Chapter 10

- 10.1 Numerical integration (overview)
- 10.2 Distributional approximations (overview, more in Chapter 4 and 13)
- 10.3 Direct simulation and rejection sampling (overview)
- 10.4 Importance sampling (used in PSIS-LOO discussed later)
- 10.5 How many simulation draws are needed? (Ex 10.1 and 10.2)
  - see chapter notes and extra slides for how many significant digits to report
- 10.6 Software (can be skipped)
- 10.7 Debugging (can be skipped)

Notation

- In this chapter, generic $p(\theta)$ is used instead of $p(\theta|y)$
- Unnormalized distribution is denoted by $q(\cdot)$
  - $\int q(\theta)d\theta \neq 1$, but finite
  - $q(\cdot) \propto p(\cdot)$
- Proposal distribution is denoted by $g(\cdot)$

Numerical accuracy – floating point

- Floating point presentation of numbers. e.g. with 64bits
  - closest value to zero is $\approx 2.2 \cdot 10^{-308}$
    - generate sample of 600 from normal distribution:
      qr=rnorm(600)
    - calculate joint density given normal:
      prod(dnorm(qr)) $\rightarrow$ 0 (underflow)

Numerical accuracy – floating point

- Floating point presentation of numbers. e.g. with 64bits
  - closest value to zero is $\approx 2.2 \cdot 10^{-308}$
    - generate sample of 600 from normal distribution: qr=rnorm(600)
    - calculate joint density given normal: prod(dnorm(qr)) $\rightarrow$ 0 (underflow)
    - see log densities in the next slide

Numerical accuracy – floating point

- Floating point presentation of numbers. e.g. with 64bits
  - closest value to zero is $\approx 2.2 \cdot 10^{-308}$
    - generate sample of 600 from normal distribution: qr=rnorm(600)
    - calculate joint density given normal: prod(dnorm(qr)) $\rightarrow$ 0 (underflow)
    - see log densities in the next slide
  - closest value to 1 is $\approx 1 \pm 2.2 \cdot 10^{-16}$
    - Laplace and ratio of girl and boy babies
    - pbeta(0.5, 241945, 251527) $\rightarrow$ 1 (rounding)

Numerical accuracy – floating point

- Floating point presentation of numbers. e.g. with 64bits
  - closest value to zero is $\approx 2.2 \cdot 10^{-308}$
    - generate sample of 600 from normal distribution:
      qr=rnorm(600)
    - calculate joint density given normal:
      prod(dnorm(qr)) $\rightarrow$ 0 (underflow)
    - see log densities in the next slide
  - closest value to 1 is $\approx 1 \pm 2.2 \cdot 10^{-16}$
    - Laplace and ratio of girl and boy babies
    - pbeta(0.5, 241945, 251527) $\rightarrow$ 1 (rounding)
    - pbeta(0.5, 241945, 251527, lower.tail=FALSE) $\approx -1.2 \cdot 10^{-42}$
      there is more accuracy near 0

Numerical accuracy – log scale

- Log densities
  - use log densities to avoid over- and underflows in floating point presentation
    - prod(dnorm(qr)) $\rightarrow$ 0 (underflow)
    - sum(dnorm(qr, log=TRUE)) $\rightarrow$ -847.3

Numerical accuracy – log scale

- Log densities
  - use log densities to avoid over- and underflows in floating point presentation
    - prod(dnorm(qr)) $\rightarrow$ 0 (underflow)
    - sum(dnorm(qr, log=TRUE)) $\rightarrow$ -847.3
    - how many observations we can now handle?

Numerical accuracy – log scale

- Log densities
  - use log densities to avoid over- and underflows in floating point presentation
    - prod(dnorm(qr)) $\rightarrow$ 0 (underflow)
    - sum(dnorm(qr, log=TRUE)) $\rightarrow$ -847.3
    - how many observations we can now handle?
  - compute exp as late as possible

Numerical accuracy – log scale

- Log densities
  - use log densities to avoid over- and underflows in floating point presentation
    - prod(dnorm(qr)) $\rightarrow$ 0 (underflow)
    - sum(dnorm(qr, log=TRUE)) $\rightarrow$ -847.3
    - how many observations we can now handle?
  - compute exp as late as possible
    - e.g. for $a > b$, compute $\log(\exp(a) + \exp(b)) = a + \log(1 + \exp(b - a))$

Numerical accuracy – log scale

- Log densities
  - use log densities to avoid over- and underflows in floating point presentation
    - prod(dnorm(qr)) $\rightarrow$ 0 (underflow)
    - sum(dnorm(qr, log=TRUE)) $\rightarrow$ -847.3
    - how many observations we can now handle?
  - compute exp as late as possible
    - e.g. for $a > b$, compute
      $\log(\exp(a) + \exp(b)) = a + \log(1 + \exp(b - a))$
      e.g. $\log(\exp(800) + \exp(800)) \rightarrow \text{Inf}$

