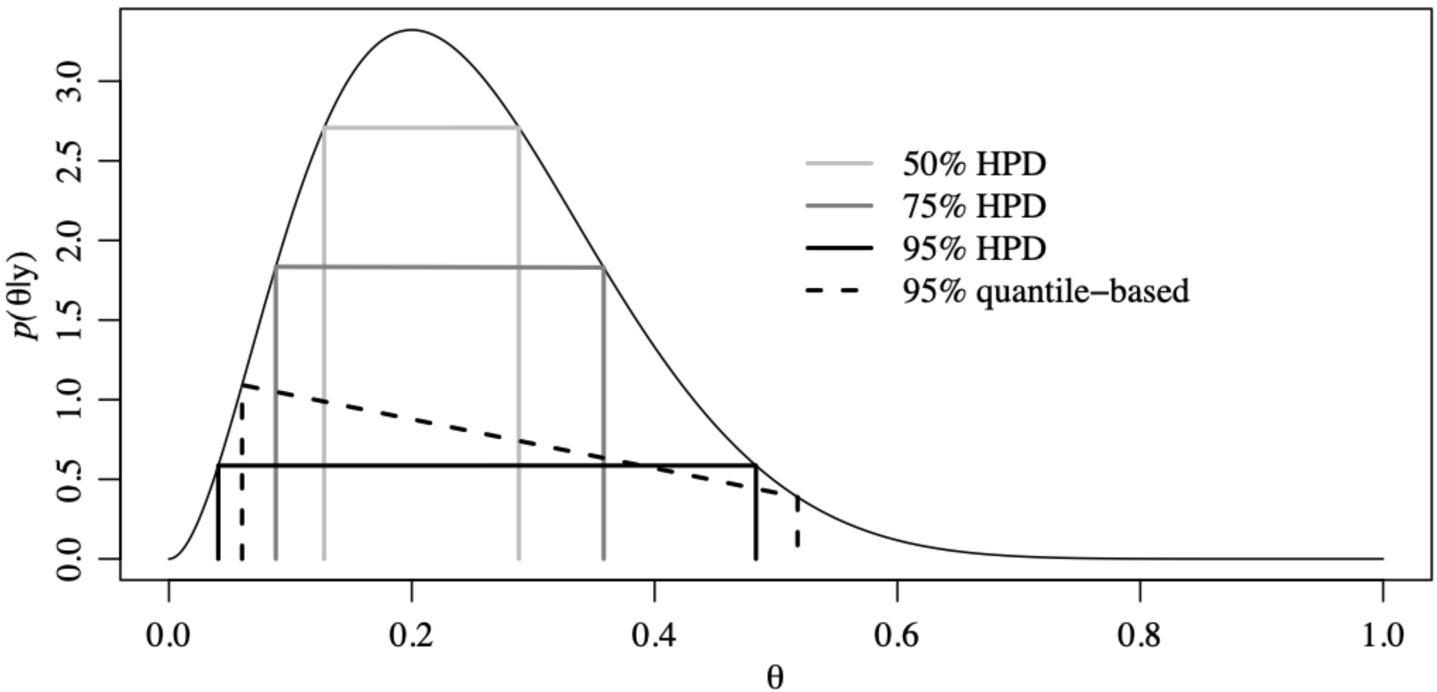
Bayesian quantile-based interval

- 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.

