Statistics 360/601 – Modern Bayesian Theory

Alexander Volfovsky

Lecture 7 - Sept 20, 2018

Monte Carlo

Monte Carlo approximation

Want to compute

$$E[\theta|y] = \int \theta p(\theta|y) d\theta$$

without worrying about actual integration...

- Let θ be a parameter of interest
- Let $y_1, \ldots, y_n$ be numerical values of a sample from a distribution $p(y_1, \ldots, y_n|\theta)$
- Let $\theta^1, \ldots, \theta^S \sim p(\theta|y_1, \ldots, y_n)$ be iid samples from the posterior.
- We estimate our posterior quantity of interest as $1/S \sum g(\theta^i)$

Full on example

- Data: birthrates and education information for women.
- Question: are there any differences in the bithrates of women without and with bachelor's degrees?
- 111 women without bachelors gave birth to 217 children.
- 44 women with bachelors degrees gave birth to 66 children.
- Model: Poisson with a mean parameter θ_1 for women without a bachelors, and θ_2 with a bachelors.
- Prior: $\theta_1, \theta_2 \sim gamma(a = 2, b = 1)$
- Posterior: $p(\theta_1 | \sum_{i=1}^{111} Y_{i,1} = 217)$ is $gamma(219, 11)$ and $p(\theta_2 | \sum_{i=1}^{44} Y_{i,2} = 66)$ is $gamma(68, 45)$.
- Want to compute $p(\theta_1 > \theta_2 | \sum Y_{i1} = 217, \sum Y_{i2} = 66)$ and $p(\tilde{Y}_1 > \tilde{Y}_2 | \sum Y_{i1} = 217, \sum Y_{i2} = 66)$.

Birthrates example

Magic answers to our questions:

$$p(\theta_1 > \theta_2 | \sum Y_{i1} = 217, \sum Y_{i2} = 66) = 0.97$$

$$p(\tilde{Y}_1 > \tilde{Y}_2 | \sum Y_{i1} = 217, \sum Y_{i2} = 66) = 0.48$$

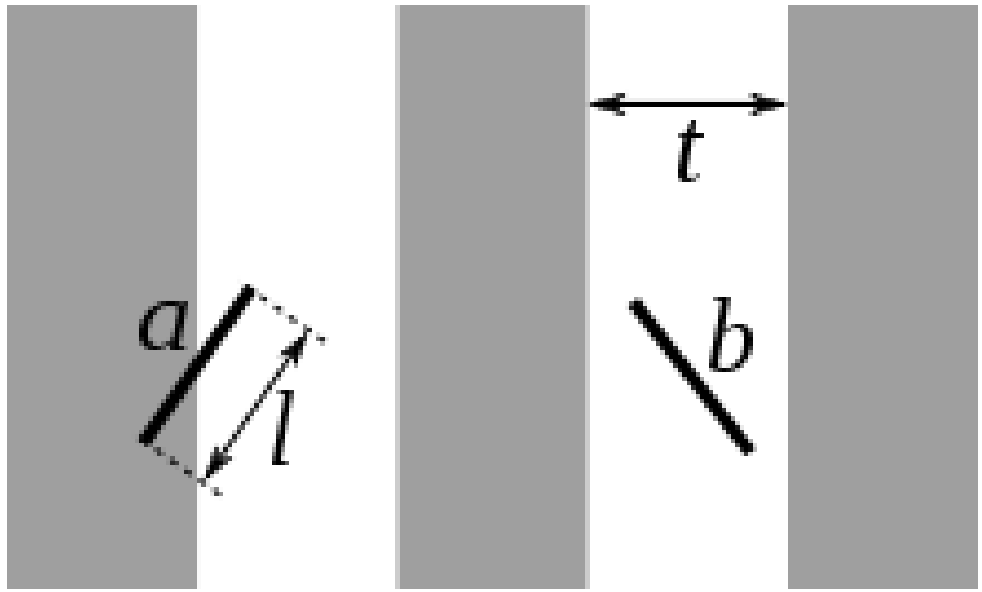
As you might imagine these are some ... sums and integrals.

Monte Carlo to the rescue

```
> a<-2 ; b<-1
> sy1<-217 ;  n1<-111
> sy2<-66  ;  n2<-44
>
> theta1.mc<-rgamma(10000,a+sy1, b+n1)
> theta2.mc<-rgamma(10000,a+sy2, b+n2)
>
> y1.mc<-rpois(10000,theta1.mc)
> y2.mc<-rpois(10000,theta2.mc)
>
>
> mean(theta1.mc>theta2.mc)
[1] 0.9745
>
> mean(y1.mc>y2.mc)
[1] 0.4768
```

Buffon's needle

A crazy man wants to estimate π and all he has is an infinite carpet and a needle...



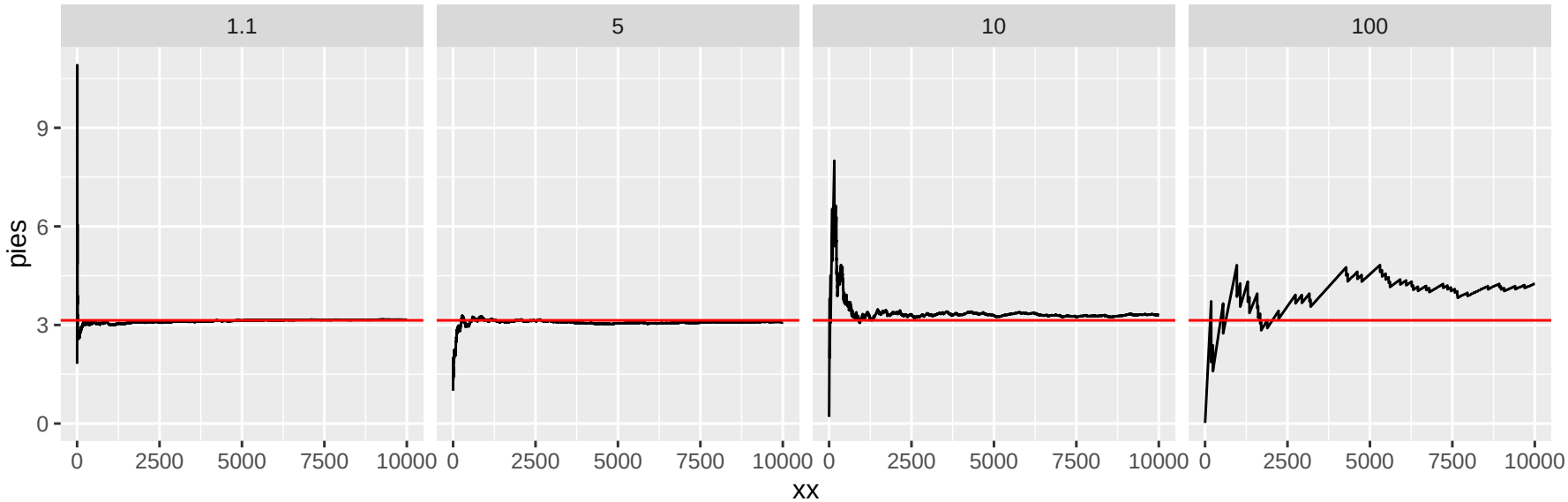
Wiki Images.

