Poisson model (wrap-up); Monte Carlo approximation and sampling

Dr. Olanrewaju Michael Akande

Jan 24, 2020

Announcements

- HW1 due midnight
- HW2 now online, due next Thursday.
  - There should be six questions.
  - If you only see five questions, refresh your browser.

Outline

- Poisson model (wrap-up)
  - Example
  - Posterior prediction
  - Other parameterizations
- Finding conjugate distributions
- Monte Carlo approximation
- Sampling methods
  - Simple accept/reject
  - Importance sampling

Poisson model (wrap-up)

Poisson-Gamma recap

Poisson data:

$$f(y_i;\theta) : y_1, \ldots, y_n \overset{iid}{\sim} \text{Po}(\theta)$$

$+$ Gamma Prior:

$$\pi(\theta) = \frac{b^a}{\Gamma(a)}\theta^{a-1}e^{-b\theta} = \text{Ga}(a,b)$$

$\Rightarrow$ Gamma posterior:

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

- Recall: for $\text{Gamma}(a,b)$,
  - $\mathbb{E}[\theta] = \dfrac{a}{b}$
  - $\mathbb{V}[\theta] = \dfrac{a}{b^2}$
  - $\text{Mode}[\theta] = \dfrac{a-1}{b}$ for $a \geq 1$

