One parameter models cont'd

Dr. Olanrewaju Michael Akande

Jan 17, 2020

Announcements

- Tentative dates for quizzes:
  - Quiz I: Wed, Feb 12
  - Quiz II: Wed, Apr 1
- Homework 1 now online, due Thursday, Jan 23.
- My office hours are confirmed:
  - Wed 9:00 - 10:00am and Thur 11:45 - 12:45pm
- No lab on Monday; MLK day.
- Occasional short (timed) "participation" quizzes from next week.
  - Will usually be available on Sakai under "Tests & Quizzes"
  - Must take at the beginning of class if present
  - Must take between 11:45am and 1:00pm if absent

Outline

- Beta-binomial (cont'd)
  - Example
  - Cautionary tale: parameters at the boundary
  - Marginal likelihood
  - Posterior prediction
  - Truncated priors
- Introduction to the Poisson-Gamma model
  - Recap of the distributions
  - Conjugacy

Beta-binomial cont'd

Beta-binomial recap

Binomial likelihood:

$$L(y;\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$$

$+$ Beta Prior:

$$\pi(\theta) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1} = \text{Beta}(a,b)$$

$\Rightarrow$ Beta posterior:

$$\pi(\theta|y) = \frac{1}{B(a+y,b+n-y)}\theta^{a+y-1}(1-\theta)^{b+n-y-1} = \text{Beta}(a+y,b+n-y).$$

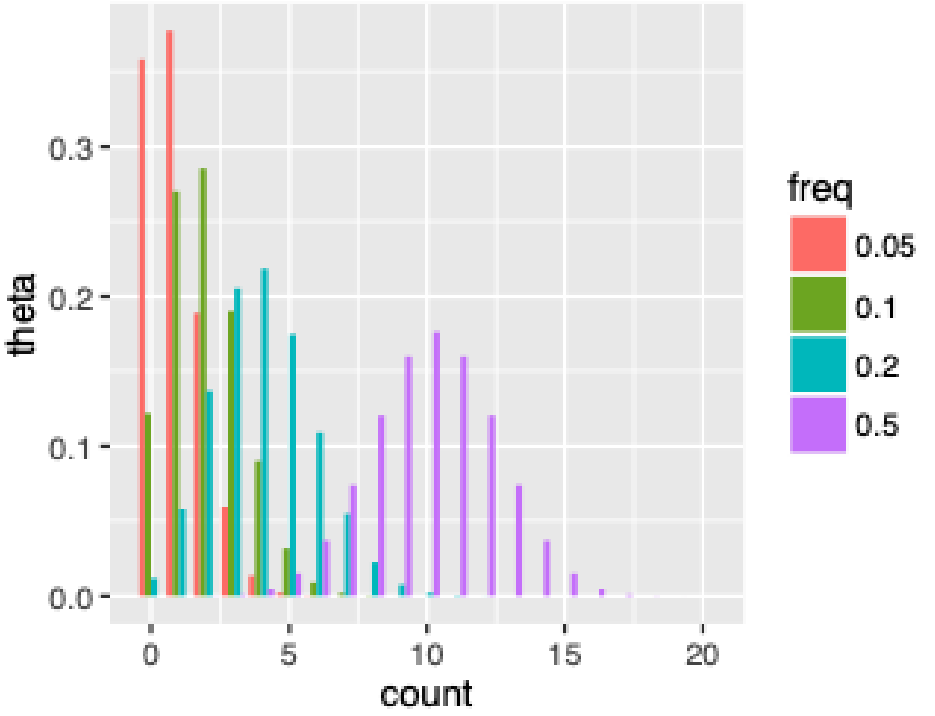
- Recall: for $\text{Beta}(a,b)$,
  - $\mathbb{E}[Y] = \dfrac{a}{a+b}$
  - $\mathbb{V}[Y] = \dfrac{ab}{(a+b)^2(a+b+1)}$

Example: rare events

- Step 1. State the question:
  - What is the prevalence of an infectious disease in a small city?
  - Why? High prevalence means more public health precautions are recommended.
- Step 2. Collect the data:
  - Suppose you collect a small random sample of 20 individuals.
- Step 3. Explore the data:
  - Count the number of infected individuals in the sample.
  - Let Y be the corresponding random variable.