Numerical accuracy – log scale

- Log densities
  - use log densities to avoid over- and underflows in floating point presentation
    - prod(dnorm(qr)) $\rightarrow$ 0 (underflow)
    - sum(dnorm(qr, log=TRUE)) $\rightarrow$ -847.3
    - how many observations we can now handle?
  - compute exp as late as possible
    - e.g. for $a > b$, compute
      $\log(\exp(a) + \exp(b)) = a + \log(1 + \exp(b - a))$
      e.g. $\log(\exp(800) + \exp(800)) \rightarrow$ Inf
      but $800 + \log(1 + \exp(800 - 800)) \approx 800.69$

Numerical accuracy – log scale

- Log densities
  - use log densities to avoid over- and underflows in floating point presentation
    - prod(dnorm(qr)) $\rightarrow$ 0 (underflow)
    - sum(dnorm(qr, log=TRUE)) $\rightarrow$ -847.3
    - how many observations we can now handle?
  - compute exp as late as possible
    - e.g. for $a > b$, compute
      $\log(\exp(a) + \exp(b)) = a + \log(1 + \exp(b - a))$
      e.g. $\log(\exp(800) + \exp(800)) \rightarrow$ Inf
      but $800 + \log(1 + \exp(800 - 800)) \approx 800.69$
    - e.g. in Metropolis-algorithm (Assignment 5) compute the log of ratio of densities using the identity
      $\log(a/b) = \log(a) - \log(b)$

It's all about expectations

$$E_{p(\theta|y)}[f(\theta)] = \int f(\theta)p(\theta|y)d\theta,$$

$$\text{where} \quad p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta}$$

It's all about expectations

$$E_{p(\theta|y)}[f(\theta)] = \int f(\theta)p(\theta|y)d\theta,$$

$$\text{where} \quad p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta}$$

We can easily evaluate $p(y|\theta)p(\theta)$ for any θ, but the integral $\int p(y|\theta)p(\theta)d\theta$ is usually difficult.

It's all about expectations

$$E_{p(\theta|y)}[f(\theta)] = \int f(\theta) p(\theta|y) d\theta,$$

$$\text{where} \quad p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta}$$

We can easily evaluate $p(y|\theta)p(\theta)$ for any θ, but the integral $\int p(y|\theta)p(\theta)d\theta$ is usually difficult.

We can use the unnormalized posterior $q(\theta|y) = p(y|\theta)p(\theta) \propto p(\theta|y)$, for example, in

It's all about expectations

$$E_{p(\theta|y)}[f(\theta)] = \int f(\theta)p(\theta|y)d\theta,$$

$$\text{where} \quad p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta}$$

We can easily evaluate $p(y|\theta)p(\theta)$ for any θ, but the integral $\int p(y|\theta)p(\theta)d\theta$ is usually difficult.

We can use the unnormalized posterior $q(\theta|y) = p(y|\theta)p(\theta) \propto p(\theta|y)$, for example, in

- Grid (equal spacing) evaluation with self-normalization

$$E_{p(\theta|y)}[f(\theta)] \approx \frac{\sum_{s=1}^{S}\left[f(\theta^{(s)})q(\theta^{(s)}|y)\right]}{\sum_{s=1}^{S} q(\theta^{(s)}|y)}$$

It's all about expectations

$$E_{p(\theta|y)}[f(\theta)] = \int f(\theta) p(\theta|y) d\theta,$$

$$\text{where} \quad p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta}$$

We can easily evaluate $p(y|\theta)p(\theta)$ for any θ, but the integral $\int p(y|\theta)p(\theta)d\theta$ is usually difficult.

We can use the unnormalized posterior $q(\theta|y) = p(y|\theta)p(\theta) \propto p(\theta|y)$, for example, in

- Grid (equal spacing) evaluation with self-normalization

$$E_{p(\theta|y)}[f(\theta)] \approx \frac{\sum_{s=1}^{S} \left[f(\theta^{(s)}) q(\theta^{(s)}|y) \right]}{\sum_{s=1}^{S} q(\theta^{(s)}|y)}$$

- Monte Carlo methods which can sample from $p(\theta^{(s)}|y)$ using only $q(\theta^{(s)}|y)$ (each draw has weight 1/S)

$$E_{p(\theta|y)}[f(\theta)] \approx \frac{1}{S} \sum_{s=1}^{S} f(\theta^{(s)})$$

It's all about expectations

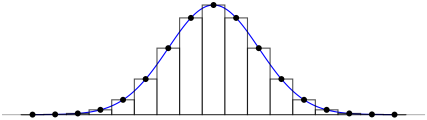
$$E_\theta[f(\theta)] = \int f(\theta)p(\theta|y)d\theta$$

- Conjugate priors and analytic solutions (Ch 1-5)
- Grid integration and other quadrature rules (Ch 3, 10)
- Independent Monte Carlo, rejection and importance sampling (Ch 10)
- Markov Chain Monte Carlo (Ch 11-12)
- Distributional approximations (Laplace, VB, EP) (Ch 4, 13)

Quadrature integration

- The simplest quadrature integration is grid integration

$$E[\theta] \approx \sum_{t=1}^{T} \theta^{(t)} w^{(t)},$$

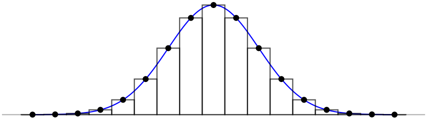


where $w^{(t)}$ is the normalized probability of a grid cell t, and $\alpha^{(t)}$ and $\beta^{(t)}$ are center locations of grid cells