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

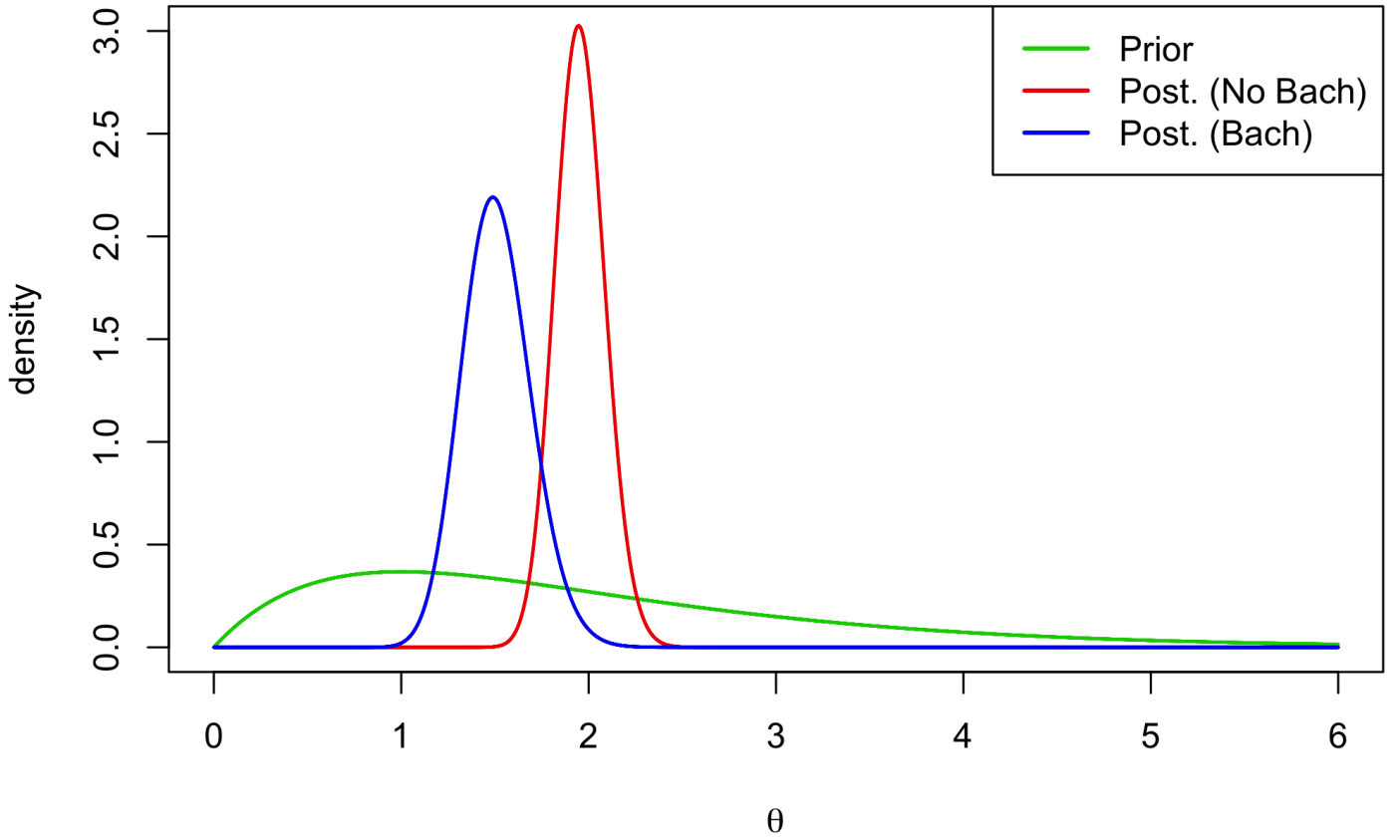
- Then,
  - $\theta_1 | \{n_1 = 111, \sum y_{i,1} = 217\} \sim \text{Ga}(2 + 217, 1 + 111) = \text{Ga}(219, 112)$.
  - $\theta_2 | \{n_2 = 44, \sum y_{i,2} = 66\} \sim \text{Ga}(2 + 66, 1 + 44) = \text{Ga}(68, 45)$.
- Use R to calculate posterior means and 95% CIs for θ_1 and θ_2.

```
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 $\Pr(\theta_1 > \theta_2 | \text{data})$.

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

  We have $\Pr(\theta_1 > \theta_2 | \text{data}) = 0.97$.

- 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

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

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

- The prior predictive distribution $f(y_{n+1}; a, b)$ 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 $(y_{n+1}|y_{1:n}) \sim \text{Neg-binomial}\left(a_n, \frac{1}{b_n+1}\right)$, we have
  - $\mathbb{E}[y_{n+1}|y_{1:n}] = \frac{a_n}{b_n} = \mathbb{E}[\theta|y_{1:n}] =$ posterior mean, and
  - $\mathbb{V}[y_{n+1}|y_{1:n}] = \frac{a_n(b_n+1)}{b_n^2} = \mathbb{E}[\theta|y_{1:n}]\left(\frac{b_n+1}{b_n}\right)$,

  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.

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

- 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

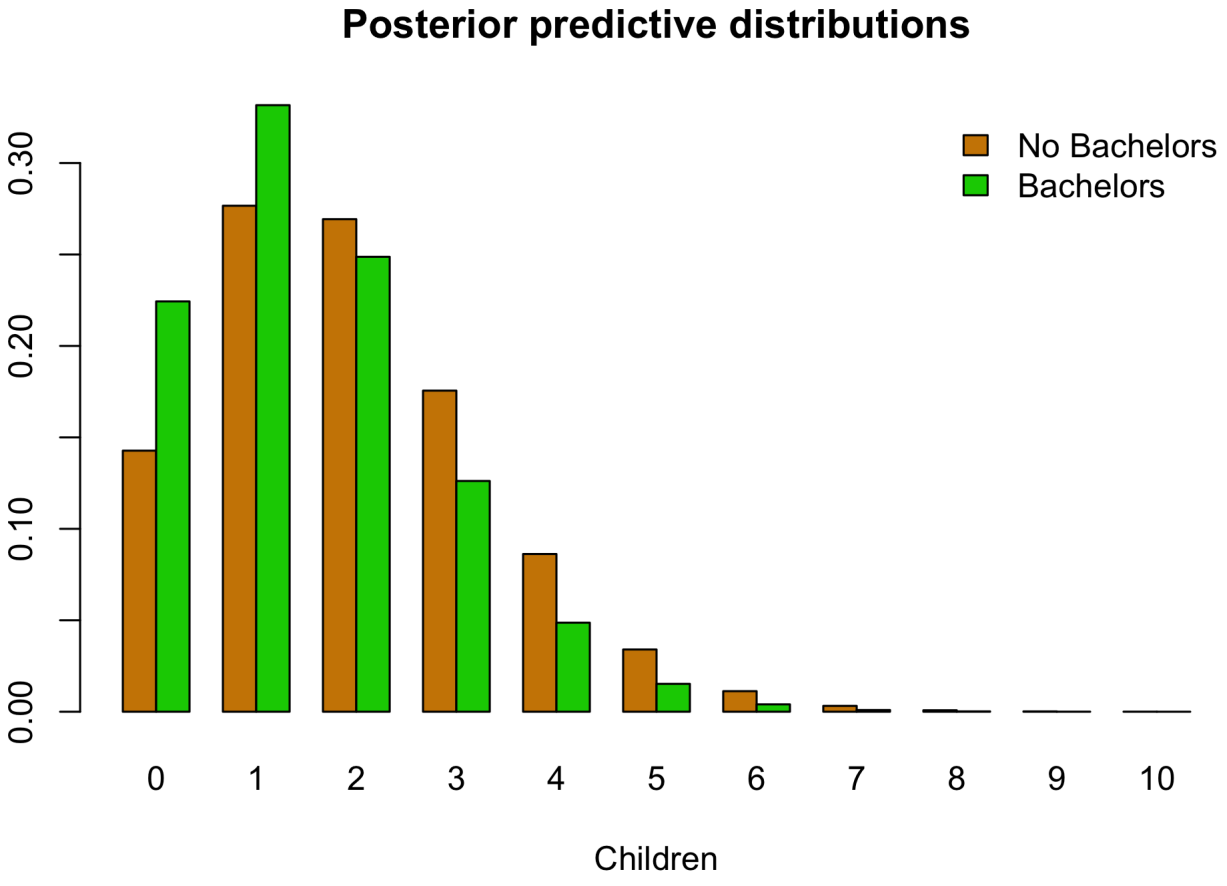
$$\frac{b_n + 1}{b_n} = \frac{b + n + 1}{b + n} \to 1$$

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

- 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.



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

$$y_i \sim \text{Po}(x_i\theta); \quad i = 1, \ldots, n.$$

  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

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

BDA EXAMPLE: ASTHMA MORTALITY RATE

- Consider an example on estimating asthma mortality rates for cities in the US.
- Since actual mortality rates can be small on the raw scale, they are often commonly estimated per 100,000 or even per one million.
- To keep it simple, let's use "per 100,000" for this example.
- For inference, ideally, we collect data which should basically count the number of asthma-related deaths per county.
- Note that inference is by county here, so county is indexes observations in the sample.
- Since we basically have count data, a Poisson model would be reasonable here.

Asthma mortality rate

- Since each city would be expected to have different populations, we might consider the sampling model:

$$y_i \sim \text{Po}(x_i\theta); \quad i = 1, \ldots, n.$$

