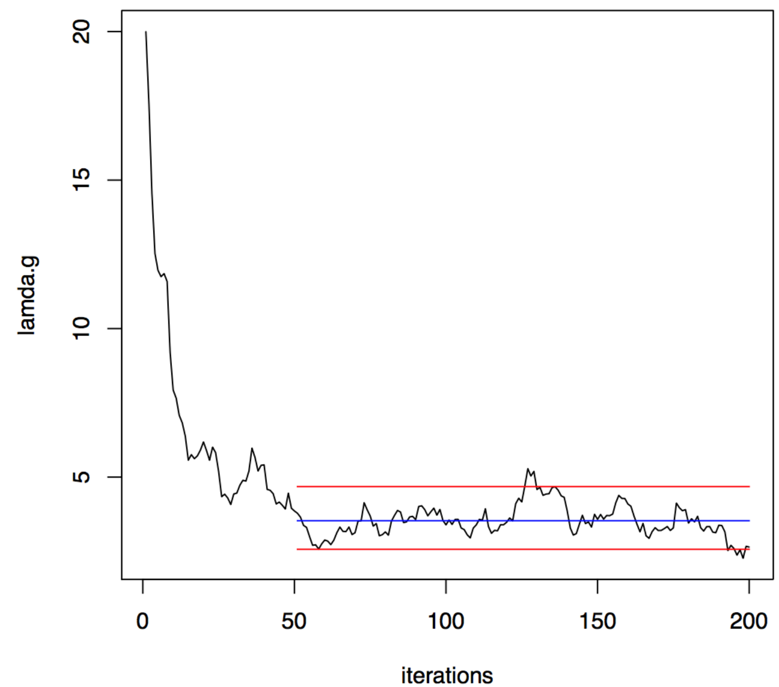
GENOME 541 Section 2

Lecture 4
Metropolis-Hastings
Sampling truncated distribution
Bayesian Hypothesis Testing

Yi Yin (yy224[at]uw.edu)
Shendure Lab
Department of Genome Sciences
University of Washington
Thursday, April 20, 2017

HW4

1. Suppose we have a model $y_i \sim \text{Pois}(\phi_i \lambda)$, with $\phi_i \sim \text{Ga}(\nu/2, \nu/2)$ & $\lambda \sim \text{Ga}(a, b)$ Use a=5, b=2 for your HW
2. Derive the full conditional posterior distributions of each ϕ_i and of λ
3. Simulate data with $\nu = 0.5$, $\lambda = 5$ and $n = 100$
4. Implement the Gibbs sampler & show a trace plot of samples of λ
5. Show an MCMC estimate of the posterior mean & 95% credible interval on this plot along with true value of λ - comment



Metropolis-(Hastings)

- With Gibbs Sampling, you still need full conditionals in closed form…
- Metropolis is an alternative that avoids this restriction
- Also start with θ_0 and sequentially update the parameters $\theta_1, \ldots, \theta_p$.

To draw θ_j^t:

1. Sample a candidate $\widetilde{\theta}_j^t \sim q_j(\cdot \mid \theta_j^{t-1})$
2. Let $\theta_j^t = \widetilde{\theta}_j^t$ with probability

$$\min\left\{1, \frac{\pi(\widetilde{\theta}_j^t)\, L(\mathbf{y} \mid \theta_j = \widetilde{\theta}_j^t, -)\, q_j(\theta_j^{t-1} \mid \widetilde{\theta}_j^t)}{\pi(\theta_j^{t-1})\, L(\mathbf{y} \mid \theta_j = \theta_j^{t-1}, -)\, q_j(\widetilde{\theta}_j^t \mid \theta_j^{t-1})}\right\},$$

$L(\mathbf{y} \mid \theta_j = \widetilde{\theta}_j^t, -)$=likelihood given $\theta_j = \widetilde{\theta}_j^t$ and current values of other parameters

3. Otherwise let $\theta_j^t = \theta_j^{t-1}$.

Performance sensitive to the proposal distributions, $q_j(\cdot \mid \theta_j^{t-1})$

Most common proposal is $N(\theta_j^{t-1}, \kappa)$, which is centered on the previous value

This results in a *Metropolis* random walk

Inefficient if κ is too small or too large

Aiming at ~20% acceptance rate is usually not bad

Warning: M-H only works if tails of proposal are at least as heavy as tails of target!