Quadrature integration

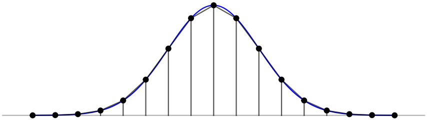
- The simplest quadrature integration is grid integration

$$\mathrm{E}[\theta] \approx \sum_{t=1}^{T} \theta^{(t)} w^{(t)},$$



where $w^{(t)}$ is the normalized probability of a grid cell t, and $\alpha^{(t)}$ and $\beta^{(t)}$ are center locations of grid cells

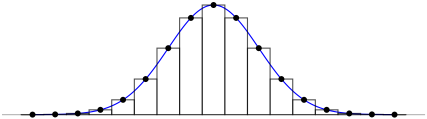
- In 1D further variations with better accuracy, e.g. trapezoid



Quadrature integration

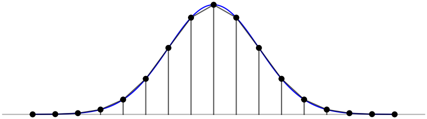
- The simplest quadrature integration is grid integration

$$E[\theta] \approx \sum_{t=1}^{T} \theta^{(t)} w^{(t)},$$



where $w^{(t)}$ is the normalized probability of a grid cell t, and $\alpha^{(t)}$ and $\beta^{(t)}$ are center locations of grid cells

- In 1D further variations with better accuracy, e.g. trapezoid

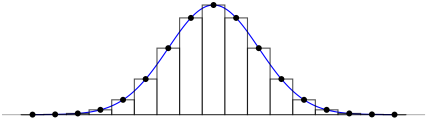


- Adaptive quadrature methods add evaluation points where needed, e.g., R function `integrate()`

Quadrature integration

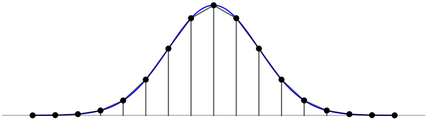
- The simplest quadrature integration is grid integration

$$E[\theta] \approx \sum_{t=1}^{T} \theta^{(t)} w^{(t)},$$



where $w^{(t)}$ is the normalized probability of a grid cell t, and $\alpha^{(t)}$ and $\beta^{(t)}$ are center locations of grid cells

- In 1D further variations with better accuracy, e.g. trapezoid



- Adaptive quadrature methods add evaluation points where needed, e.g., R function `integrate()`
- In 2D and higher
  - nested quadrature
  - product rules

Monte Carlo - history

- Used already before computers
  - Buffon (18th century; needles)
  - De Forest, Darwin, Galton (19th century)
  - Pearson (19th century; roulette)
  - Gosset (Student, 1908; hat)

Monte Carlo - history

- Used already before computers
  - Buffon (18th century; needles)
  - De Forest, Darwin, Galton (19th century)
  - Pearson (19th century; roulette)
  - Gosset (Student, 1908; hat)
- ”Monte Carlo method” term was proposed by Metropolis, von Neumann or Ulam in the end of 1940s
  - they worked together in atomic bomb project
  - Metropolis and Ulam, ”The Monte Carlo Method”, 1949

Monte Carlo - history

- Used already before computers
  - Buffon (18th century; needles)
  - De Forest, Darwin, Galton (19th century)
  - Pearson (19th century; roulette)
  - Gosset (Student, 1908; hat)
- ”Monte Carlo method” term was proposed by Metropolis, von Neumann or Ulam in the end of 1940s
  - they worked together in atomic bomb project
  - Metropolis and Ulam, ”The Monte Carlo Method”, 1949
- Bayesians started to have enough cheap computation time in 1990s
  - BUGS project started 1989 (last OpenBUGS release 2014)
  - Gelfand & Smith, 1990
  - Stan initial release 2012

Monte Carlo

- Simulate draws from the target distribution
  - these draws can be treated as any observations
  - a collection of draws is sample
- Use these draws, for example,
  - to compute means, deviations, quantiles
  - to draw histograms
  - to marginalize
  - etc.

Monte Carlo vs. deterministic

- Monte Carlo = simulation methods
  - evaluation points are selected stochastically (randomly)
- Deterministic methods (e.g. grid)
  - evaluation points are selected by some deterministic rule
  - good deterministic methods converge faster (need less function evaluations)

How many simulation draws are needed?

- How many draws or how big sample size?
- If draws are independent
  - usual methods to estimate the uncertainty due to a finite number of observations (finite sample size)
- Markov chain Monte Carlo produces dependent draws
  - requires additional work to estimate the effective sample size

How many simulation draws are needed?

- Expectation of unknown quantity

$$\mathrm{E}(\theta) \approx \frac{1}{S} \sum_{s=1}^{S} \theta^{(s)}$$

  if S is big and $\theta^{(s)}$ are independent, way may assume that the distribution of the expectation approaches normal distribution (see BDA3 Ch 4) with variance σ_θ^2/S (asymptotic normality)
    - this variance is independent on dimensionality of θ

How many simulation draws are needed?