where

  - x_i is the "exposure" for county i, that is, population of county i is $x_i \times 100,000$; and
  - θ is the unknown "true" city mortality rate per 100,000.

- Suppose

  - we pick a city in the US with population of 200,000;
  - we find that 3 people died of asthma, i.e., roughly 1.5 cases per 100,000.

- Thus, we have one single observation with $x_i = 2$ and $y_i = 3$ for this city.

Asthma mortality rate

- Next, we need to specify a prior. What is a sensible prior here?
- Perhaps we should look at mortality rates around the world or in similar countries.
- Suppose reviews of asthma mortality rates around the world suggest rates above 1.5 per 100,000 are very rare in Western countries, with typical rates around 0.6 per 100,000.
- Let's try a gamma distribution with $\mathbb{E}[\theta] = 0.6$ and $\Pr[\theta \geq 1.5]$ very low!
- A few options here, but let's go with $\text{Ga}(3, 5)$, which has $\mathbb{E}[\theta] = 0.6$ and $\Pr[\theta \geq 1.5] \approx 0.02$.
- Using trial-and error, explore more options in R!

Asthma mortality rate

- Therefore, our posterior takes the form

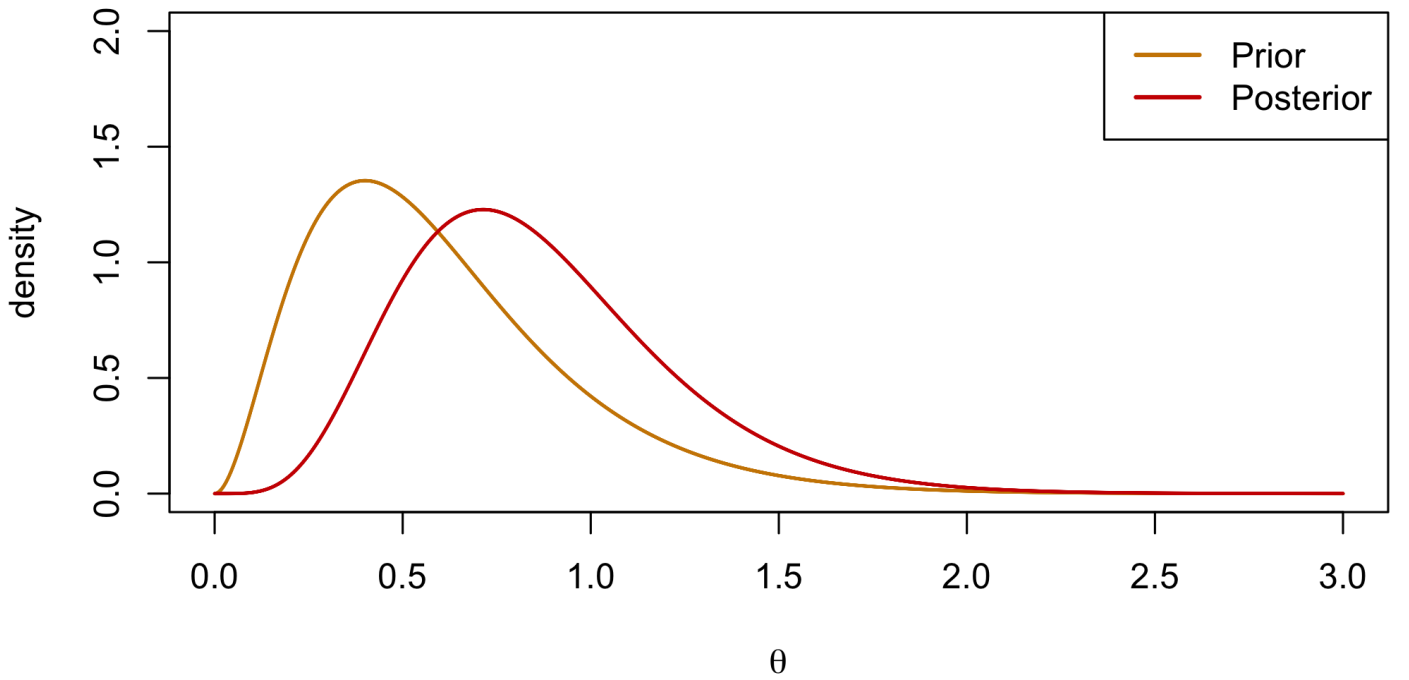
$$\pi(\theta|\{x_i, y_i\}) : \theta|\{x_i, y_i\} \sim \text{Ga}(a + \sum_i y_i, b + \sum_i x_i)$$

  which is actually

$$\pi(\theta|x, y) = \text{Ga}(a + y, b + x) = \text{Ga}(3 + 3, 5 + 2) = \text{Ga}(6, 7).$$

- $\mathbb{E}[\theta|x, y] = 6/7 = 0.86$ so that we expect less than 1 (0.86 to be exact) asthma-related deaths per 100,000 people in this city.
- In fact, the posterior probability that the long term death rate from asthma in this city is more than 1 per 100,000, $\Pr[\theta > 1|x, y]$, is 0.3.
- Also, $\Pr[\theta \geq 2|x, y] = 0.99$, so that there is very little chance that we see more than 2 asthma-related deaths per 100,000 people in this city.
- Use `pgamma` in R to compute the probabilities.

Prior vs posterior



Posterior is to the right of the prior since the data suggests higher mortality rates are more likely than the prior suggests. However, we only have one data point!

Finding conjugate distributions

Finding conjugate distributions

- In the conjugate examples we have looked at so far, how did we know the prior distributions we chose would result in conjugacy?
- Can we figure out the family of distributions that would be conjugate for arbitrary densities?
- Let's explore this using the exponential distribution. The exponential distribution is often used to model "waiting times" or other random variables (with support $(0, \infty)$) often measured on a time scale.