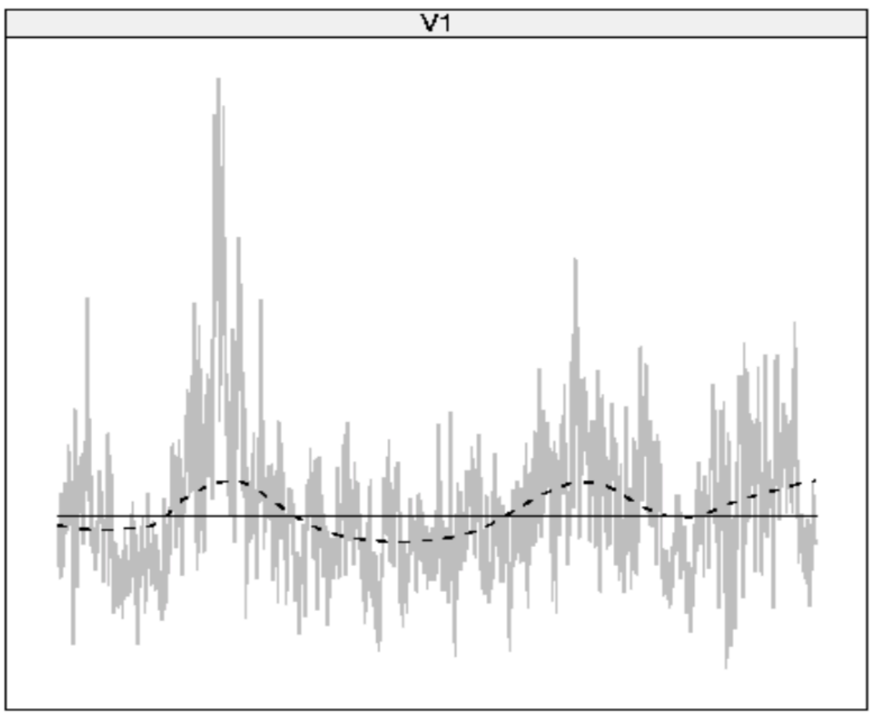
Convergence: initial drift in the samples towards a stationary distribution

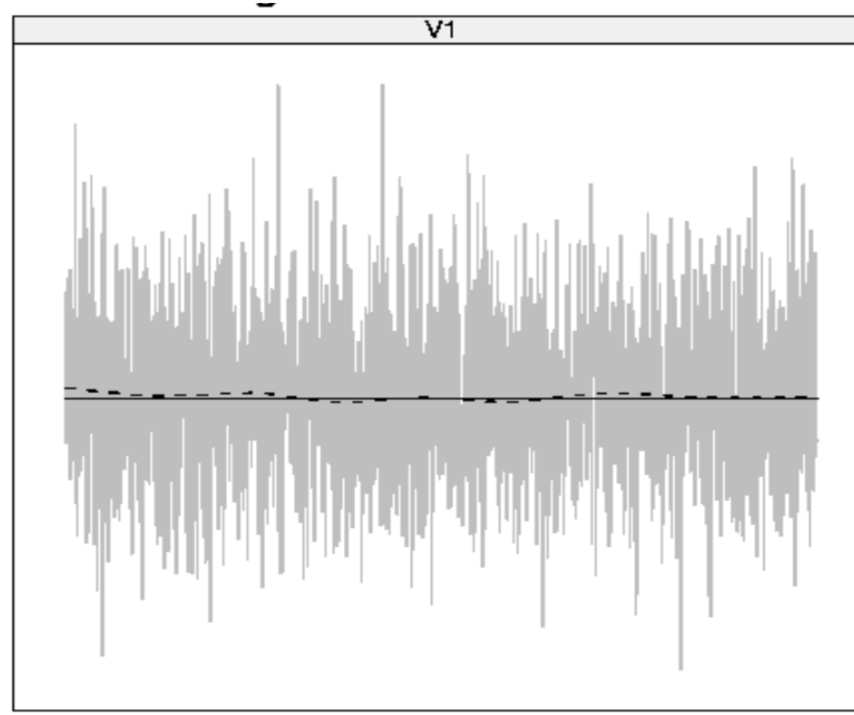
Burn-in: samples at start of the chain that are discarded to allow convergence

Slow mixing: tendency for high autocorrelation in the samples.