- Expectation of unknown quantity

$$E(\theta) \approx \frac{1}{S}\sum_{s=1}^{S}\theta^{(s)}$$

  if S is big and $\theta^{(s)}$ are independent, way may assume that the distribution of the expectation approaches normal distribution (see BDA3 Ch 4) with variance σ_θ^2/S (asymptotic normality)
  - this variance is independent on dimensionality of θ
  - total variance is sum of the epistemic uncertainty in the posterior and the uncertainty due to using finite number of Monte Carlo draws

$$\sigma_\theta^2 + \sigma_\theta^2/S$$

How many simulation draws are needed?

- Expectation of unknown quantity

$$E(\theta) \approx \frac{1}{S}\sum_{s=1}^{S}\theta^{(s)}$$

  if S is big and $\theta^{(s)}$ are independent, way may assume that the distribution of the expectation approaches normal distribution (see BDA3 Ch 4) with variance σ_θ^2/S (asymptotic normality)
  - this variance is independent on dimensionality of θ
  - total variance is sum of the epistemic uncertainty in the posterior and the uncertainty due to using finite number of Monte Carlo draws

$$\sigma_\theta^2 + \sigma_\theta^2/S = \sigma_\theta^2(1 + 1/S)$$

How many simulation draws are needed?

- Expectation of unknown quantity

$$\mathrm{E}(\theta) \approx \frac{1}{S}\sum_{s=1}^{S}\theta^{(s)}$$

if S is big and $\theta^{(s)}$ are independent, way may assume that the distribution of the expectation approaches normal distribution (see BDA3 Ch 4) with variance σ_θ^2/S (asymptotic normality)

  - this variance is independent on dimensionality of θ
  - total variance is sum of the epistemic uncertainty in the posterior and the uncertainty due to using finite number of Monte Carlo draws

$$\sigma_\theta^2 + \sigma_\theta^2/S = \sigma_\theta^2(1 + 1/S)$$

  - e.g. if $S = 100$, deviation increases by $\sqrt{1 + 1/S} = 1.005$ i.e. Monte Carlo error is very small (for the expectation)

How many simulation draws are needed?

- Expectation of unknown quantity

$$\mathrm{E}(\theta) \approx \frac{1}{S}\sum_{s=1}^{S}\theta^{(s)}$$

if S is big and $\theta^{(s)}$ are independent, way may assume that the distribution of the expectation approaches normal distribution (see BDA3 Ch 4) with variance σ_θ^2/S (asymptotic normality)

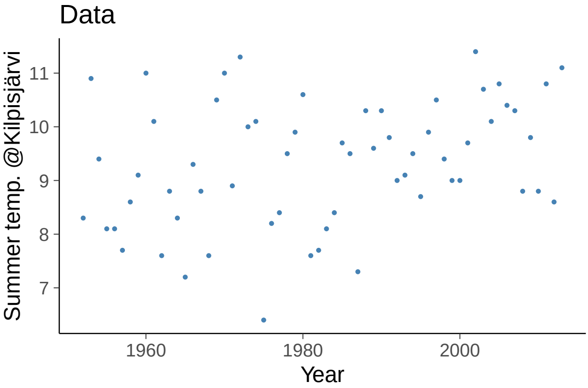
  - this variance is independent on dimensionality of θ
  - total variance is sum of the epistemic uncertainty in the posterior and the uncertainty due to using finite number of Monte Carlo draws

$$\sigma_\theta^2 + \sigma_\theta^2/S = \sigma_\theta^2(1 + 1/S)$$

  - e.g. if $S = 100$, deviation increases by $\sqrt{1 + 1/S} = 1.005$ i.e. Monte Carlo error is very small (for the expectation)
  - See BDA3 Ch 4 for counter-examples for asymptotic normality

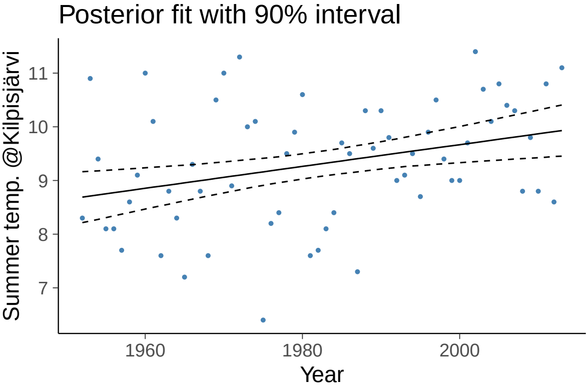
Example: Kilpisjärvi summer temperature

Average temperature in June, July, and August at Kilpisjärvi, Finland in 1952–2013