- If $y \sim \text{Exp}(\theta)$, we have the pdf

$$f(y) = \theta e^{-y\theta}; \quad y > 0.$$

  where θ is the rate parameter, and $\mathbb{E}[y] = 1/\theta$.
- Recall, if $Y \sim \text{Ga}(1, \theta)$, then $Y \sim \text{Exp}(\theta)$. What is $\mathbb{V}[y]$ then?
- Let's figure out what the conjugate prior for this density would look like (to be done in class).

Monte Carlo approximation

Monte Carlo approximation

- Monte Carlo integration is very key for Bayesian computation and using simulations in general.
- While we will focus on using Monte Carlo integration for Bayesian inference, the development is general and applies to any pdf/pmf $p(\theta)$.
- For our purposes, we will want to evaluate expectations of the form

$$H = \int h(\theta)p(\theta)d\theta,$$

  for many different functions $h(.)$ (usually scalar for us).

- Procedure:

  1. Generate a random sample $\theta_1, \ldots, \theta_m \overset{ind}{\sim} p(\theta)$.
  2. Estimate H using

$$\bar{h} = \frac{1}{m}\sum_{i=1}^{m} h(\theta_i).$$

Monte Carlo approximation

- We have $\mathbb{E}[h(\theta_i)] = H$.
- Assuming $\mathbb{E}[h^2(\theta_i)] < \infty$, so that the variance of each $h(\theta_i)$ is finite, we have

  1. LLN: $\bar{h} \overset{a.s.}{\to} H$.
  2. CLT: $\bar{h} - H$ is is asymptotically normal, with asymptotic variance

     $$\frac{1}{m} \int (h(\theta) - H)^2 p(\theta) d\theta,$$

     which can be approximated by

     $$v_m = \frac{1}{m^2} \sum_{i=1}^{m} (h(\theta_i) - \bar{h})^2.$$

- $\sqrt{v_m}$ is often called the Monte Carlo standard error.

Monte Carlo approximation

- That is, generally, taking large Monte Carlo sample sizes m (in the thousands or tens of thousands) can yield very precise, and cheaply computed, numerical approximations to mathematically difficult integrals.
- What this means for us: we can approximate just about any aspect of the posterior distribution with a large enough Monte Carlo sample.
- For samples $\theta_1, \ldots, \theta_m$ drawn iid from $p(\theta|y)$, as $m \to \infty$, we have
  - $\bar{\theta} = \frac{1}{m}\sum_{i=1}^{m}\theta_i \to \mathbb{E}[\theta|y]$
  - $\hat{\sigma}_\theta = \frac{1}{m-1}\sum_{i=1}^{m}(\theta_i - \bar{\theta})^2 \to \mathbb{V}[\theta|y]$
  - $\frac{1}{m}\sum_{i=1}^{m}1[\theta_i \leq c] = \frac{\#\theta_i \leq c}{m} \to \Pr[\theta \leq c|y]$
  - $[\frac{\alpha}{2}\text{th percentile of } (\theta_1, \ldots, \theta_m), (1-\frac{\alpha}{2})\text{th percentile of } (\theta_1, \ldots, \theta_m)]$ $\to 100 \times (1-\alpha)$ quantile-based credible interval.

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,