Example: rare events

- Step 4. Formulate and state a modeling framework:
  - Parameter of interest: θ is the fraction of infected individuals in the city.
  - Sampling model: a reasonable model for Y can be $\text{Bin}(20, \theta)$



Example: rare events

- Step 4. Formulate and state a modeling framework:
  - Prior specification: information from previous studies – infection rate in "comparable cities" ranges from 0.05 to 0.20 with an average of 0.10. So maybe a standard deviation of roughly 0.05?
  - What is a good prior? The **expected value** of θ close to 0.10 and the **variance** close to 0.05.
  - Possible option: $\text{Beta}(3.5, 31.5)$ or maybe even $\text{Beta}(3, 32)$?

Example: rare events

- Step 4. Formulate and state a modeling framework:
  - Under $\text{Beta}(3, 32)$, $\Pr(\theta < 0.1) \approx 0.67$.
  - Posterior distribution for the model:
    $(\theta|Y = y) = \text{Beta}(a + y, b + n - y)$
  - Suppose $Y = 0$. Then, $(\theta|Y = y) = \text{Beta}(3, 32 + 20)$

Example: rare events

- Step 5. Check your models:
  - Compare performance of posterior mean and posterior probability that $\theta < 0.1$. See pages 5 to 7 of the Hoff book (the section on sensitivity analysis).
  - Under $\text{Beta}(3, 52)$,
    - $\Pr(\theta < 0.1 | Y = y) \approx 0.92$. More confidence in low values of θ.
    - For $\mathbb{E}(\theta | Y = y)$, we have

$$\mathbb{E}(\theta|y) = \frac{a+y}{a+b+n} = \frac{3}{52} = 0.058.$$

    - Recall that the prior mean is $a/(a+b) = 0.09$. Thus, we can see how that contributes to the prior mean.

$$\begin{aligned}\mathbb{E}(\theta|y) &= \frac{a+b}{a+b+n} \times \text{prior mean} + \frac{n}{a+b+n} \times \text{sample mean} \\ &= \frac{a+b}{a+b+n} \times \frac{a}{a+b} + \frac{n}{a+b+n} \times \frac{y}{n} \\ &= \frac{35}{52} \times \frac{3}{35} + \frac{20}{52} \times \frac{0}{n} = \frac{3}{52} = 0.058.\end{aligned}$$

Example: rare events

- Step 6. Answer the question:
  - People with low prior expectations are generally at least 90% certain that the infection rate is below 0.10. Again, see pages 5 to 7 of the Hoff book.
  - $\pi(\theta|Y)$ is to the left of $\pi(\theta)$ because the observation $Y = 0$ provides evidence of a low value of θ.
  - $\pi(\theta|Y)$ is more peaked than $\pi(\theta)$ because it combines information and so contains more information than $\pi(\theta)$ alone.
  - The posterior expectation is 0.058.
  - The posterior mode is 0.04.
    - Note, for $\text{Beta}(a, b)$, the mode is $\dfrac{a-1}{a+b-2}$.
  - The posterior probability that $\theta < 0.1$ is 0.92.