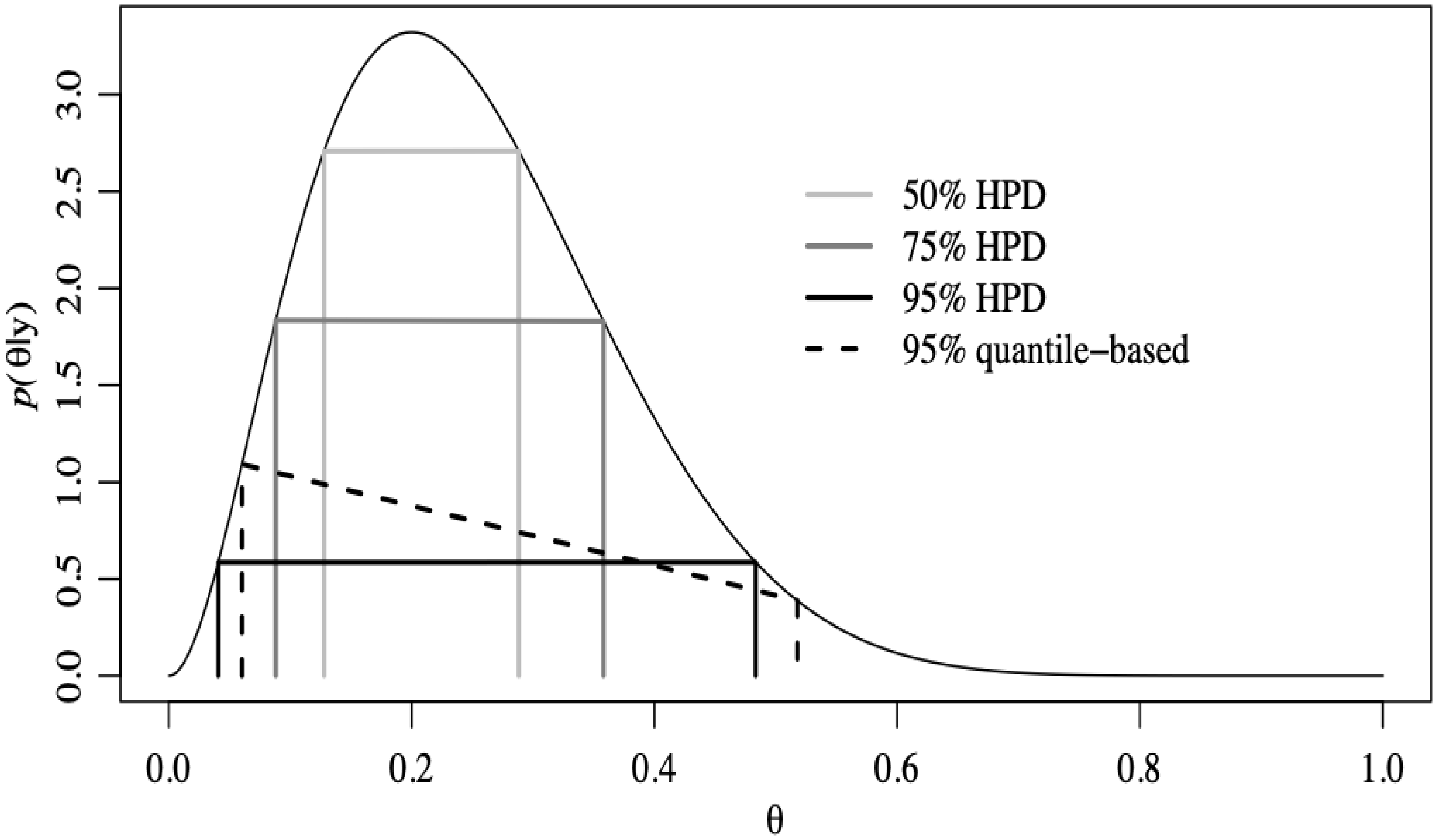
  Note this region is not necessarily a single interval (e.g., in the case of a multimodal posterior).