```
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
```

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

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

- That is, $\Pr(\tilde{y}_1 > \tilde{y}_2) \approx 0.48$ and $\Pr(\tilde{y}_1 > \tilde{y}_2) \approx 0.22$.
- Strong evidence of difference between two populations does not really imply the difference in predictions is large.

Monte Carlo approximation

- This general idea of using samples to "approximate" averages (expectations) is also useful when trying to approximate posterior predictive distributions.
- Quite often, we are able to sample from $f(y_i;\theta)$ and $\pi(\theta|\{y_i\})$ but not from $f(y_{n+1}|y_{1:n})$ directly.
- We can do so indirectly using the following Monte Carlo procedure:

$$\begin{aligned}
&\text{sample } \theta^{(1)} \sim \pi(\theta|\{y_i\}), \quad \text{then sample } y_{n+1}^{(1)} \sim f(y_{n+1};\theta^{(1)}) \\
&\text{sample } \theta^{(2)} \sim \pi(\theta|\{y_i\}), \quad \text{then sample } y_{n+1}^{(2)} \sim f(y_{n+1};\theta^{(2)}) \\
&\qquad\vdots \qquad\qquad\qquad\qquad\qquad\qquad\vdots \\
&\text{sample } \theta^{(m)} \sim \pi(\theta|\{y_i\}), \quad \text{then sample } y_{n+1}^{(m)} \sim f(y_{n+1};\theta^{(m)}).
\end{aligned}$$

- The sequence $\{(\theta, y_{n+1})^{(1)}, \ldots, (\theta, y_{n+1})^{(m)}\}$ constitutes m independent samples from the joint posterior of (θ, Y_{n+1}).
- In fact, $\{y_{n+1}^{(1)}, \ldots, y_{n+1}^{(m)}\}$ are independent draws from the posterior predictive distribution we care about.

Sampling methods

Rejection sampling (simple accept/reject)

- Setup:
  - $p(\theta)$ is some density we are interested in sampling from;
  - $p(\theta)$ is tough to sample from but we are able to evaluate $p(\theta)$ as a function at any point; and
  - $g(\theta)$ is some proposal distribution or importance sampling distribution that is easier to sample from.
- Two key requirements:
  - $g(\theta)$ is easy to sample from, and
  - $g(\theta)$ is easy to evaluate at any point as for $p(\theta)$.
- Usually, the context is one in which $g(\theta)$ has been derived as an analytic approximation to $p(\theta)$; and the closer the approximation, the more accurate the resulting Monte Carlo analysis will be.

Rejection sampling (simple accept/reject)

- Procedure:

  1. Define $w(\theta) = p(\theta)/g(\theta)$.

  2. Assume that $w(\theta) = p(\theta)/g(\theta) < M$ for some constant M. If $g(\theta)$ represents a good approximation to $p(\theta)$, then M should not be too far from 1.

  3. Generate a *candidate* value $\theta \sim g(\theta)$ and **accept** with probability $w(\theta)/M$: if accepted, θ is a draw from $p(\theta)$; otherwise **reject** and try again.

     Equivalently, generate $u \sim U(0, 1)$ independently of θ. Then **accept** θ as a draw from $p(\theta)$ if, and only if, $u < w(\theta)/M$.

- For those interested, the proof that all accepted θ values are indeed from $p(\theta)$ is on the last slide.

- **Drawback**: we need M for this to work. However, in the case of truncated densities, we actually have M.

Rejection sampling for truncated densities

- The inverse CDF method works well for truncated densities but what happens when we can not write down the truncated CDF?
- Suppose we want to sample from $f_{[a,b]}(\theta)$, that is, a known pdf $f(\theta)$ truncated to $[a, b]$.
  - Recall that $f_{[a,b]}(\theta) \propto f(\theta)1[\theta \in [a, b]]$. Using the notation for rejection sampling, $p(\theta) = f_{[a,b]}(\theta)$ and $g(\theta) = f(\theta)$.
  - Set $1/M = \int_a^b f(\theta^\star)\mathrm{d}\theta^\star$, so that M is a constant. Then, $w(\theta) = p(\theta)/g(\theta) = M1[\theta \in [a, b]] \leq M$ as required.
- We can then use the procedure on the previous page to generate the required samples. Specifically,
  - For each $i = 1, \ldots, m$, generate $\theta_i \sim f$. If $\theta_i \in [a, b]$, accept θ_i, otherwise **reject** and try again.
  - Easy to show that this is equivalent to accepting each θ_i with probability $w(\theta)/M$.

Example

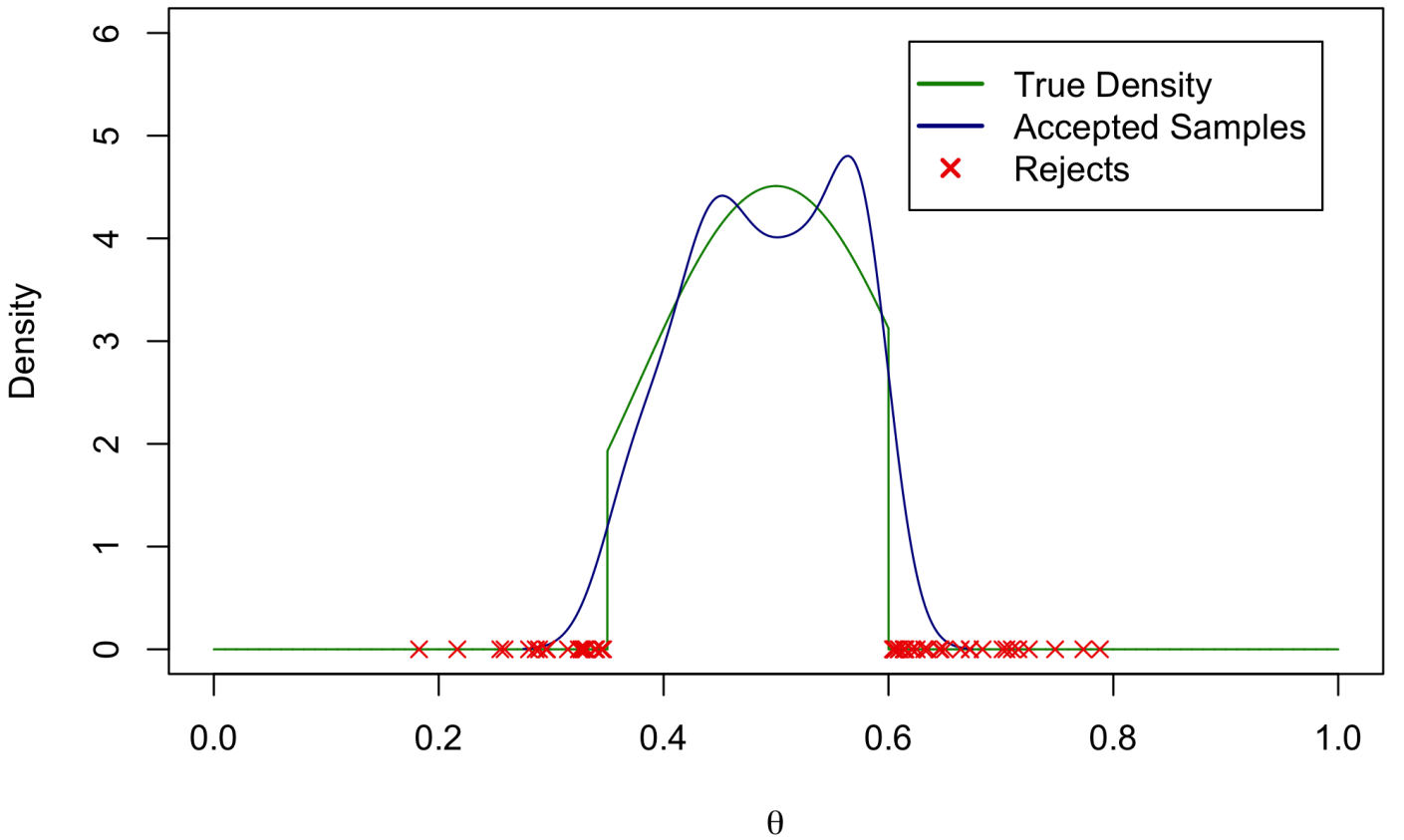
```
#Simple code for using rejection sampling to generate m samples
#from the Beta[10,10] density truncated to (0.35,0.6).
set.seed(12345)
#NOTE: there are more efficient ways to write this code!