Practical example:

`http://mste.illinois.edu/activity/buffon/`

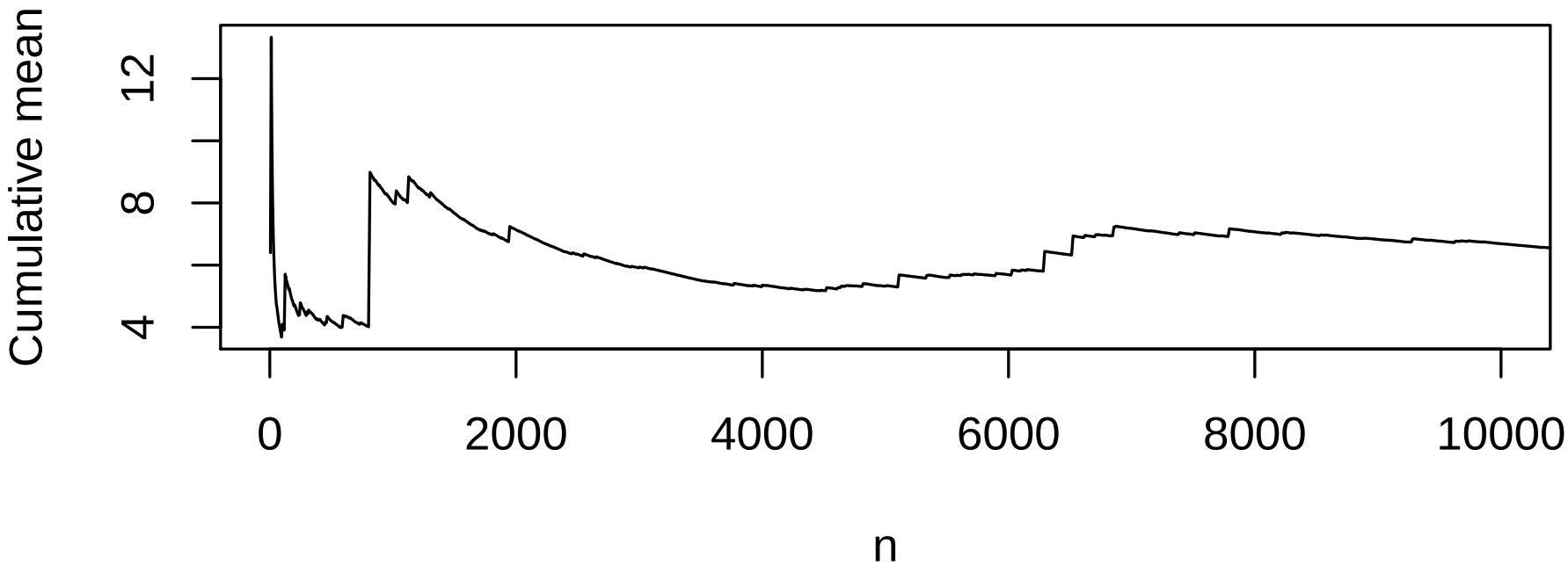
Buffon's needle

- Real world: throw a bunch of needles and estimate.
- Computer world: pretend to throw a bunch of needles and estimate.
- Sample (x, θ) uniformly from $[0, t/2] \times [0, \pi/2]$.



Monte Carlo does not solve everything

Monte Carlo failure



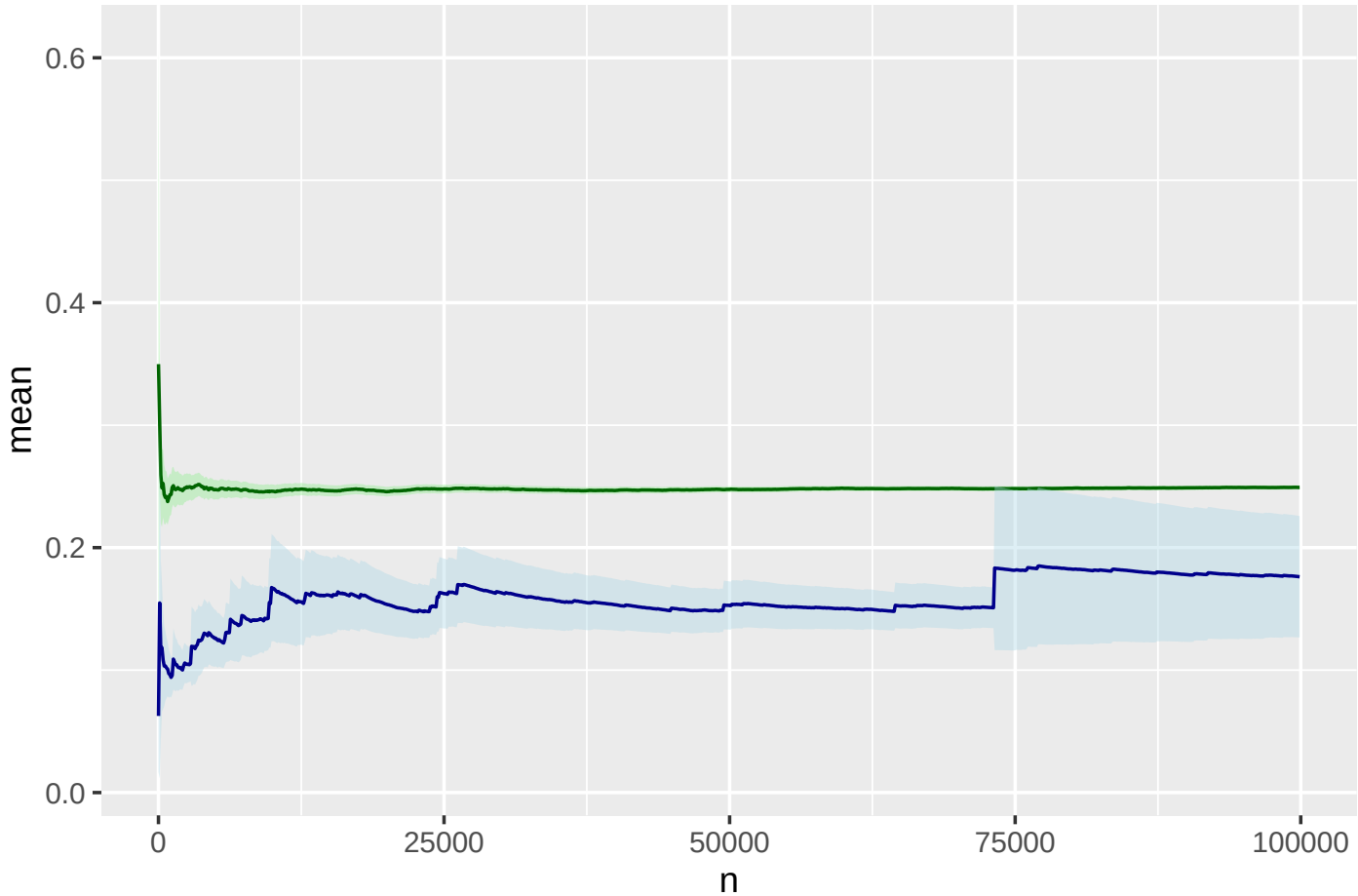
Vanilla MC vs Importance sampling

- Want to estimate $I = \int_0^1 x^3 dx$.
- Method 1:
  1. Sample $u_1, \ldots, u_S \sim Unif(0, 1)$
  2. Estimate $\hat{I} = \sum(u_i)^3/S$
- Method 2:
  1. Sample $b_1, \ldots, b_S \sim Beta(a, b)$
  2. Estimate $\hat{I} = \sum((b_i)^3/dbeta(b_i, a, b))/S$
- Which one is better?

  (note that the optimal choice for importance sampling has the density $g^\star(x) = 4x^3$ – obviously $x^3/g^\star = 1/4$ is the answer)