Thinning: practice of collecting every kth iteration to reduce autocorrelation

Trace plot: plot of sampled values of a parameter vs iteration #





Back to Lecture 1: How to sample from a truncated distribution?

- It is often useful to incorporate truncation
- Suppose we want to choose a prior for a probability of an event
- Let θ = probability Duke basketball wins
- Theoretically, it's between 0 and 1, but you may want to rule out very low values and very high values – say, $\theta \in [0.65, 0.95]$ with probability 1.
- How to choose a prior restricted to this interval?
- Of course, you can use unif(0.65, 0.95) instead of unif(0,1), which is also Beta(1,1), but you may want more flexibility than that.
- Say, you have ample evidence from previous Duke games, and want to summarize your prior belief that Duke wins with Beta(80, 20). (equivalent to 100 prior games worth of evidence, 80 wins, 20 losses)… and you want to truncate the Beta(80, 20) to [0.65, 0.95], instead of [0, 1].

- Suppose we have some arbitrary random variable $\theta \sim f$ with support on $\mathcal{Y}$
- For example $\theta \sim \text{beta}(a, b)$ has support on (0,1)
- Then, we can modify the density $f(\theta)$ to have support on a sub-interval $[a, b] \in \mathcal{Y}$
- The density $f(\theta)$ <u>truncated</u> to $[a, b]$ is

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

with $1(A)$ being the indicator function that returns 1 if A is true & 0 otherwise

Method 1: Rejection sampling, sample from Beta(80, 20), throw away all the samples that are <0.65 or >0.95. → could be inefficient

Method 2: the Inverse CDF method

The Inverse CDF method (without truncation)

- Suppose we have $\theta \sim f$, for some arbitrary continuous density f
- To sample θ, we can first sample $u \sim \text{Unif}(0,1)$ and then let $\theta = F^{-1}(u)$
- This is referred to as the inverse-cdf method
- We can sample $\theta \sim \text{beta}(c, d)$ through the inverse cdf method

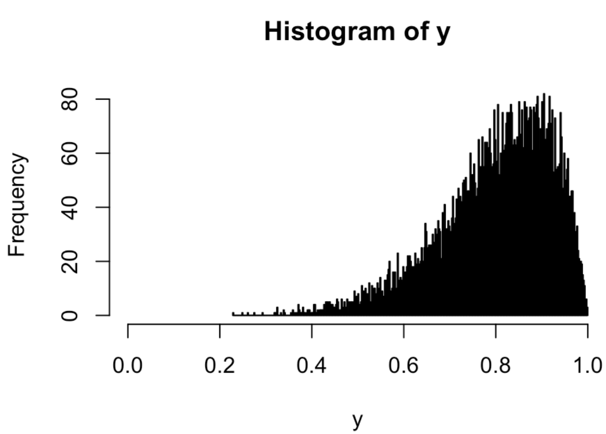
$$u = runif(1, 0, 1), \quad \theta = qbeta(u, c, d),$$

- We sample a uniform & then transform it using the inverse cdf of $\text{beta}(c, d)$

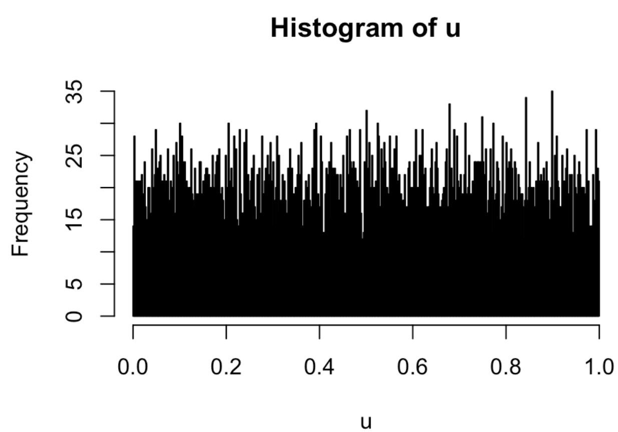
Let's pick c=8, d=2 for Beta(c, d).