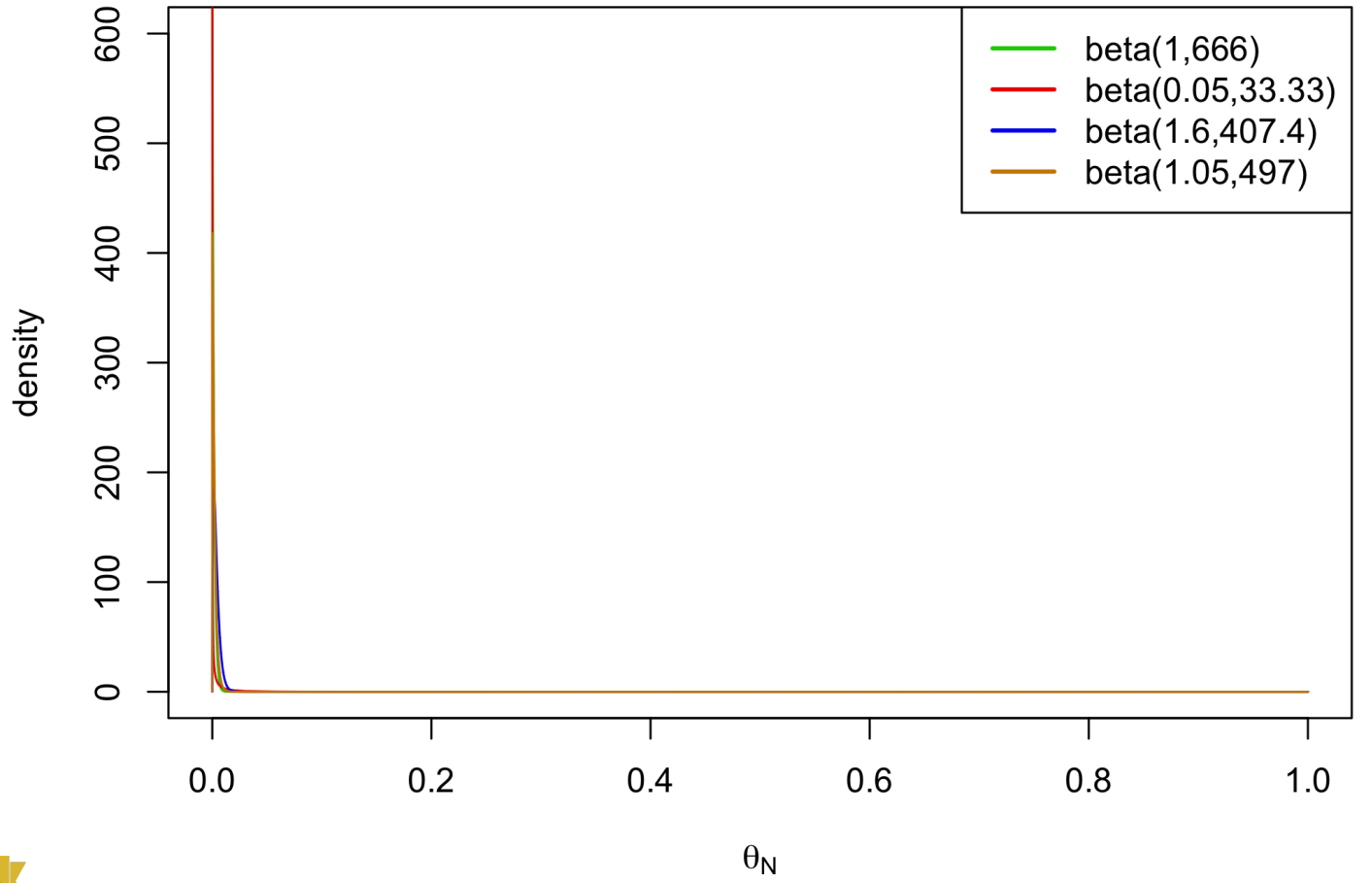
Cautionary tale: parameters at the boundary

- Tuyl et al. (2008) discuss potential dangers of using priors that have $a < 1$ with data that are all 0's (or $b < 1$ with all 1's). They consider data on adverse reactions to a new radiological contrast agent.
- Let θ_N: probability of a bad reaction using the new agent.
- Current standard agent causes bad reactions about 15 times in 10000, so one might think 0.0015 is a good guess for θ_N.
- How do we choose a prior distribution?