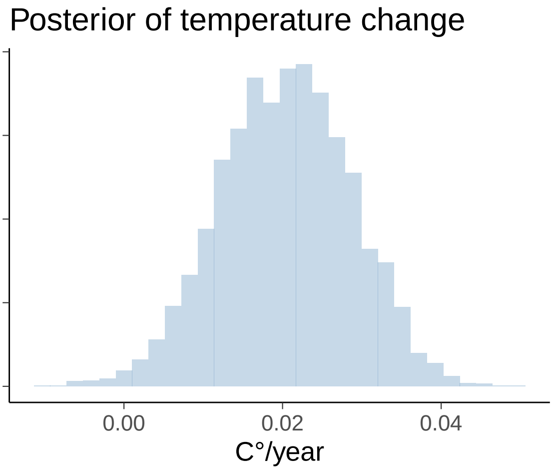


Example: Kilpisjärvi summer temperature

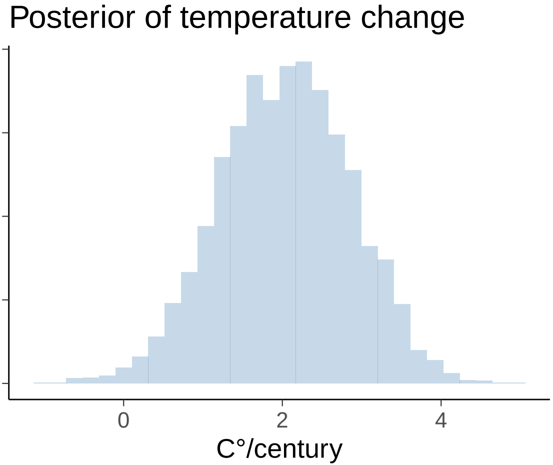
Average temperature in June, July, and August at Kilpisjärvi, Finland in 1952–2013



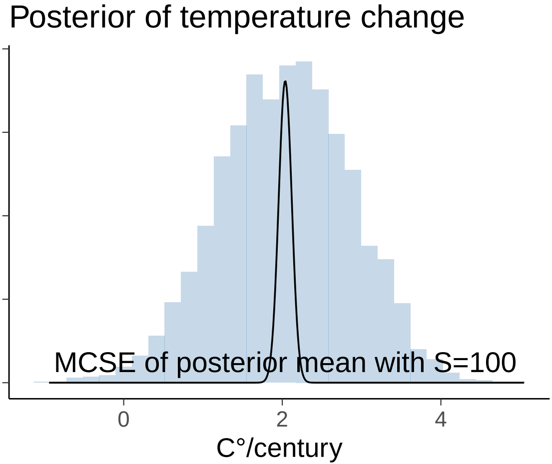
Example: Kilpisjärvi summer temperature



Example: Kilpisjärvi summer temperature



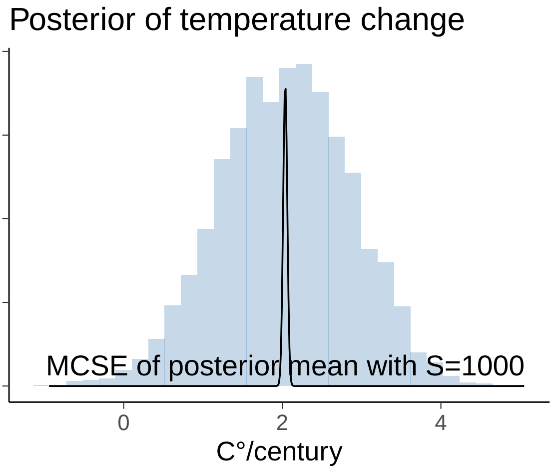
Example: Kilpisjärvi summer temperature



$\sigma_\theta \approx 0.827$, MCSE ≈ 0.0827, total deviation ≈ 0.831

total deviation$^2 = \sigma_\theta^2 +$ MCSE^2

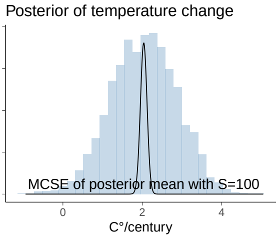
Example: Kilpisjärvi summer temperature

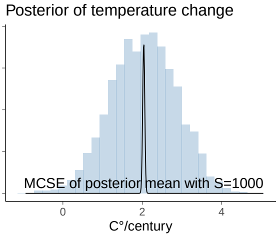


$\sigma_\theta \approx 0.827$, MCSE ≈ 0.0261, total deviation ≈ 0.827

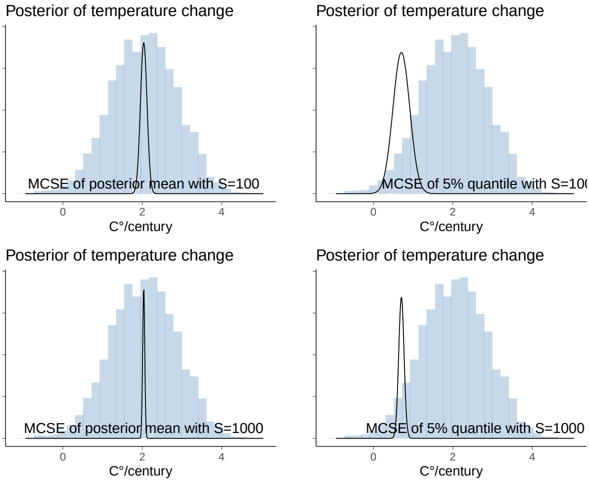
total deviation$^2 = \sigma_\theta^2 +$ MCSE^2