Vanilla MC vs Importance sampling

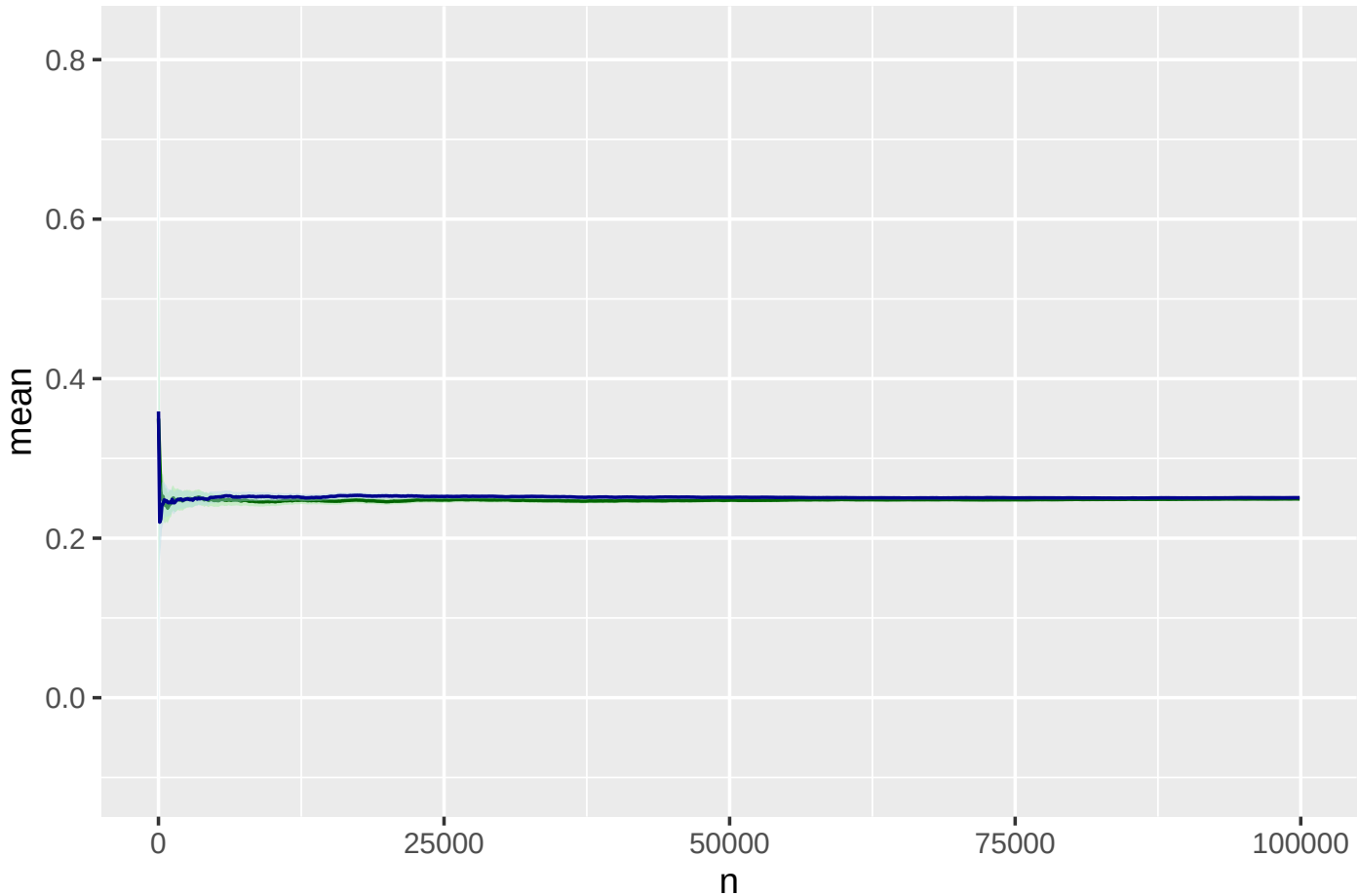
Setting $a = 10, b = 10$



(green is vanilla, blue is importance with beta(a, b))

Vanilla MC vs Importance sampling

Setting $a = 1/2, b = 1/2$

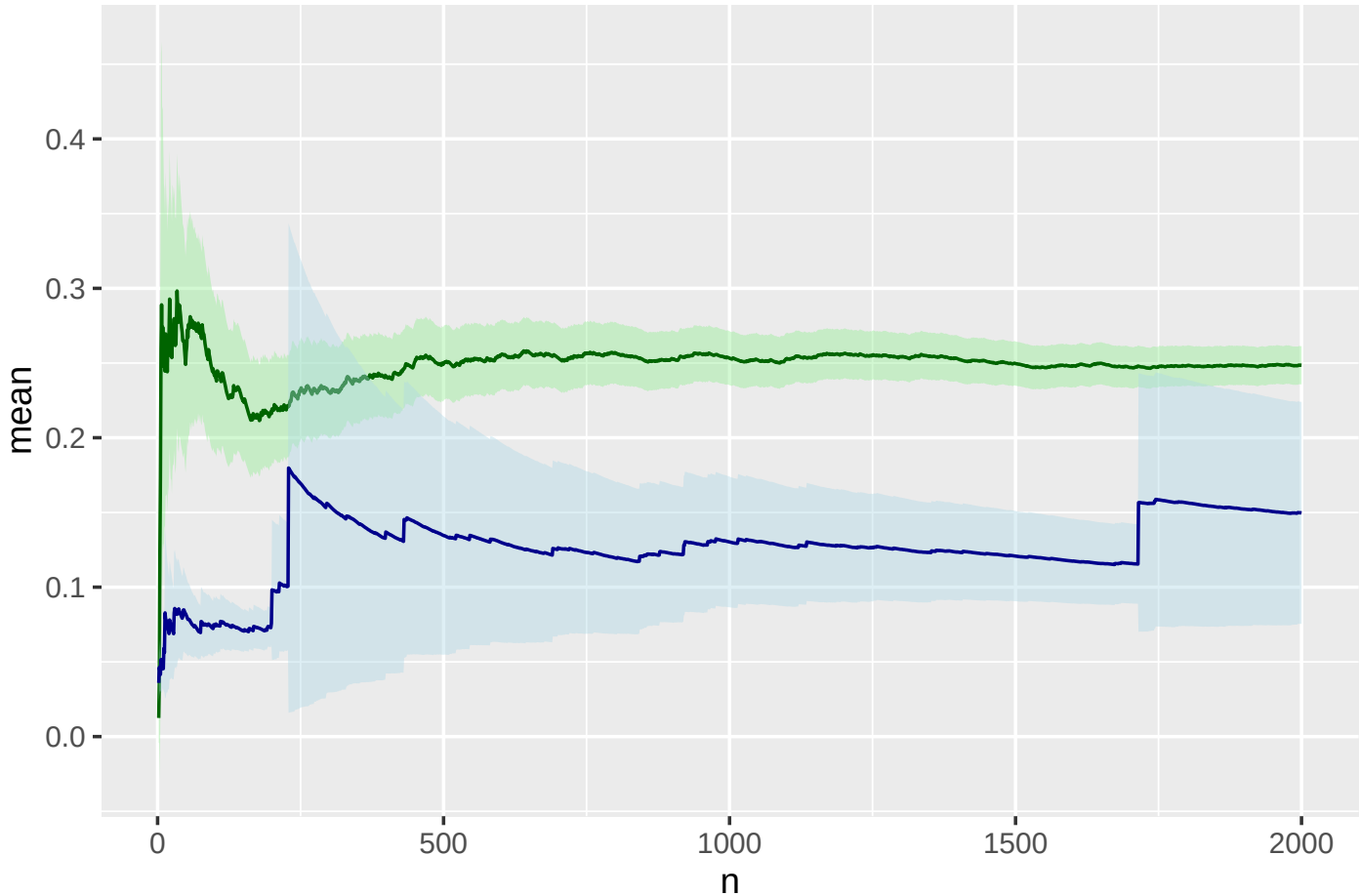


(green is vanilla, blue is importance with beta(a, b))