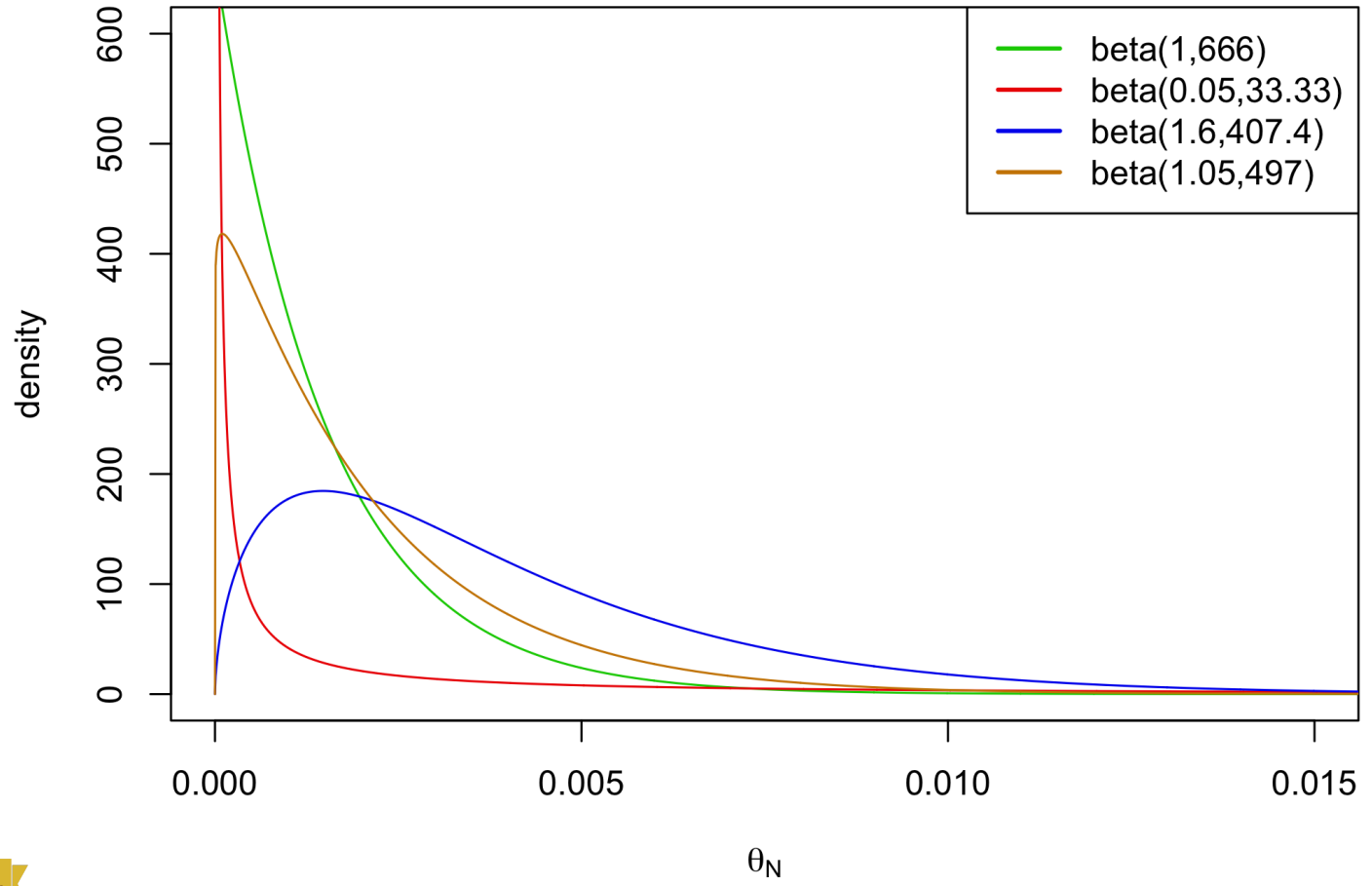
Potential prior distributions

- One might consider a variety of choices centered on $15/10000 = 0.0015$:
  - Prior 1: Beta(1,666) (mean 0.0015; 1 prior bad reaction in 667 administrations)
  - Prior 2: Beta(0.05,33.33) (mean 0.0015; 0.05 prior bad reactions in 33.38 administrations)
  - Prior 3: Beta(1.6, 407.4) (mode 0.0015; 409 prior administrations)
  - Prior 4: Beta(1.05, 497) (median 0.0015; 498.05 prior administrations)
- Does it matter which one we choose?

Potential prior distributions



Potential prior distributions



Potential prior distributions

- Let's take a closer look at properties of these four prior distributions, concentrating on the probability that $\theta_N < 0.0015$.
- That is, new agent not more dangerous than old agent.

	Be(1,666)	**Be(0.05,33.33)**	**Be(1.6,407.4)**	**Be(1.05,497)**
Prior prob	0.632	0.882	0.222	0.500
Post prob (0 events)	0.683	0.939	0.289	0.568
Post prob (1 event)	0.319	0.162	0.074	0.213

- Suppose we have a single arm study of 100 subjects.
- Consider the two most likely potential outcomes:
  - 0 adverse outcomes observed
  - 1 adverse outcome observed

Problems with the priors

- After just 100 trials with no bad reactions, the low weight (33.38 prior observations) prior indicates one should be 94% sure the new agent is equally safe as (or safer than) the old one.
- The low weight prior largely assumes the conclusion we actually hope for (that the new agent is safer), thus it takes very little confirmatory data to reach that conclusion.
- Is this the behavior we want?
- Take home message: be very careful with priors that have $a < 1$ with data that are all 0's (or $b < 1$ with all 1's).
- Given that we know the adverse event rate should be small, we might try a restricted prior e.g. Unif(0,0.1).
- In all cases, how many trials would we need, assuming no adverse reactions, to be 95% sure that the new agent is as safe as (or safer than) the old one? (Homework question!)