Example: Kilpisjärvi summer temperature

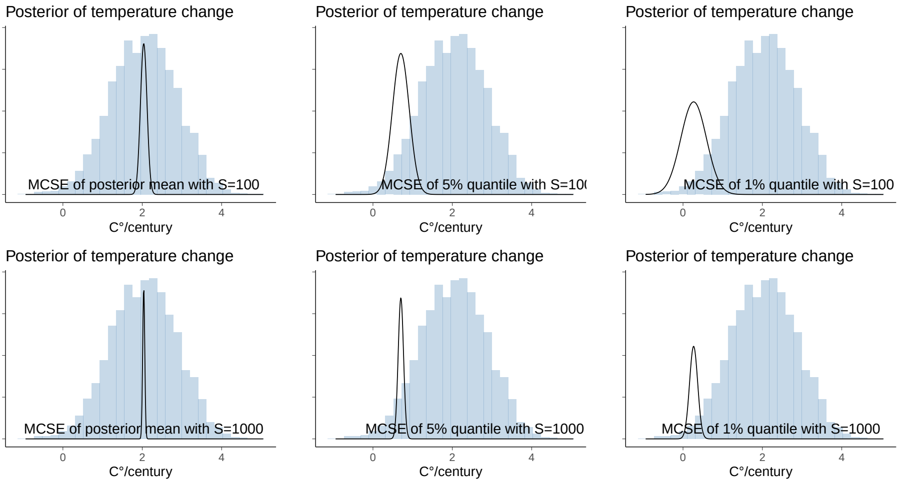




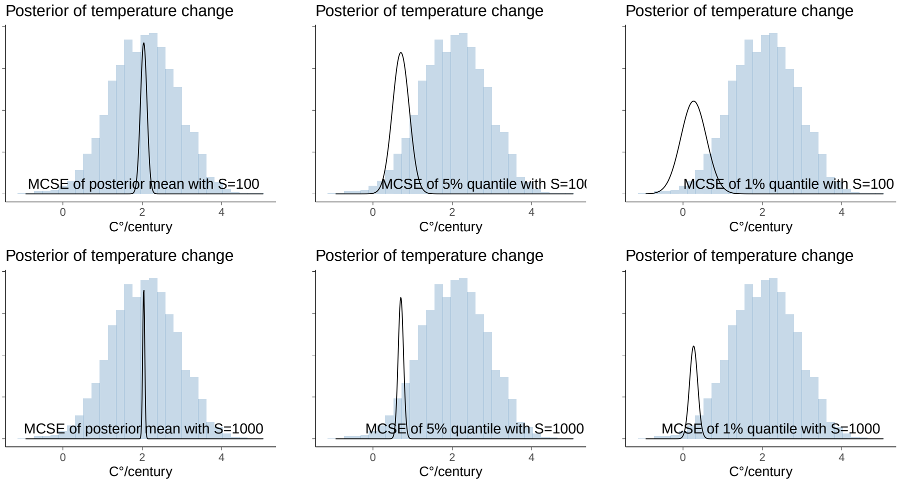
Example: Kilpisjärvi summer temperature



Example: Kilpisjärvi summer temperature



Example: Kilpisjärvi summer temperature



Tail quantiles are more difficult to estimate

How many simulation draws are needed?

- Posterior probability

$$p(\theta \in A) \approx \frac{1}{S} \sum_{l} I(\theta^{(s)} \in A)$$

where $I(\theta^{(s)} \in A) = 1$ if $\theta^{(s)} \in A$

- $I(\cdot)$ is binomially distributed as $p(\theta \in A)$
  - $\rightarrow$ $\mathrm{var}(I(\cdot)) = p(1-p)$ (Appendix A, p. 579)
  - $\rightarrow$ standard deviation of p is $\sqrt{p(1-p)/S}$

How many simulation draws are needed?

- Posterior probability

$$p(\theta \in A) \approx \frac{1}{S} \sum_{l} I(\theta^{(s)} \in A)$$

where $I(\theta^{(s)} \in A) = 1$ if $\theta^{(s)} \in A$

  - $I(\cdot)$ is binomially distributed as $p(\theta \in A)$
    - $\rightarrow$ $\mathrm{var}(I(\cdot)) = p(1-p)$ (Appendix A, p. 579)
    - $\rightarrow$ standard deviation of p is $\sqrt{p(1-p)/S}$
  - if $S = 100$ and $p \approx 0.5$, $\sqrt{p(1-p)/S} = 0.05$
    i.e. accuracy is about 5% units

How many simulation draws are needed?

- Posterior probability

$$p(\theta \in A) \approx \frac{1}{S} \sum_{l} I(\theta^{(s)} \in A)$$

where $I(\theta^{(s)} \in A) = 1$ if $\theta^{(s)} \in A$

- $I(\cdot)$ is binomially distributed as $p(\theta \in A)$
  - $\rightarrow$ $\mathrm{var}(I(\cdot)) = p(1-p)$ (Appendix A, p. 579)
  - $\rightarrow$ standard deviation of p is $\sqrt{p(1-p)/S}$
- if $S = 100$ and $p \approx 0.5$, $\sqrt{p(1-p)/S} = 0.05$
  i.e. accuracy is about 5% units
- $S = 2500$ draws needed for 1% unit accuracy

How many simulation draws are needed?