Vanilla MC vs Importance sampling

zooming in

Setting $a = 10, b = 10$

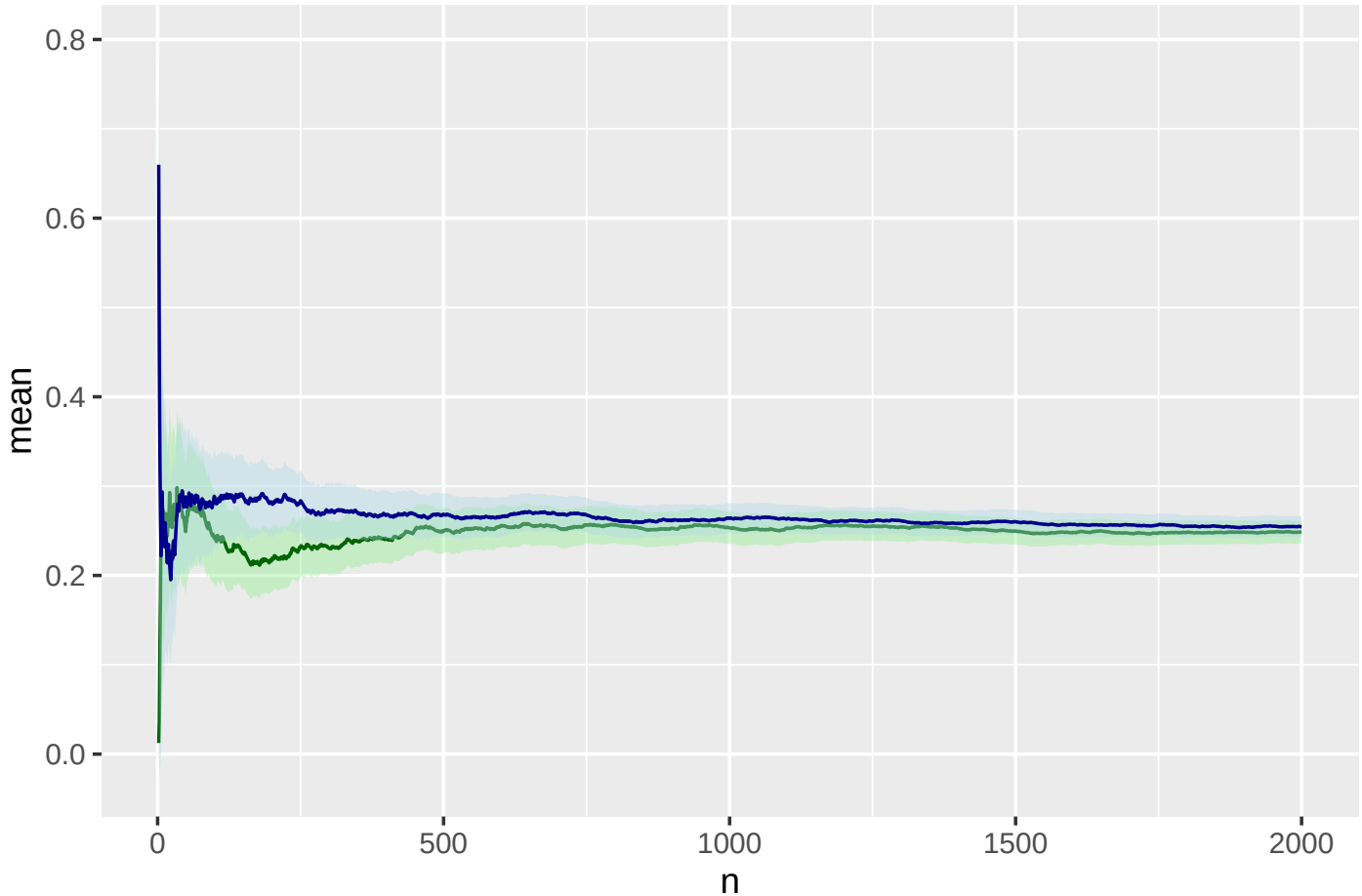


(green is vanilla, blue is importance with beta(a, b))

Vanilla MC vs Importance sampling

zooming in

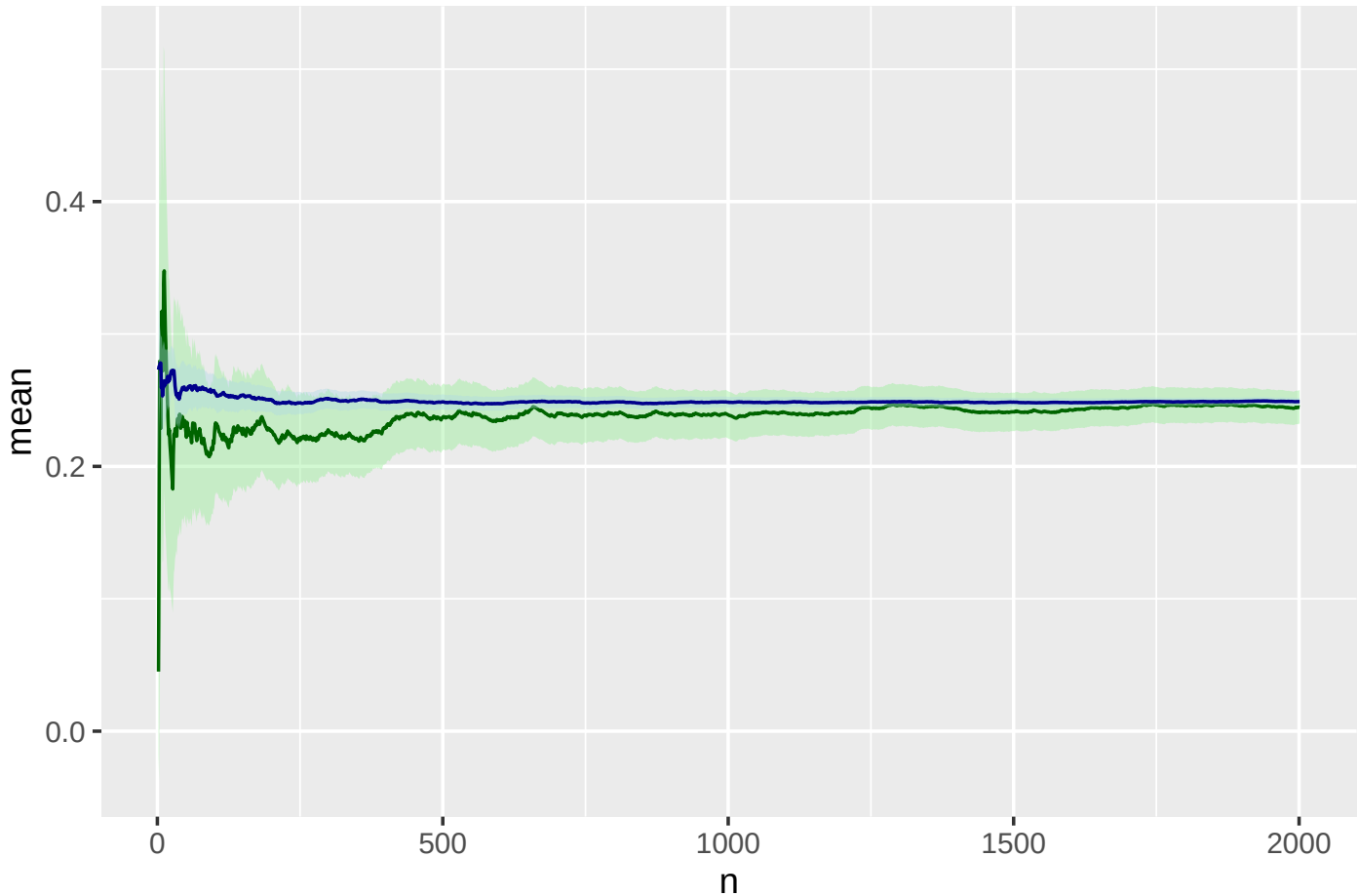
Setting $a = 1/2, b = 1/2$



(green is vanilla, blue is importance with beta(a, b))