You can either directly sample from Beta(8,2) using rbeta(), or you can sample from runif(), then use qbeta(), i.e., inverse cdf to sample z.

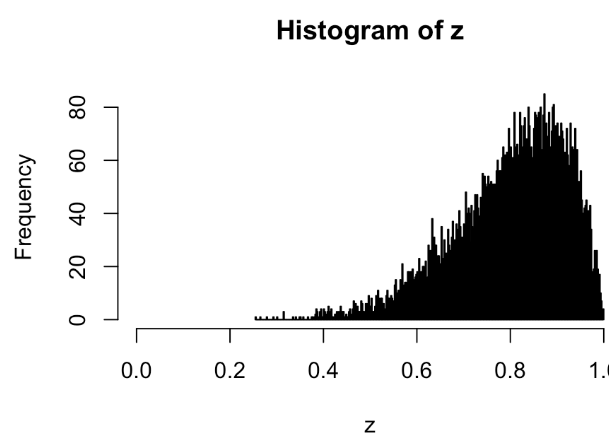
```
y <- rbeta(10000, 8, 2)
```




```
u <- runif(10000, 0, 1)
```




```
z <- qbeta(u, 8, 2)
```



The Inverse CDF method (How to apply truncation?)

- If we had the inverse cdf of beta(c, d) truncated to $[a, b]$ then we could use this
- Let f, F, F^{-1} denote the pdf, cdf, inverse-cdf without truncation & A=[a,b]. Then,

$$f_A(\theta) = \frac{f(\theta)1(\theta \in A)}{F(b) - F(a)}, \quad F_A(z) = \Pr(\theta \leq z) = \frac{F(z) - F(a)}{F(b) - F(a)}.$$

f corresponds to dbeta()
F corresponds to pbeta()
F^{-1} corresponds to qbeta()

- To find the inverse cdf $F_A^{-1}(p)$, we let

$$p = \frac{F(z) - F(a)}{F(b) - F(a)}, \qquad (1)$$

  and solve for z as a function of p
- Re-expressing (1) as a function of $F(z)$,

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

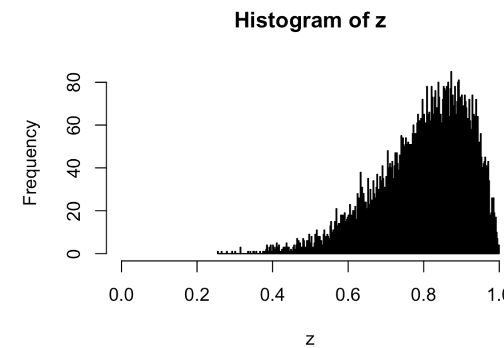
- Apply the untruncated inverse cdf F^{-1} to both sides,

$$z = F^{-1}\left[\{F(b) - F(a)\}p + F(a)\right] = F_A^{-1}(p)$$

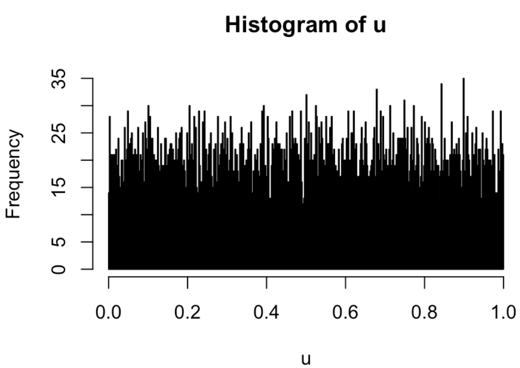
How to sample from truncated distribution using inverse cdf method