- Posterior probability

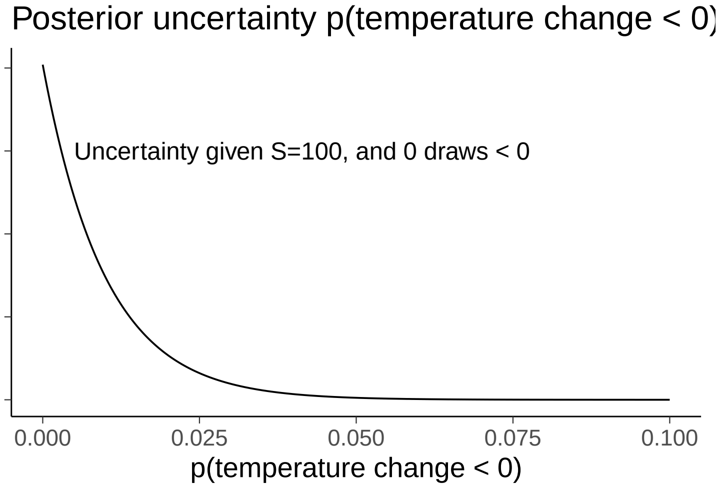
$$p(\theta \in A) \approx \frac{1}{S} \sum_{l} I(\theta^{(s)} \in A)$$

where $I(\theta^{(s)} \in A) = 1$ if $\theta^{(s)} \in A$

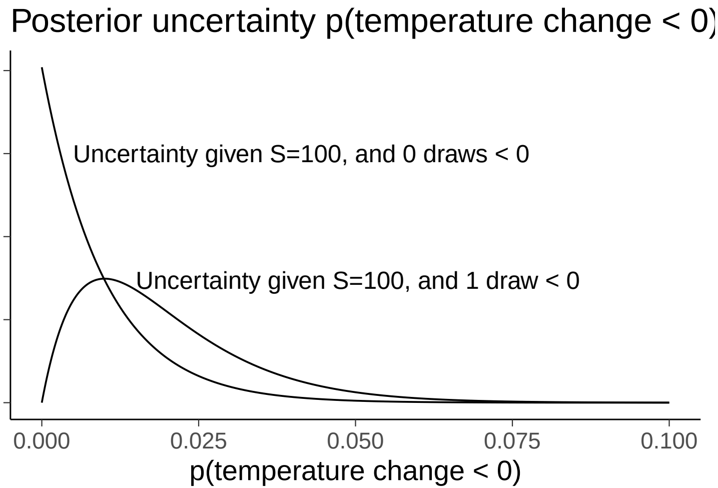
- $I(\cdot)$ is binomially distributed as $p(\theta \in A)$
  - $\rightarrow$ $\mathrm{var}(I(\cdot)) = p(1-p)$ (Appendix A, p. 579)
  - $\rightarrow$ standard deviation of p is $\sqrt{p(1-p)/S}$
- if $S = 100$ and $p \approx 0.5$, $\sqrt{p(1-p)/S} = 0.05$
  i.e. accuracy is about 5% units
- $S = 2500$ draws needed for 1% unit accuracy

- To estimate small probabilities, a large number of draws is needed
  - to be able to estimate p, need to get draws with $\theta^{(l)} \in A$, which in expectation requires $S \gg 1/p$