#set sample size and reate vector to store sample
m <- 10000; THETA <- rep(0,m)
#keep track of rejects
TotalRejects <- 0; Rejections <- NULL
#now the 'for loop'
for(i in 1:m){
  t <- 0
  while(t < 1){
    theta <- rbeta(1,10,10)
    if(theta > 0.35 & theta < 0.6){
      THETA[i] <- theta
      t <- 1
    } else {
    TotalRejects <- TotalRejects + 1
    Rejections <- rbind(Rejections,theta)
  }
}
}
#How many rejections in all, to generate m=10000 samples?
TotalRejects
```

```
## [1] 3740
```

Clearly less efficient than inverse-CDF method which we already know we can use for this exercise. Acceptance rate ≈ 0.726.

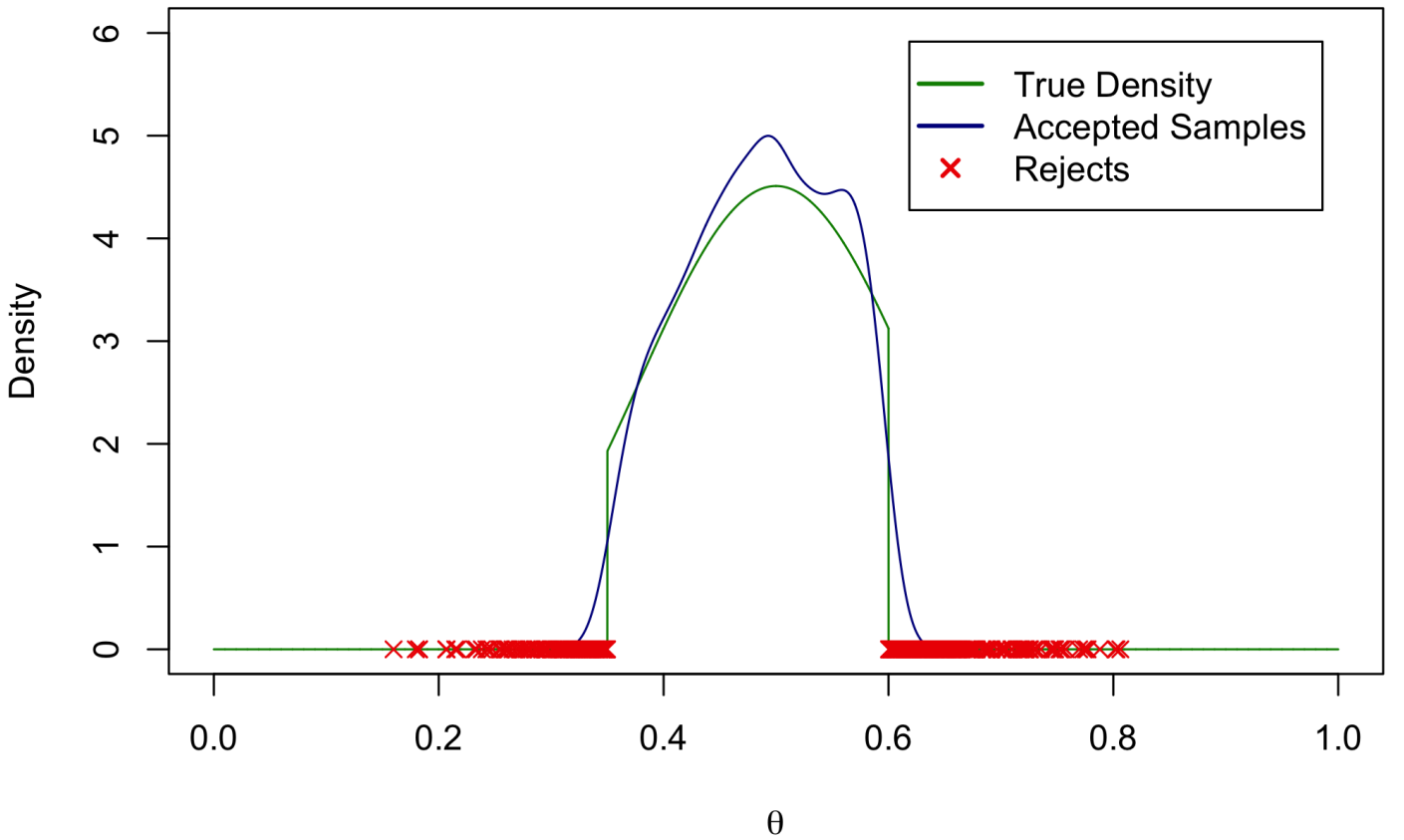
Example

How does our sample compare to the true truncated density? $m = 100$



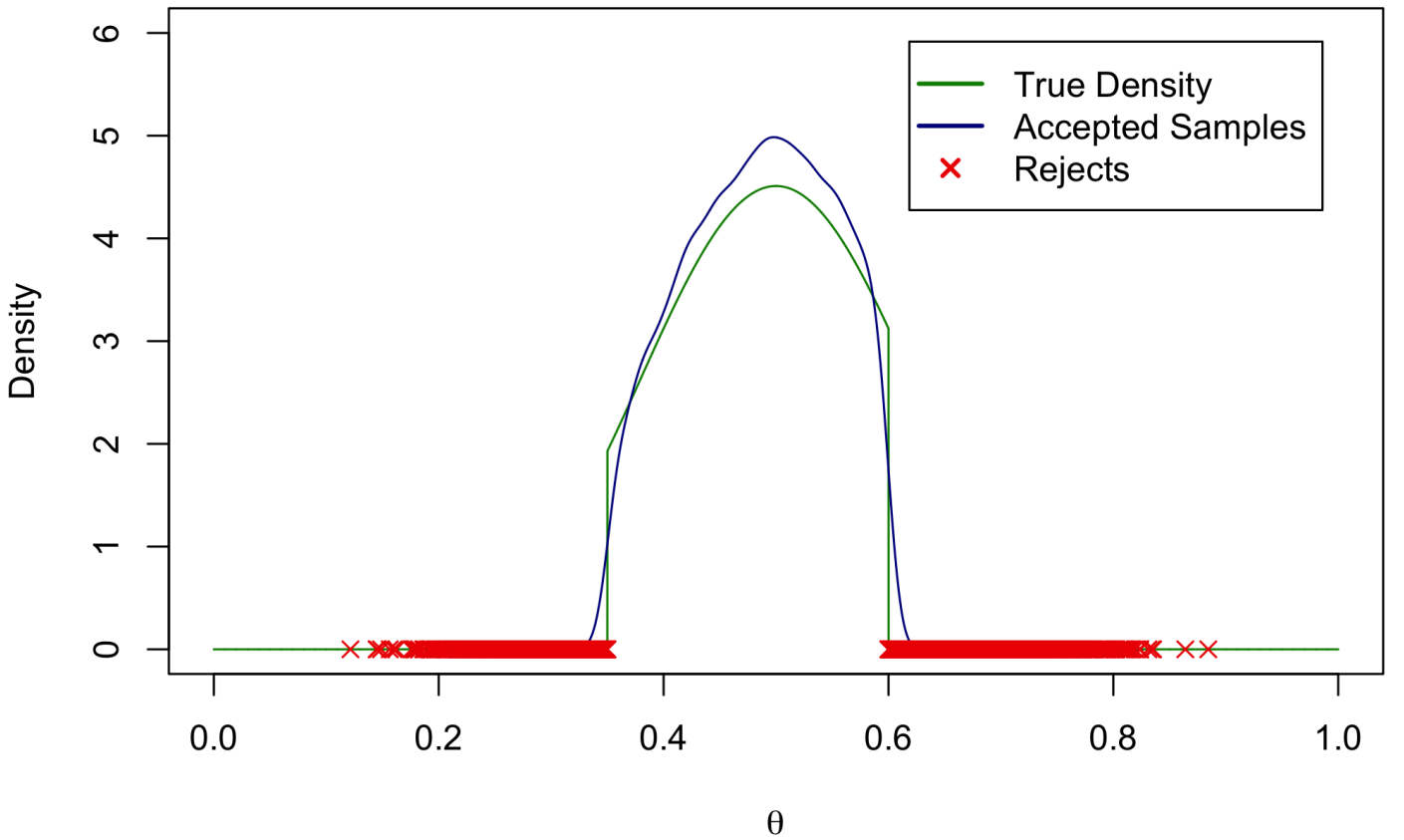
Example

How does our sample compare to the true truncated density? $m = 1000$



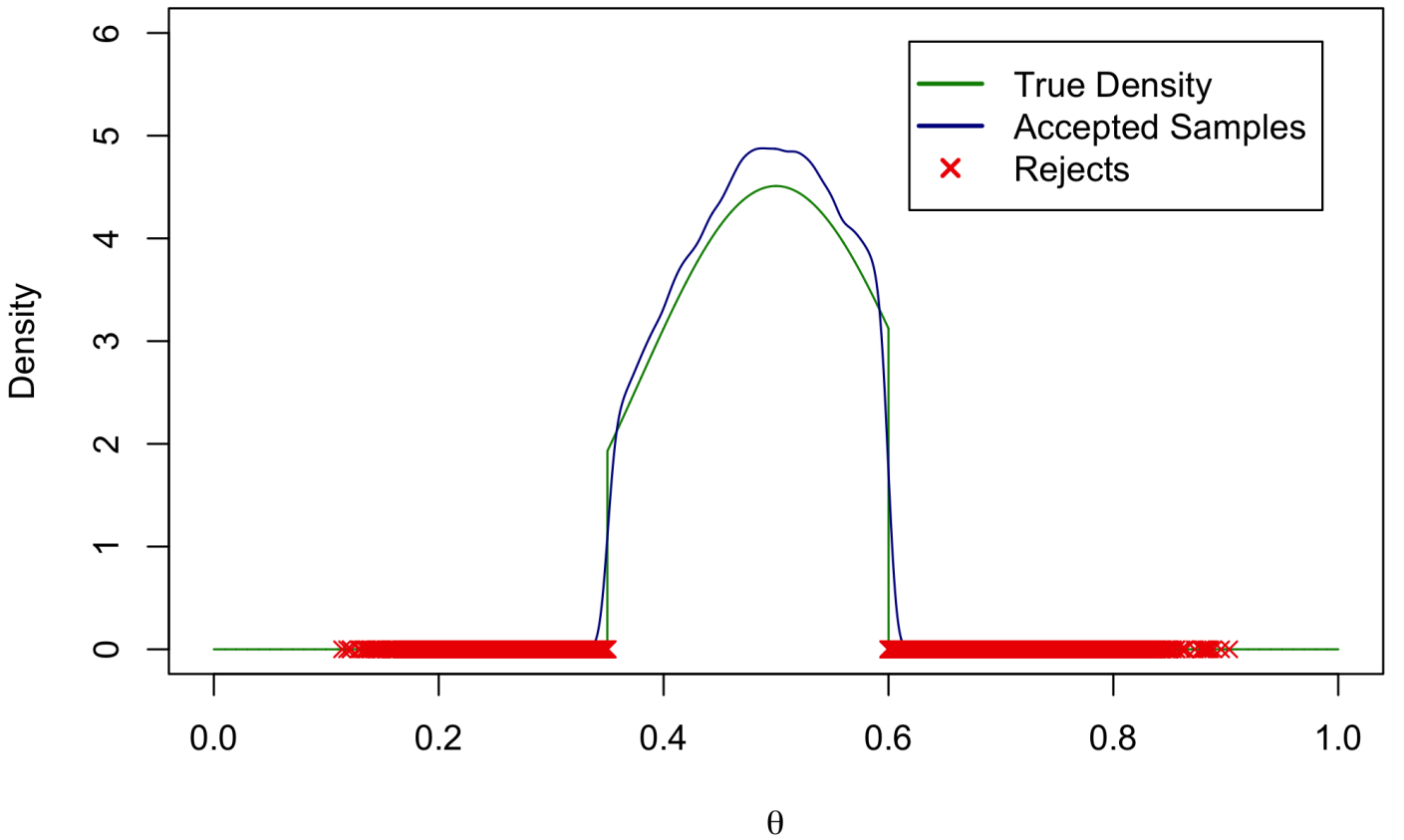
Example

How does our sample compare to the true truncated density? $m = 10000$



Example

How does our sample compare to the true truncated density? $m = 100000$



Importance sampling

- Importance sampling is actually one of the first steps into Monte Carlo analysis, in which simulated values from one distribution are used to explore another.
- Simulation from the "wrong distribution" can be incredibly useful as we have seen with rejection sampling and will also see later in this course.
- Not used as often anymore but still of practical interest in
  - fairly small problems, in terms of dimension,
  - in which the density of the distribution of interest can be easily evaluated, but when it is difficult to sample from directly, and
  - when it is relatively easy to identify and simulate from distributions that approximate the distribution of interest.