Vanilla MC vs Importance sampling

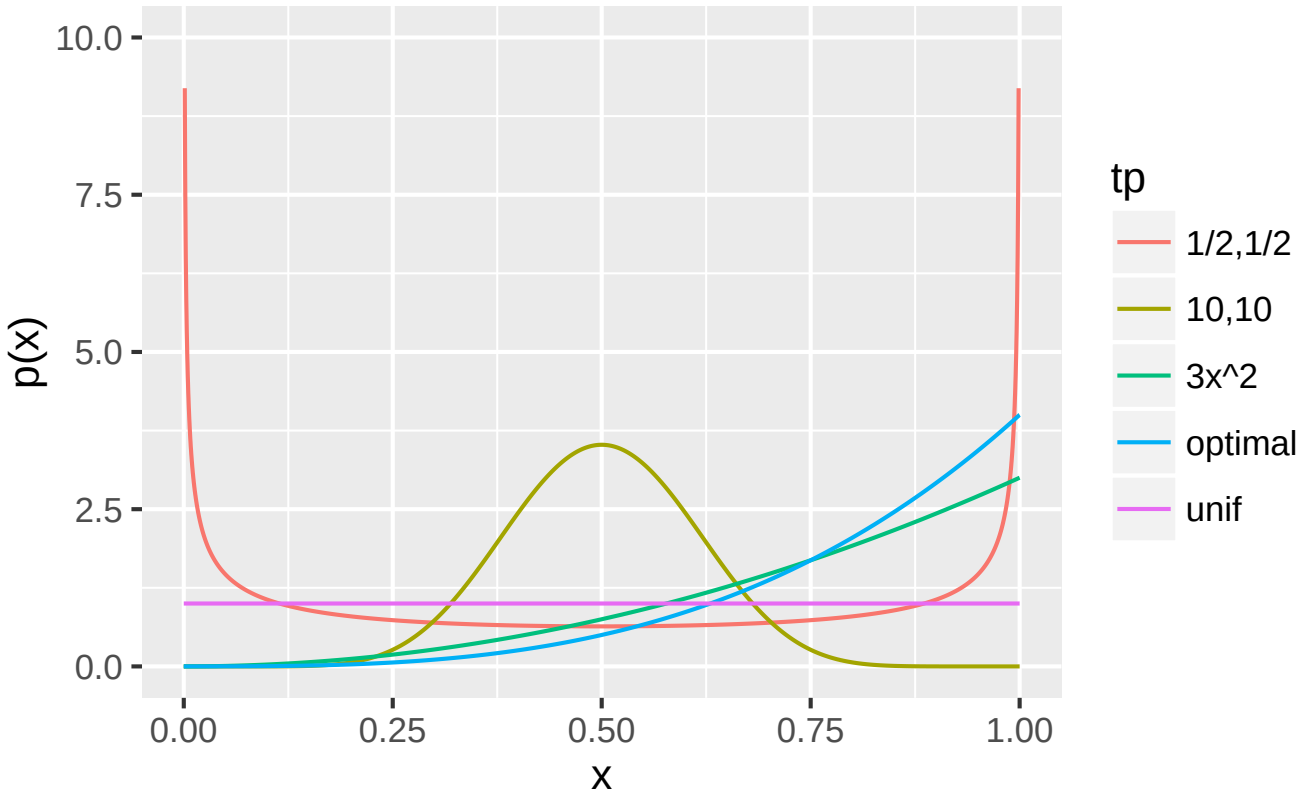
So much better... $g(x) = 3x^2$



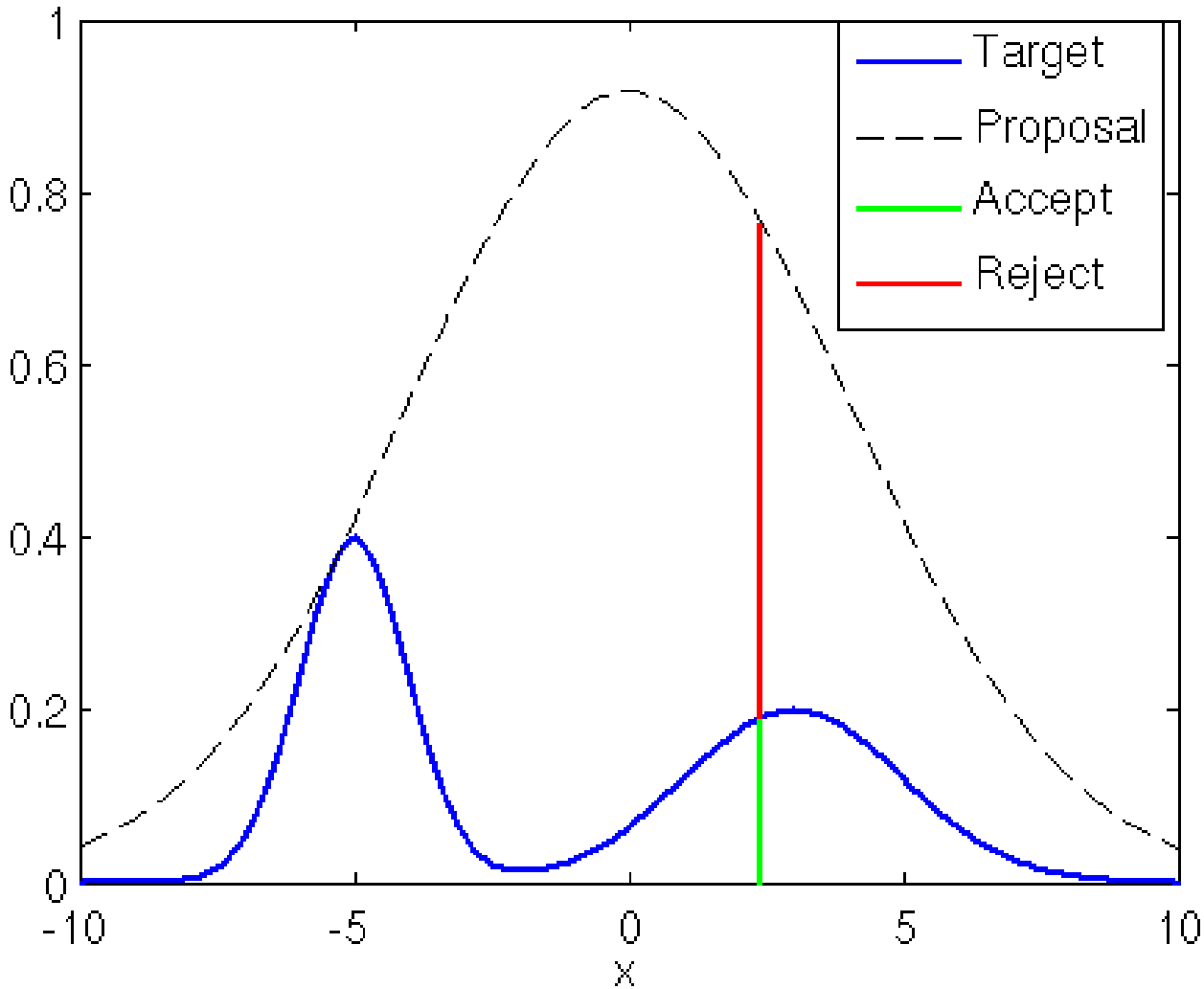
(green is vanilla, blue is importance with beta(a, b))

Vanilla MC vs Importance sampling

Why is there a difference?



Rejection sampling



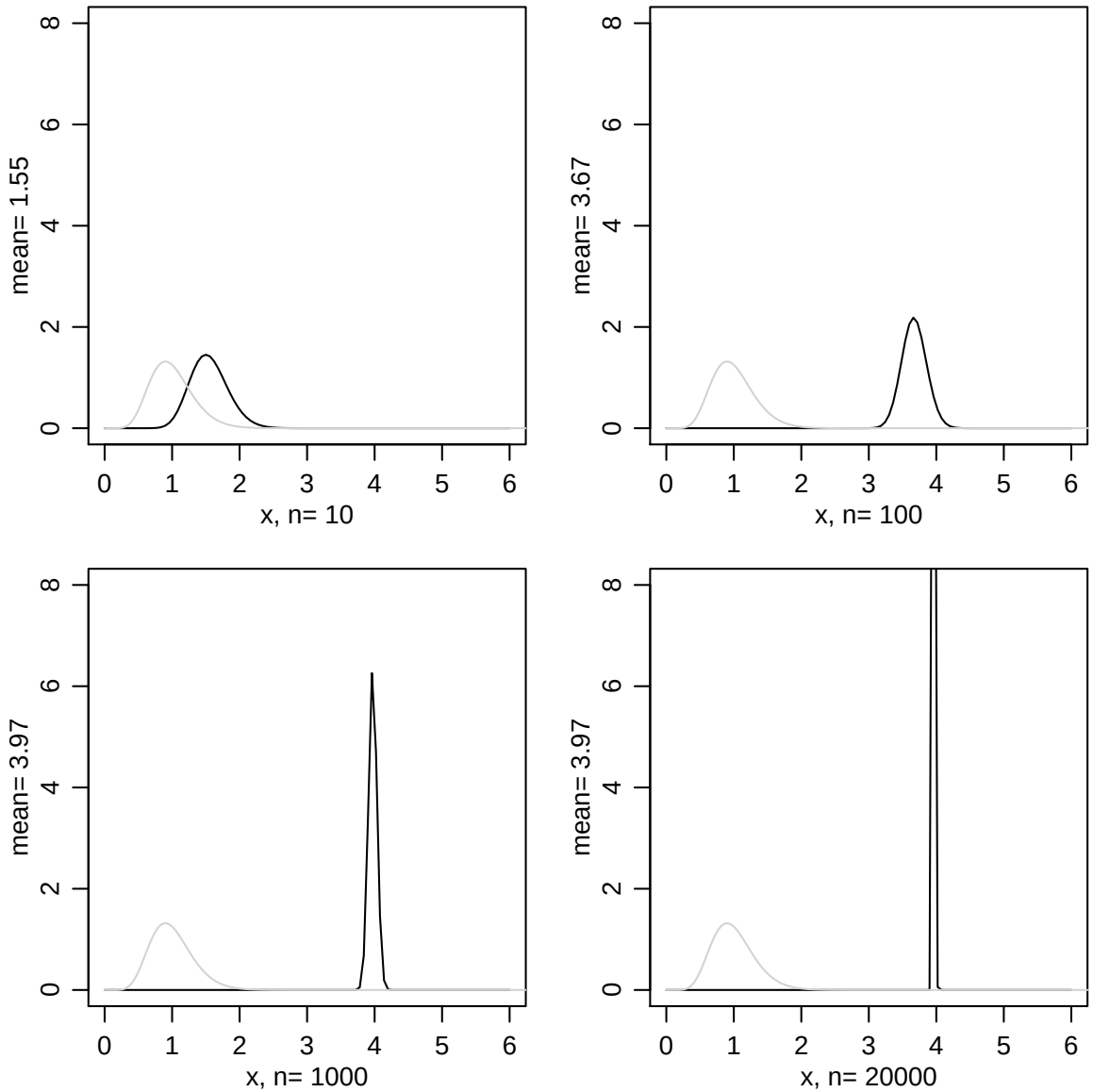
Consistent parameters??