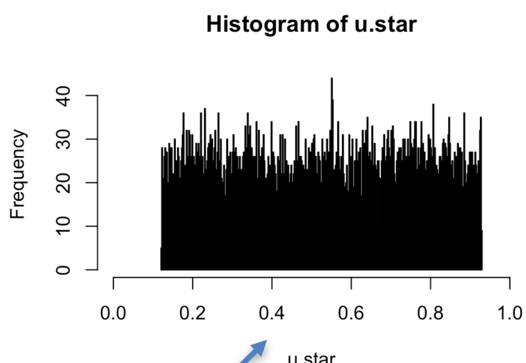
- We now have all the pieces to use the inverse-cdf method to sample $\theta \sim f_A$ (f truncated to A)
- We first draw a uniform$(0,1)$ random variable, $u \sim runif(1, 0, 1)$
- We then apply the linear transformation, $u^* = \{F(b) - F(a)\}u + F(a)$
- Finally we plug u^* into the untruncated cdf $\theta = F^{-1}(u^*)$

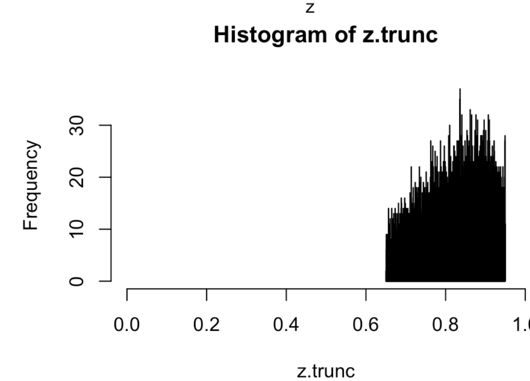
z <- qbeta(u, 8, 2)



u <- runif(10000, 0, 1)







u.star = (pbeta(0.95,8,2) - pbeta(0.65,8,2)) * u + pbeta(0.65,8,2)

z.trunc <- qbeta(u.star, 8, 2)

Bayesian Hypothesis Testing

Let's go back to the biased coin example. As a reminder, we have

1. A binomial sampling model
2. A Beta prior distribution Beta(a,b) such that we have conjugacy, i.e., the posterior distribution is also a Beta distribution Beta(a.hat, b.hat).

Now, say we want to formalize a hypothesis testing procedure under this Bayesian paradigm on whether the coin is biased.

Consider the simple case of testing a point null hypothesis:

$$H_0 : p = 0.5 \quad \text{against} \quad H_1 : p \neq 0.5.$$

Note that this is also useful in Bayesian variable selection (or feature selection) in regression problems

1. you have p parameters
2. you have $\beta_1, \ldots, \beta_p$, which are the regression coefficients (parameters of interest)
3. if some of them are shrunk to 0 (for example, in Lasso Regression), you don't select those features.
4. essentially testing whether $\beta_j=0$ (H_{0j}: $\beta_j = 0$; H_{1j}: $\beta_j \neq 0$).

Bayesian Hypothesis Testing

$$p(H_1|Y) = \frac{p(H_1)L(Y|H_1)}{L(Y)}$$

$$p(H_0|Y) = \frac{p(H_0)L(Y|H_0)}{L(Y)}$$

$$\frac{p(H_1|Y)}{p(H_0|Y)} = \frac{p(H_1)L(Y|H_1)}{p(H_0)L(Y|H_0)} = \frac{p(H_1)}{p(H_0)} \times \frac{L(Y|H_1)}{L(Y|H_0)}$$

$$posterior\ odds = prior\ odds \times BF$$

This is the marginal likelihood we usually see. Make sure you understand the difference between this term and the L(Y) term above.

Bayes Factors

- The Bayes factor (BF) can be used as a summary of the weight of evidence in the data in favor of model γ_1 over model γ_2.
- The BF for model γ_1 over γ_2 is defined as the ratio of posterior to prior odds, which is simply:

  $$BF_{12} = \frac{L_1(\mathbf{y})}{L_2(\mathbf{y})},$$

  a ratio of marginal likelihoods.
- Values of $BF_{12} > 1$ suggest that model m_1 is preferred, with the weight of evidence in favor of m_1 increasing as BF_{12} increases.