Marginal likelihood

- Recall that the marginal likelihood is

$$L(y) = \int L(y;\theta)p(\theta)\mathrm{d}\theta.$$

- What is the marginal likelihood for the Beta-binomial?
- We have

$$\begin{aligned} L(y) &= \int L(y;\theta)p(\theta)\mathrm{d}\theta \\ &= \int_0^1 \binom{n}{y}\theta^y(1-\theta)^{n-y}\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}d\theta \\ &= \binom{n}{y}\frac{B(a+y,\, b+n-y)}{B(a,b)}, \end{aligned}$$

  by the integral definition of the Beta function.

- Deriving the marginal likelihood for conjugate distributions is often relatively straightforward.

Posterior predictive distribution

- Let's go back to Bernoulli data. Suppose $y_i, \ldots, y_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$.
- We may wish to predict a new data point y_{n+1}.
- We can do so using the posterior predictive distribution $p(y_{n+1}|y_{1:n})$.
- Why are we not including the parameter in the posterior predictive distribution?
- Recall that we have conditional independence of the y's given θ.
- Generally,

$$\begin{aligned} p(y_{n+1}|y_{1:n}) &= \int p(y_{n+1}, \theta|y_{1:n})\, d\theta \\ &= \int p(y_{n+1}|\theta, y_{1:n}) p(\theta|y_{1:n})\, d\theta \\ &= \int p(y_{n+1}|\theta) p(\theta|y_{1:n})\, d\theta. \end{aligned}$$

Posterior predictive distribution

- When we talk about the posterior predictive distribution for Bernoulli data, we are really referring to $p(y_{n+1} = 1|y_{1:n})$ and $p(y_{n+1} = 0|y_{1:n})$.
- Then,

$$\begin{aligned} p(y_{n+1} = 1|y_{1:n}) &= \int p(y_{n+1} = 1|\theta) p(\theta|y_{1:n}) \, d\theta \\ &= \ldots \\ &= \ldots \end{aligned}$$

which simplifies to what? In class!

- What then is $p(y_{n+1} = 0|y_{1:n})$?
- Posterior predictive pmf therefore takes the form

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

- What are a_n and b_n?

Priors with restricted support

- As we have seen, when dealing with rare events, we might expect the true proportion to be very small.
- In that case, we might want to try a restricted prior, e.g. Unif(0,0.1).
- Even when we don't have rare events, we might still desire truncation if we are certain the true proportion lies within (a, b) with $0 < a < b < 1$.
- It is therefore often really useful to incorporate truncation.
- Let $\theta =$ probability of a randomly-selected student making an A in this course.
- You may want to rule out very low & very high values – perhaps $\theta \in [0.35, 0.6]$ with probability one.
- How to choose a prior restricted to this interval?

Uniform priors

- One possibility is to just choose a uniform prior.
- When the parameter θ is a probability, the typical uniform prior would correspond to beta(1,1).
- This is uniform on the entire (0,1) interval.
- However, we can just as easily choose a uniform prior on a narrower interval Unif(a,b) with $0 < a < b < 1$.
- Perhaps not flexible enough.
- Would be nice if we could pick a flexible beta density and then truncate it to (a, b).

Truncated random variables

- Suppose we have some arbitrary random variable $\theta \sim f$ with support Θ.
- For example, $\theta \sim \text{Beta}(c, d)$ has support on (0,1).
- Then, we can modify the density $f(\theta)$ to have support on a sub-interval $[a, b] \in \Theta$.
- The density $f(\theta)$ **truncated** to $[a, b]$ is

$$f_{[a,b]}(\theta) = \frac{f(\theta)1[\theta \in [a, b]]}{\int_a^b f(\theta^\star)\mathrm{d}\theta^\star},$$

with $1[A]$ being the indicator function that returns 1 if A is true & 0 otherwise.