- Let the data come from $Y \sim$Geometric(p).
- Recall that a geometric variable has pdf

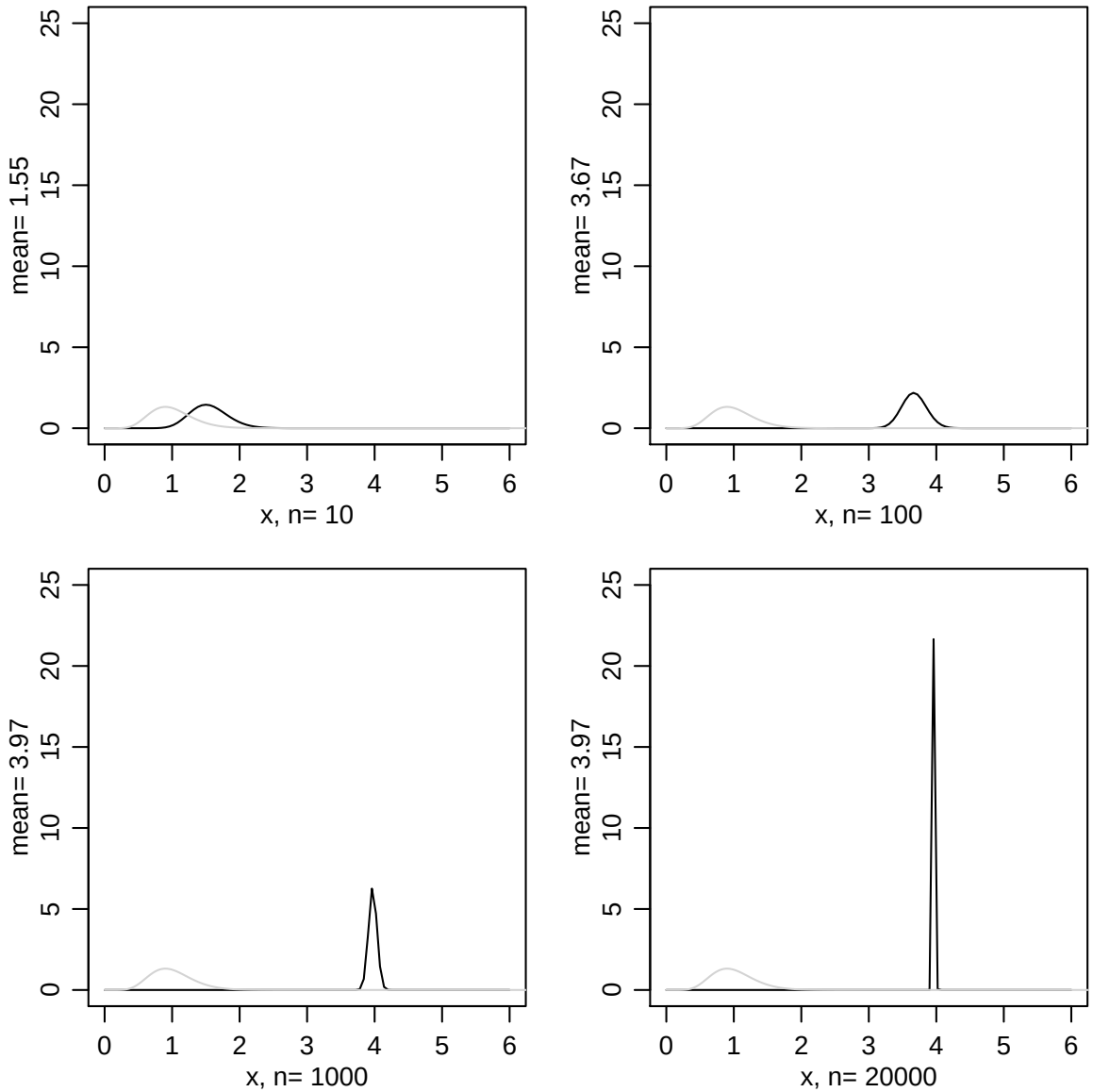
$$\Pr(Y = k) = (1 - p)^k p$$

  and it captures a success on the $k + 1$ trial after k failures.
- The mean of a geometric is $(1 - p)/p$.
- Let the model we think the data comes from be *Poisson*(θ).
- Let the prior be a Gamma(α, β).
- We know the posterior is Gamma$(\alpha + \sum y_i, \beta + n)$.

Consistent parameters??



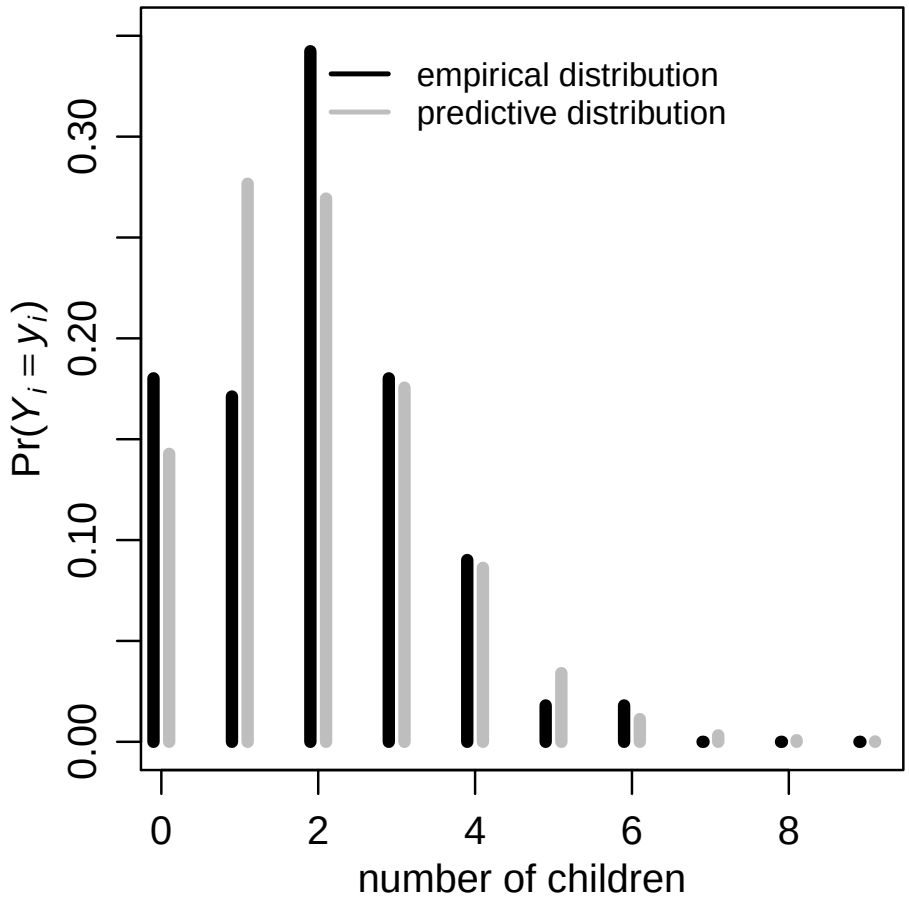
Consistent parameters??