- Importance sampling and Rejection sampling use the same importance ratio ideas, but the latter leads to exact corrections and so exact samples from $p(\theta)$.

Importance sampling

- Interest lies in expectations of the form (instead of the actual samples)

$$H = \int h(\theta)p(\theta)d\theta,$$

- Write

$$H = \int h(\theta)w(\theta)g(\theta)d\theta \quad \text{with} \quad w(\theta) = p(\theta)/g(\theta)$$

that is, $\mathbb{E}[h(\theta)]$ under $p(\theta)$ is just $\mathbb{E}[h(\theta)w(\theta)]$ under $g(\theta)$.

- Using direct Monte Carlo integration

$$\bar{h} = \frac{1}{m}\sum_{i=1}^{m} w(\theta_i)h(\theta_i).$$

where $\theta_1, \ldots, \theta_m \overset{ind}{\sim} g(\theta)$. We are sampling from the "wrong" distribution.

Importance sampling

- The measure of "how wrong" we are at each simulated θ_m value is the importance weight

$$w(\theta_i) = p(\theta_i)/g(\theta_i).$$

  These ratios weight the sample estimates $h(\theta_i)$ to "correct" for the fact that we sampled the wrong distribution.

- See Lopes & Gamerman (Ch 3.4) and Robert and Casella (Ch. 3.3) for discussion of convergence and optimality.

- Clearly, the closer g is to p, the better the results, just as we had with rejection sampling.

Importance sampling

- Key considerations:
  - MC estimate $\bar{h}$ has the expectation H; and is generally almost surely convergent to H (under certain conditions of course but we will not dive into those).
  - $\mathbb{V}[\bar{h}]$ is often going to be finite in cases in which, generally, $w(\theta) = p(\theta)/g(\theta)$ is bounded and decays rapidly in the tails of $p(\theta)$.
  - Thus, superior MC approximations, are achieved for choices of $g(\theta)$ whose tails dominate those of the target $p(\theta)$.