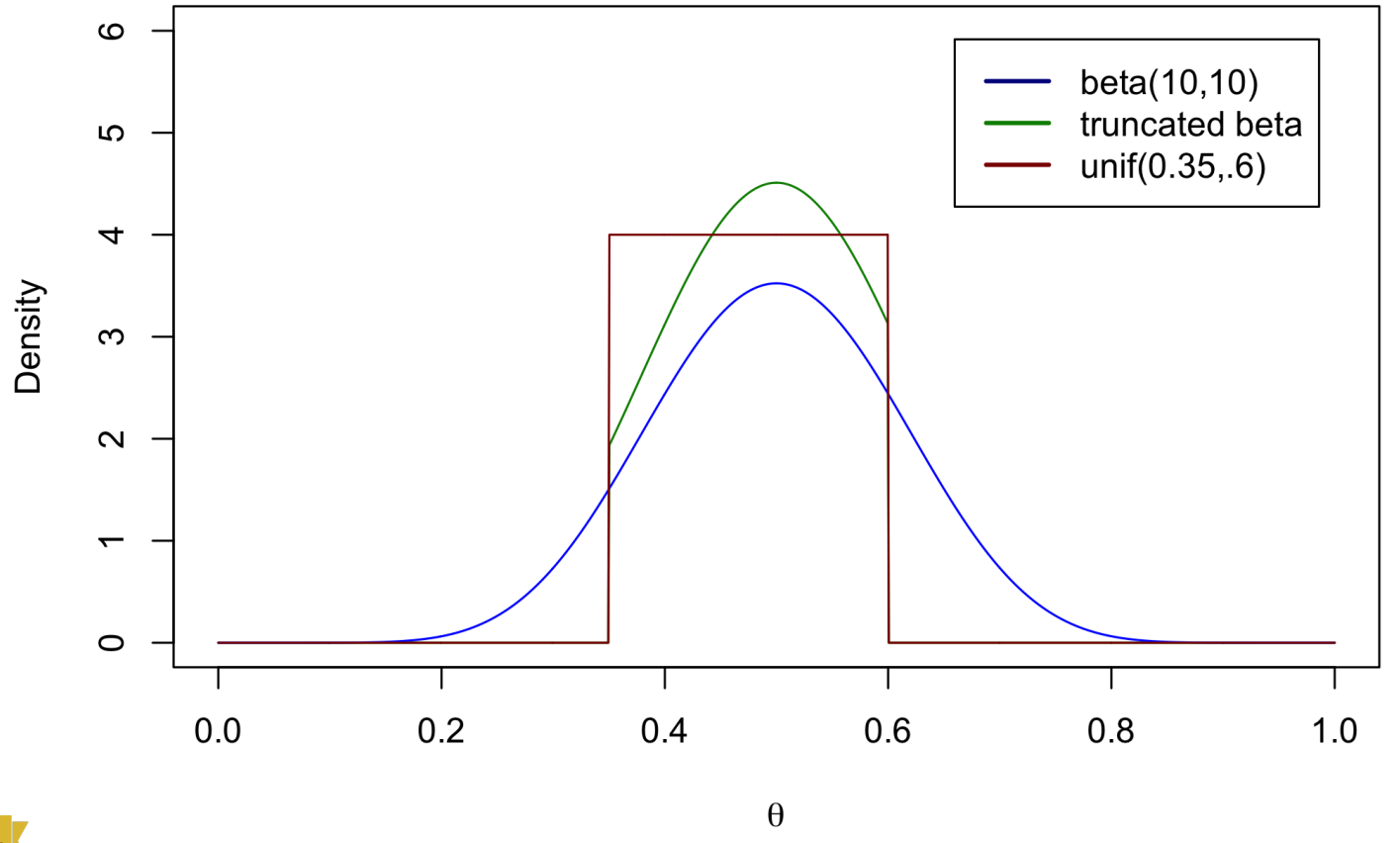
Truncated beta density

- Suppose to characterize the prior probability of earning an A, you poll a sample of students from a former STA 602 course and find that 10 earned an A and 10 earned a B (or lower).
- Therefore, you go with a beta(10,10) prior truncated to [0.35, 0.6].
- In R we can calculate the truncated beta density at p via

```
p <- seq(0,1,length=1000)
f1 <- dbeta(p,10,10)
f2 <- dbeta(p,10,10)*as.numeric(p>0.35 & p<0.6)/(pbeta(0.6,10,10) - pbeta(0.3,10,10))
f3 <- dunif(p,0.35,.6)
plot(p,f2,type='l',col='green4',xlim=c(0,1),ylab='Density', xlab=expression(theta),
  ylim=c(0,6))
lines(p,f1,type='l',col='blue')
lines(p,f3,type='l',col='red4')
labels <- c("beta(10,10)", "truncated beta","unif(0.35,.6)")
legend("topright", inset=.05, labels, lwd=2, lty=c(1,1,1), col=c('blue4','green4','red
```