Posterior Predictive Checks

Posterior predictive checking

Lets look at the data (from the book).

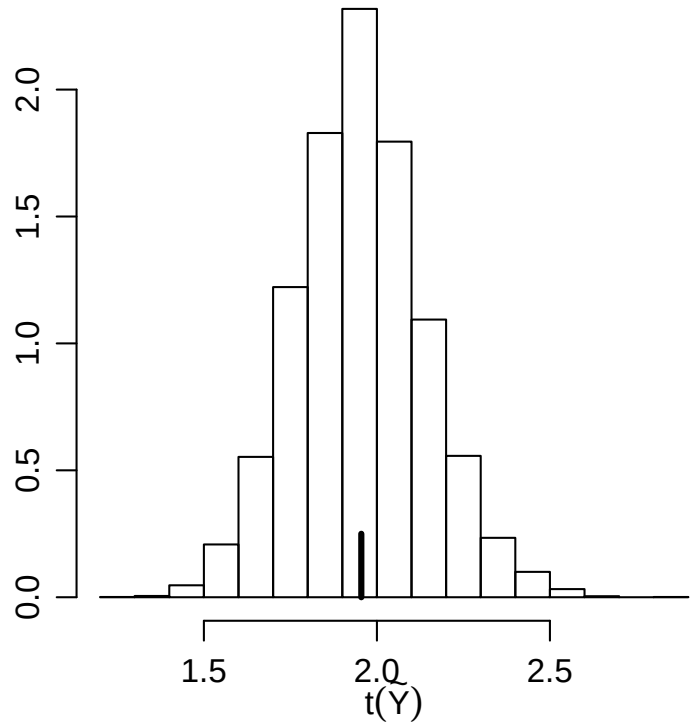


Data: twice as many women with 2 children as with 1.

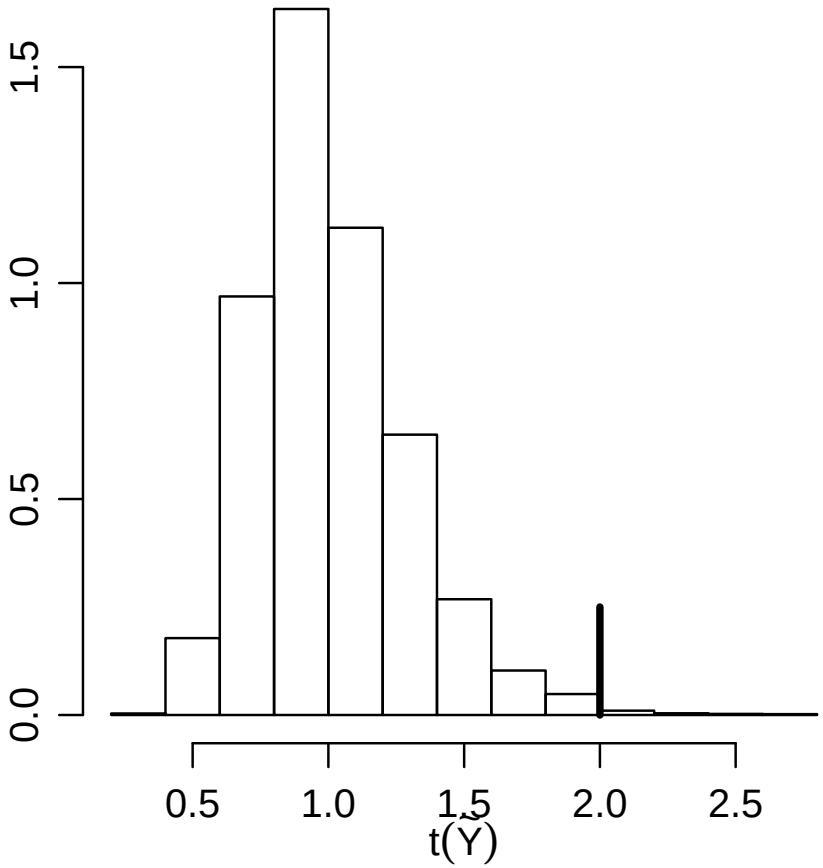
Posterior predictive: fewer women with 2 children than with 1.

Posterior predictive checking

```
> t.mc <- t2.mc <-NULL
> for(s in 1:10000) {
>  theta1<-rgamma(1,a+sum(y1), b+length(y1))
>  y1.mc<-rpois(length(y1),theta1)
>  t.mc <- c(t.mc,mean(y1.mc))
>  t2.mc<-c(t2.mc,sum(y1.mc==2)/sum(y1.mc==1))
> }
```



Posterior predictive checking



Where $t(y)$ is the ratio of 2's to 1's in a dataset.

Posterior predictive checking

- Lets look at the data (Gelman, Meng and Stern).

Posterior predictive checking

- Lets look at the data (Gelman, Meng and Stern).
- Gelman (1990,92) describe positron emission tomography experiment.

Posterior predictive checking

- Lets look at the data (Gelman, Meng and Stern).
- Gelman (1990,92) describe positron emission tomography experiment.
- Goal: Estimate the density of a radioactive isotope in a cross-section of the brain.

Posterior predictive checking

- Lets look at the data (Gelman, Meng and Stern).
- Gelman (1990,92) describe positron emission tomography experiment.
- Goal: Estimate the density of a radioactive isotope in a cross-section of the brain.
- 2-d image is estaimte from gamma-ray counts in a rigng of detectors around the head.

Posterior predictive checking

- Lets look at the data (Gelman, Meng and Stern).
- Gelman (1990,92) describe positron emission tomography experiment.
- Goal: Estimate the density of a radioactive isotope in a cross-section of the brain.
- 2-d image is estaimte from gamma-ray counts in a rigng of detectors around the head.
- n bins of counts based on positions of detectors — 6 million counts.

Posterior predictive checking

- Lets look at the data (Gelman, Meng and Stern).
- Gelman (1990,92) describe positron emission tomography experiment.
- Goal: Estimate the density of a radioactive isotope in a cross-section of the brain.
- 2-d image is estaimte from gamma-ray counts in a rigng of detectors around the head.
- n bins of counts based on positions of detectors — 6 million counts.
- Bin count y_i are modeled as independent Poisson(θ_i)

Posterior predictive checking

- Lets look at the data (Gelman, Meng and Stern).
- Gelman (1990,92) describe positron emission tomography experiment.
- Goal: Estimate the density of a radioactive isotope in a cross-section of the brain.
- 2-d image is estaimte from gamma-ray counts in a rigng of detectors around the head.
- n bins of counts based on positions of detectors — 6 million counts.
- Bin count y_i are modeled as independent Poisson(θ_i)
- $\Theta = Ag + r$ where g is the unknown image, A is a known linear operator and r are known corrections.

Posterior predictive checking

- Lets look at the data (Gelman, Meng and Stern).
- Gelman (1990,92) describe positron emission tomography experiment.
- Goal: Estimate the density of a radioactive isotope in a cross-section of the brain.
- 2-d image is estaimte from gamma-ray counts in a rigng of detectors around the head.
- n bins of counts based on positions of detectors — 6 million counts.
- Bin count y_i are modeled as independent Poisson(θ_i)
- $\Theta = Ag + r$ where g is the unknown image, A is a known linear operator and r are known corrections.
- A, g and r are non-negative.