  - That is, importance sampling distributions should be chosen to have tails at least as fat as the target (think normal distribution vs t-distribution).
  - Obviously require the support of $g(\theta)$ to be the same as, or contain, that of $p(\theta)$.
- These also clearly apply to rejection sampling too.

Importance sampling

- Problems in which $w(\theta) = p(\theta)/g(\theta)$ can be computed are actually rare.
- As you will see when we move away from conjugate distributions, we usually only know $p(\theta)$ up to a normalizing constant.
- When this is the case, simply "re-normalize" the importance weights, so that

$$\bar{h} = \frac{1}{m}\sum_{i=1}^{m} w_i h(\theta_i) \quad \text{where} \quad w_i = \frac{w(\theta_i)}{\sum_{i=1}^{m} w(\theta_i)}.$$

- Generally, in importance sampling, weights that are close to uniform are desirable, and very unevenly distributed weights are not.

Proof for simple accept/reject

- We need to show that all accepted θ values are indeed from $p(\theta)$. Equivalently, show that $f(\theta|u < w(\theta)/M) = p(\theta)$.
- By Bayes' theorem,

$$f(\theta|u < w(\theta)/M) = \frac{\Pr(\theta \text{ and } u < w(\theta)/M)}{\Pr(u < w(\theta)/M)} = \frac{\Pr(u < w(\theta)/M \mid \theta)g(\theta)}{\Pr(u < w(\theta)/M)}.$$

- But,
  - $\Pr(u < w(\theta)/M \mid \theta) = w(\theta)/M$ since $u \sim U(0,1)$, and
  - $$\begin{aligned}\Pr(u < w(\theta)/M) &= \int \Pr(u < w(\theta)/M \mid \theta)g(\theta)d\theta \\ &= \int w(\theta)/Mg(\theta)d\theta = 1/M \int w(\theta)g(\theta)d\theta = 1/M \int p(\theta)d\theta = 1/M.\end{aligned}$$
- Therefore,

$$f(\theta|u < w(\theta)/M) = \frac{\Pr(u < w(\theta)/M \mid \theta)g(\theta)}{\Pr(u < w(\theta)/M)} = \frac{w(\theta)/Mg(\theta)}{1/M} = w(\theta)g(\theta) = p(\theta).$$