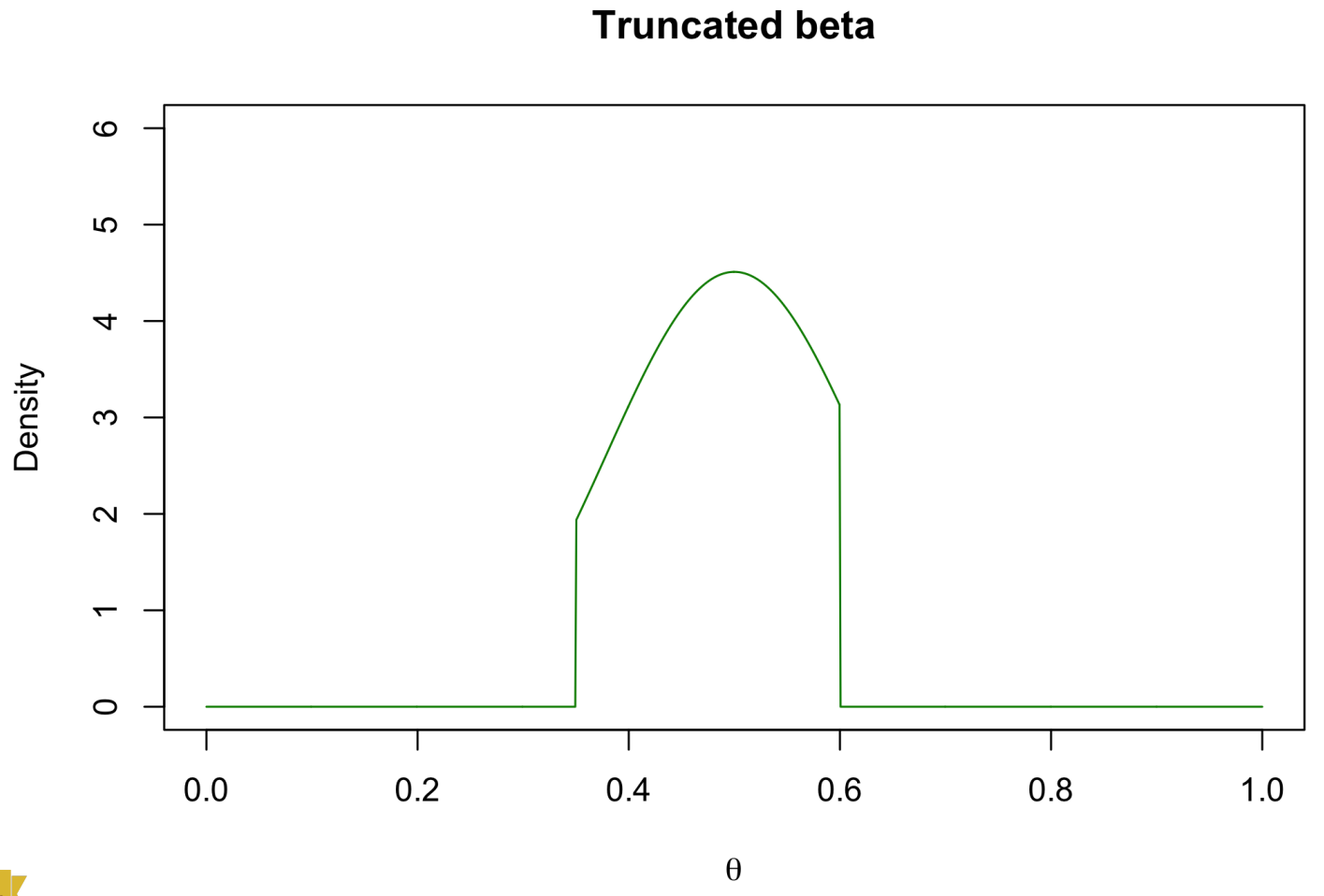
TRUNCATED BETA DENSITY

What would that look like?