Posterior predictive checking

- Lets look at the data (Gelman, Meng and Stern).
- Gelman (1990,92) describe positron emission tomography experiment.
- Goal: Estimate the density of a radioactive isotope in a cross-section of the brain.
- 2-d image is estaimte from gamma-ray counts in a rigng of detectors around the head.
- n bins of counts based on positions of detectors — 6 million counts.
- Bin count y_i are modeled as independent Poisson(θ_i)
- $\Theta = Ag + r$ where g is the unknown image, A is a known linear operator and r are known corrections.
- A, g and r are non-negative.
- This is an easy problem without the constraint.

Posterior predictive checking

- Poisson noise + model problems makes exact non-negative solutions impossible.

Posterior predictive checking

- Poisson noise + model problems makes exact non-negative solutions impossible.
- Use an estimate $\hat{g}$ and capture the discrepancy between y and $\hat{\theta} = A\hat{g}_r$.

Posterior predictive checking

- Poisson noise + model problems makes exact non-negative solutions impossible.
- Use an estimate $\hat{g}$ and capture the discrepancy between y and $\hat{\theta} = A\hat{g}_r$.
- Consider χ^2 discrepancy.

$$X^2(y;\hat{\theta}) = \sum \frac{(y_i - \hat{\theta}_i)^2}{\hat{\theta}_i}$$

Posterior predictive checking

- Poisson noise + model problems makes exact non-negative solutions impossible.
- Use an estimate $\hat{g}$ and capture the discrepancy between y and $\hat{\theta} = A\hat{g}_r$.
- Consider χ^2 discrepancy.

$$X^2(y;\hat{\theta}) = \sum \frac{(y_i - \hat{\theta}_i)^2}{\hat{\theta}_i}$$

- Fitting to real data: y with $n = 22464$.

Posterior predictive checking

- Poisson noise + model problems makes exact non-negative solutions impossible.
- Use an estimate $\hat{g}$ and capture the discrepancy between y and $\hat{\theta} = A\hat{g}_r$.
- Consider χ^2 discrepancy.

$$X^2(y;\hat{\theta}) = \sum \frac{(y_i - \hat{\theta}_i)^2}{\hat{\theta}_i}$$

- Fitting to real data: y with $n = 22464$.
- The best-fit non-negative image $\hat{g}$ was not an exact fit leading to the discrepancy between y and $\hat{\theta}$ to be $X^2(y;\hat{\theta}) \approx 30,000$.

Posterior predictive checking

- Poisson noise + model problems makes exact non-negative solutions impossible.
- Use an estimate $\hat{g}$ and capture the discrepancy between y and $\hat{\theta} = A\hat{g}_r$.
- Consider χ^2 discrepancy.

$$X^2(y;\hat{\theta}) = \sum \frac{(y_i - \hat{\theta}_i)^2}{\hat{\theta}_i}$$

- Fitting to real data: y with $n = 22464$.
- The best-fit non-negative image $\hat{g}$ was not an exact fit leading to the discrepancy between y and $\hat{\theta}$ to be $X^2(y;\hat{\theta}) \approx 30,000$.
- Reject the model!

Posterior predictive checking

- Poisson noise + model problems makes exact non-negative solutions impossible.
- Use an estimate $\hat{g}$ and capture the discrepancy between y and $\hat{\theta} = A\hat{g}_r$.
- Consider χ^2 discrepancy.

$$X^2(y;\hat{\theta}) = \sum \frac{(y_i - \hat{\theta}_i)^2}{\hat{\theta}_i}$$

- Fitting to real data: y with $n = 22464$.
- The best-fit non-negative image $\hat{g}$ was not an exact fit leading to the discrepancy between y and $\hat{\theta}$ to be $X^2(y;\hat{\theta}) \approx 30,000$.
- Reject the model!
- Possible failures:

Posterior predictive checking

- Poisson noise + model problems makes exact non-negative solutions impossible.
- Use an estimate $\hat{g}$ and capture the discrepancy between y and $\hat{\theta} = A\hat{g}_r$.
- Consider χ^2 discrepancy.

$$X^2(y;\hat{\theta}) = \sum \frac{(y_i - \hat{\theta}_i)^2}{\hat{\theta}_i}$$

- Fitting to real data: y with $n = 22464$.
- The best-fit non-negative image $\hat{g}$ was not an exact fit leading to the discrepancy between y and $\hat{\theta}$ to be $X^2(y;\hat{\theta}) \approx 30,000$.
- Reject the model!
- Possible failures:
  - Error in the specification of A, r

Posterior predictive checking

- Poisson noise + model problems makes exact non-negative solutions impossible.
- Use an estimate $\hat{g}$ and capture the discrepancy between y and $\hat{\theta} = A\hat{g}_r$.
- Consider χ^2 discrepancy.

$$X^2(y;\hat{\theta}) = \sum \frac{(y_i - \hat{\theta}_i)^2}{\hat{\theta}_i}$$

- Fitting to real data: y with $n = 22464$.
- The best-fit non-negative image $\hat{g}$ was not an exact fit leading to the discrepancy between y and $\hat{\theta}$ to be $X^2(y;\hat{\theta}) \approx 30,000$.
- Reject the model!
- Possible failures:
  - Error in the specification of A, r
  - Lack of independence

Posterior predictive checking

- Poisson noise + model problems makes exact non-negative solutions impossible.
- Use an estimate $\hat{g}$ and capture the discrepancy between y and $\hat{\theta} = A\hat{g}_r$.
- Consider χ^2 discrepancy.

$$X^2(y;\hat{\theta}) = \sum \frac{(y_i - \hat{\theta}_i)^2}{\hat{\theta}_i}$$

- Fitting to real data: y with $n = 22464$.
- The best-fit non-negative image $\hat{g}$ was not an exact fit leading to the discrepancy between y and $\hat{\theta}$ to be $X^2(y;\hat{\theta}) \approx 30,000$.
- Reject the model!
- Possible failures:
  - Error in the specification of A, r
  - Lack of independence
  - super-Poisson variance in the counts

Posterior predictive checking

- Poisson noise + model problems makes exact non-negative solutions impossible.
- Use an estimate $\hat{g}$ and capture the discrepancy between y and $\hat{\theta} = A\hat{g}_r$.
- Consider χ^2 discrepancy.

$$X^2(y;\hat{\theta}) = \sum \frac{(y_i - \hat{\theta}_i)^2}{\hat{\theta}_i}$$

- Fitting to real data: y with $n = 22464$.
- The best-fit non-negative image $\hat{g}$ was not an exact fit leading to the discrepancy between y and $\hat{\theta}$ to be $X^2(y;\hat{\theta}) \approx 30,000$.
- Reject the model!
- Possible failures:
  - Error in the specification of A, r
  - Lack of independence
  - super-Poisson variance in the counts
  - Error from discretizing the continuous g.

Posterior predictive checking

- The "critically bad" level of $X^2(y;\hat{\theta})$ is $n + 2\sqrt{2n} \approx 23,000$.

Posterior predictive checking

- The "critically bad" level of $X^2(y;\hat{\theta})$ is $n + 2\sqrt{2n} \approx 23,000$.
- We reassess and change the model.

Posterior predictive checking

- The "critically bad" level of $X^2(y;\hat{\theta})$ is $n + 2\sqrt{2n} \approx 23,000$.
- We reassess and change the model.
- We get a χ^2 discrepancy that is $22,000$ or even $20,000$.

Posterior predictive checking

- The "critically bad" level of $X^2(y;\hat{\theta})$ is $n + 2\sqrt{2n} \approx 23,000$.
- We reassess and change the model.
- We get a χ^2 discrepancy that is $22,000$ or even $20,000$.
- Should we just accept this new model?

Posterior predictive checking

- The "critically bad" level of $X^2(y;\hat{\theta})$ is $n + 2\sqrt{2n} \approx 23,000$.
- We reassess and change the model.
- We get a χ^2 discrepancy that is $22,000$ or even $20,000$.
- Should we just accept this new model?
- No.

Posterior predictive checking

- The "critically bad" level of $X^2(y;\hat{\theta})$ is $n + 2\sqrt{2n} \approx 23,000$.
- We reassess and change the model.
- We get a χ^2 discrepancy that is $22,000$ or even $20,000$.
- Should we just accept this new model?
- No.
- Be skeptical!