- 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

- $Y_1, \ldots, Y_n \overset{iid}{\sim} \text{Po}(\theta)$ 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

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

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

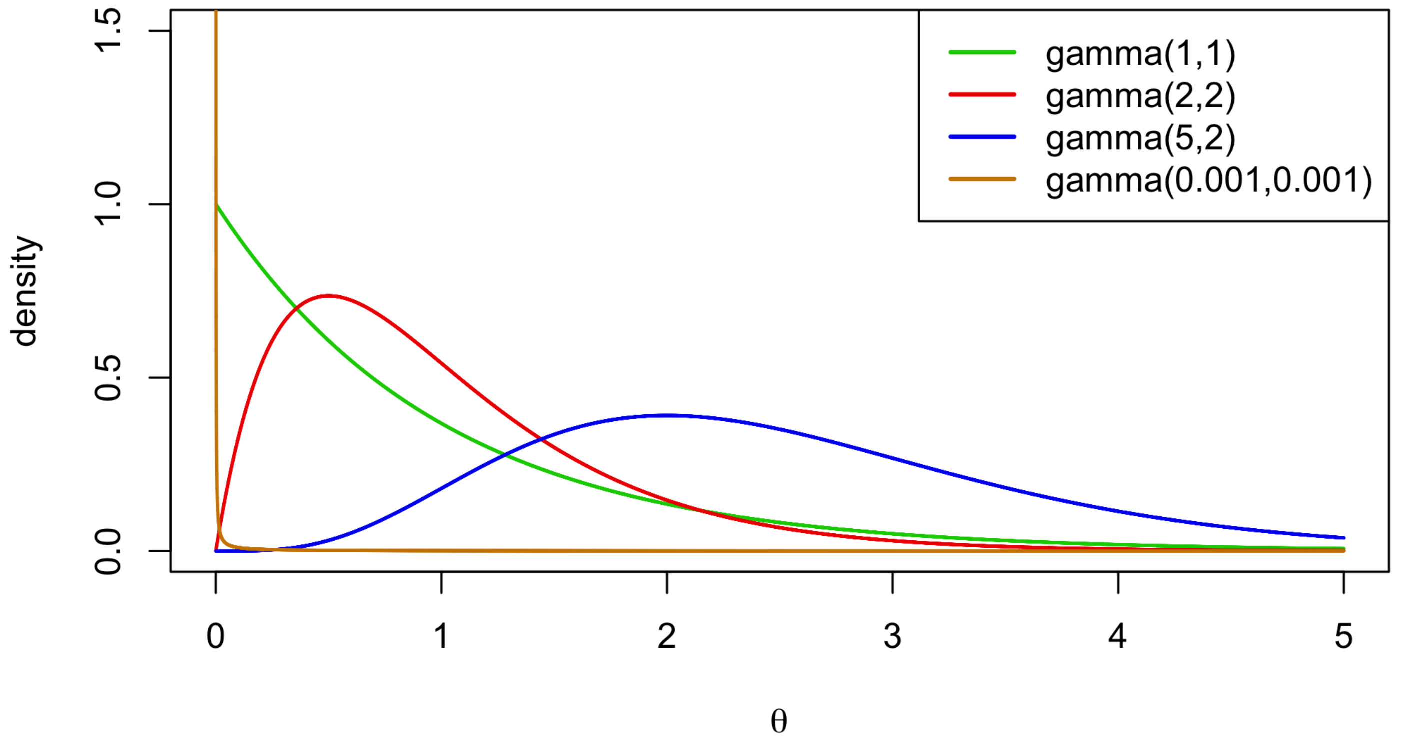
$$\mathbb{E}[\theta] = \mu = \frac{a}{b}; \ \ \mathbb{V}[\theta] = \mu\phi = \frac{a}{b^2},$$

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

- More properties:
  - If $\theta_1, \ldots, \theta_p \overset{ind}{\sim} \mathrm{Ga}(a_i, b)$, then $\sum_i \theta_i \sim \mathrm{Ga}(\sum_i a_i, b)$.
  - If $\theta \sim \mathrm{Ga}(a, b)$, then for any $c > 0$, $c\theta \sim \mathrm{Ga}(a, b/c)$.
  - If $\theta \sim \mathrm{Ga}(a, b)$, 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!.