- In this problem, we essentially have $\Gamma = \{0, 1\}$, with $\gamma = 0$ if H_0 is true and $\gamma = 1$ if H_1 is true
- We let $\Pr(\gamma = 1) = 0.5$ to assign equal prior probability to each hypothesis
- We require a prior for the probability of heads under H_1:

$$\pi \sim \text{beta}(a, b)$$

Remember that E(π)=a/b and a+b can be viewed as the prior sample size
We usually center the prior distribution on the null, which is p(H)=p(T)=0.5, therefore, we can pick Beta(1,1) or Beta(0.5, 0.5) as a weakly informative prior (we will come back to this choice of a and b).

Question in HW4:
Can you make the prior sample size infinitely small, i.e., let a, b → 0?
What influence will it have on the hypothesis testing problem?

$$H_0 : p = 0.5 \quad \text{against} \quad H_1 : p \neq 0.5.$$

- The marginal likelihood under the null hypothesis is simply binomial with

$$L(\mathbf{y} \,|\, \gamma = 0) = \binom{n}{x} 0.5^n.$$

- The marginal likelihood under the alternative hypothesis is

$$\begin{aligned} L(\mathbf{y} \,|\, \gamma = 1) &= \int \binom{n}{x} \pi^x (1-\pi)^{n-x} \frac{\pi^{a-1}(1-\pi)^{b-1}}{B(a,b)} d\pi \\ &= \frac{B(\hat{a},\hat{b})}{B(a,b)} \binom{n}{x} \underbrace{\int \frac{1}{B(\hat{a},\hat{b})} \pi^{\hat{a}-1}(1-\pi)^{\hat{b}-1} d\pi}_{\text{This part integrates to 1}} \end{aligned}$$

$\hat{a} = a + x, \; \hat{b} = b + n - x$

a=b=1 in the Beta(a,b) prior

Hence, the marginal likelihood under the alternative is

$$\begin{aligned} L(\mathbf{y} \,|\, \gamma = 1) &= \frac{B(\hat{a},\hat{b})n!}{B(a,b)x!(n-x)!} = \frac{G(1+x)G(1+n-x)n!}{G(2+n)x!(n-x)!} \\ &= \frac{x!(n-x)!n!}{(n+1)!x!(n-x)!} = \frac{1}{n+1}, \end{aligned}$$

G() denotes the gamma function

Under the Beta(1,1) prior, we have

$$L(\mathbf{y} \mid \gamma = 0) = \binom{n}{x} 0.5^n \qquad L(\mathbf{y} \mid \gamma = 1) = \frac{1}{n+1}$$

The Bayes factor in favor of H_0 is:

$$BF_{12} = \frac{L_1(\mathbf{y})}{L_2(\mathbf{y})} = \frac{n!(n+1)}{2^n x!(n-x)!}$$

(H_0 over H_1)

$$p(H_1 \mid x, n) = \frac{p(H_1)\, L(x \mid H_1, n)}{p(H_1)\, L(x \mid H_1, n) + p(H_0)\, L(x \mid H_0, n)} = \frac{1}{1 + BF}$$

“lgamma” means log of gamma function, o.w. may run into numerical problems

- The rcode to calculate the BF in favor of H_0 is:

  *exp(lgamma(n+1) + log(n+1) - n*log(2) - lgamma(x+1) - lgamma(n-x+1))*

- If we let $n = 100$ and $x = 20$ we obtain $BF = 4.27e-08$, which converts to $\Pr(H_1 \mid x, n) \approx 1$, which is very strong evidence the coin is biased
- If we let $n = 10$ and $x = 8$ we obtain $BF = 0.483$ and $\Pr(H_1 \mid x, n) = 0.674$, which is weak evidence of bias
- Unlike the p-value, which is used to assess whether there is significance evidence to reject the null hypothesis, Bayes factors provide a weight of evidence in favor of the null
- As an exercise, calculate the BF for general a, b and observe what happens as $a, b \to 0$ - do we encounter Lindley's paradox?

Go back to Slide 14 and 15, there we let a=b=1, what will happen if we let a, b→0? (We assume that n is finite). Use beta(a,b) in R to experiment how B(a,b) changes when a, b → 0.