Truncated beta density

The truncated density by itself would look like



The inverse cdf method

- How to sample truncated random variables?
- First start with the pdf for an untruncated distribution such as $\theta \sim \text{Beta}(c, d)$.
- Suppose we then want to sample $\theta \sim \text{Beta}_{[a,b]}(c, d)$. How can we do that? One popular method is the inverse-cdf method.
- The inverse cdf is useful for generating random variables in general, especially for generating truncated random variables.
- Suppose we have $\theta \sim f$, for some arbitrary continuous density f.
- According to probability integral transform, for any continuous random variable θ, the random variable $U = F(\theta)$ has a Unif(0,1) distribution. Note that F is the cdf.
- Thus, to use the inverse-cdf method to sample $\theta \sim f$, first sample $u \sim \text{Unif}(0, 1)$, then set $\theta = F^{-1}(u)$.

The inverse cdf method

- As an example, suppose we want to sample $\theta \sim \text{Beta}(c, d)$ through the inverse cdf method.
- Very easy. Just do the following in R.

```
u <- runif (1, 0, 1)
theta <- qbeta(u,c,d)
```

- That is, first sample from a uniform distribution.
- Then, transform it using the inverse cdf of the $\text{Beta}(c, d)$ distribution.
- Viola!

The inverse cdf method

- Back to the original problem: how to sample $\theta \sim \text{Beta}_{[a,b]}(c, d)$?
- If we had the inverse cdf of $\text{Beta}(c, d)$ truncated to $[a, b]$, then we could use the inverse cdf method. Easy enough! Let's find that inverse cdf.
- Let f, F and F^{-1} denote the pdf, cdf and inverse-cdf without truncation and let $A = [a, b]$.
- Recall that the density $f(\theta)$ **truncated** to $[a, b]$ is

$$f_A(\theta) = f_{[a,b]}(\theta) = \frac{f(\theta)1[\theta \in [a, b]]}{\int_a^b f(\theta^\star)\mathrm{d}\theta^\star} = \frac{f(\theta)1[\theta \in [a, b]]}{F(b) - F(a)}.$$

- Therefore, the truncated cdf

$$F_A(z) = \Pr[\theta \leq z] = \frac{F(z) - F(a)}{F(b) - F(a)}.$$

- Not enough though. We need the truncated inverse cdf.

The inverse cdf method

- To find the inverse cdf $F_A^{-1}(u)$, let $F_A(z) = u$. That is, set

$$u = F_A(z) = \frac{F(z) - F(a)}{F(b) - F(a)}$$

  and solve for z as a function of u.

- Re-expressing as a function of $F(z)$,

$$F(z) = \{F(b) - F(a)\}u + F(a).$$

- Applying the untruncated inverse cdf F^{-1} to both sides, we have

$$z = F^{-1}[\{F(b) - F(a)\}u + F(a)] = F_A^{-1}(u).$$

The inverse cdf method

- We now have all the pieces to use the inverse-cdf method to sample $\theta \sim f_A$, that is, f truncated to A.
- First draw a $\text{Unif}(0, 1)$ random variable

```
u <- runif (1, 0, 1)
```

- Next, apply the linear transformation:

$$u^\star = \{F(b) - F(a)\}u + F(a).$$

- Finally, plug $u^\star$ into the untruncated cdf $\theta = F^{-1}(u^\star)$.
- Note we can equivalently sample $u^\star \sim runif(1, F(a), F(b))$.

Intro to the Poisson-Gamma model

Poisson distribution recap

- $Y \sim \text{Po}(\theta)$ denotes that Y 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.
- The Poisson distribution is parameterized by θ and the pmf is given by

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

- Also,

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

- Suppose $y_i, \ldots, y_n \overset{iid}{\sim} \text{Po}(\theta)$.

What is the best guess (MLE) for the Poisson parameter?

Gamma density recap

- The gamma density will be useful as a prior for parameters that are strictly positive.
- If $\theta \sim \text{Ga}(a, b)$, we have the pdf

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

- Properties:

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

Poisson-Gamma model

- Generally, it turns out that if
  - $f(y_i;\theta) : y_i, \ldots, y_n \overset{iid}{\sim} \text{Po}(\theta)$, 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? 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.

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

- This is the Poisson-Gamma model. We will dive deeper with examples next time.