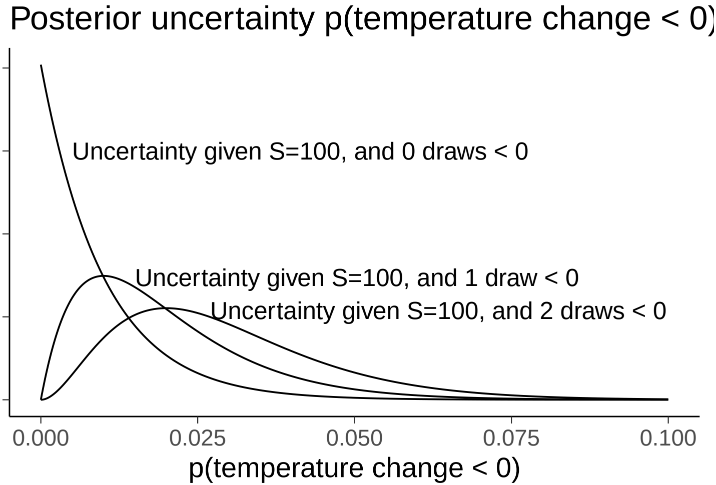
Example: Kilpisjärvi summer temperature



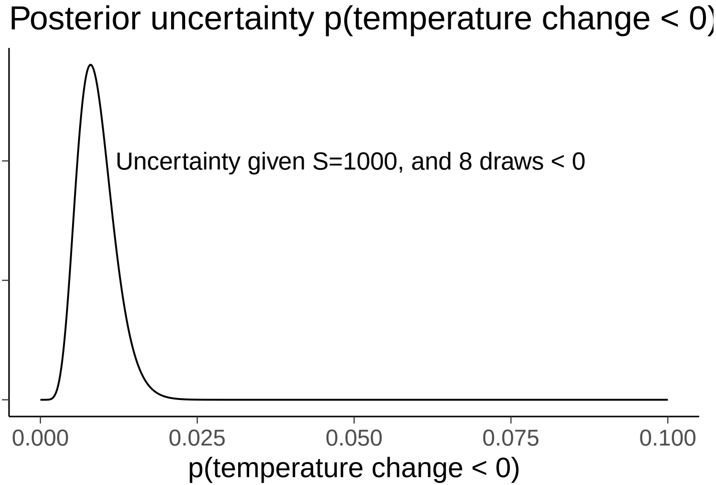
Example: Kilpisjärvi summer temperature



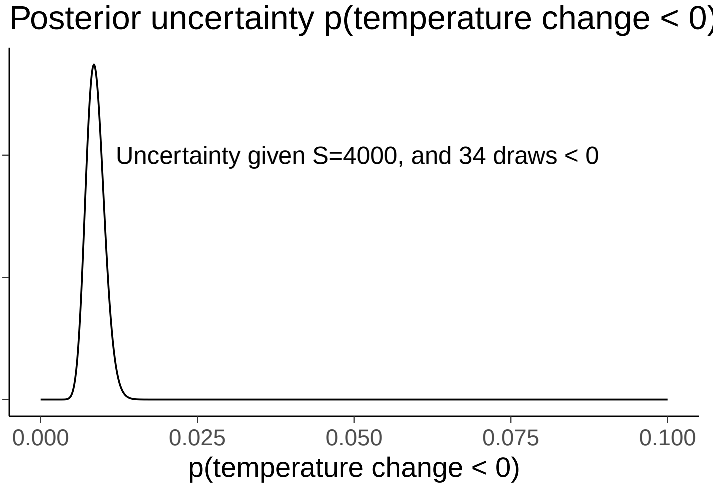
Example: Kilpisjärvi summer temperature



Example: Kilpisjärvi summer temperature



Example: Kilpisjärvi summer temperature



How many digits to show in reports?

- Too many digits make reading of the results slower and give false impression of the accuracy

How many digits to show in reports?

- Too many digits make reading of the results slower and give false impression of the accuracy
- Don’t show digits which are just random noise
  - check what is the Monte Carlo standard error

How many digits to show in reports?

- Too many digits make reading of the results slower and give false impression of the accuracy
- Don't show digits which are just random noise
  - check what is the Monte Carlo standard error
- Show meaningful digits given the posterior uncertainty

How many digits to show in reports?

- Too many digits make reading of the results slower and give false impression of the accuracy
- Don't show digits which are just random noise
  - check what is the Monte Carlo standard error
- Show meaningful digits given the posterior uncertainty
- Example: The mean and 90% central posterior interval for temperature increase C$^\circ$/century based on posterior draws

How many digits to show in reports?

- Too many digits make reading of the results slower and give false impression of the accuracy
- Don't show digits which are just random noise
  - check what is the Monte Carlo standard error
- Show meaningful digits given the posterior uncertainty
- Example: The mean and 90% central posterior interval for temperature increase C$^\circ$/century based on posterior draws
  - 2.050774 and [0.7472868 3.3017524] (NO!)

How many digits to show in reports?

- Too many digits make reading of the results slower and give false impression of the accuracy
- Don't show digits which are just random noise
  - check what is the Monte Carlo standard error
- Show meaningful digits given the posterior uncertainty
- Example: The mean and 90% central posterior interval for temperature increase C°/century based on posterior draws
  - 2.050774 and [0.7472868 3.3017524] (NO!)
  - 2.1 and [0.7 3.3]

How many digits to show in reports?

- Too many digits make reading of the results slower and give false impression of the accuracy
- Don't show digits which are just random noise
  - check what is the Monte Carlo standard error
- Show meaningful digits given the posterior uncertainty
- Example: The mean and 90% central posterior interval for temperature increase C$^\circ$/century based on posterior draws
  - 2.050774 and [0.7472868 3.3017524] (NO!)
  - 2.1 and [0.7 3.3]
  - 2 and [1 3] (depends on the context)

How many digits to show in reports?

- Too many digits make reading of the results slower and give false impression of the accuracy
- Don't show digits which are just random noise
  - check what is the Monte Carlo standard error
- Show meaningful digits given the posterior uncertainty
- Example: The mean and 90% central posterior interval for temperature increase C$^\circ$/century based on posterior draws
  - 2.050774 and [0.7472868 3.3017524] (NO!)
  - 2.1 and [0.7 3.3]
  - 2 and [1 3] (depends on the context)
- Example: The probability that temp increase is positive

How many digits to show in reports?

- Too many digits make reading of the results slower and give false impression of the accuracy
- Don't show digits which are just random noise
  - check what is the Monte Carlo standard error
- Show meaningful digits given the posterior uncertainty
- Example: The mean and 90% central posterior interval for temperature increase C°/century based on posterior draws
  - 2.050774 and [0.7472868 3.3017524] (NO!)
  - 2.1 and [0.7 3.3]
  - 2 and [1 3] (depends on the context)
- Example: The probability that temp increase is positive
  - 0.9960000 (NO!)

How many digits to show in reports?

- Too many digits make reading of the results slower and give false impression of the accuracy
- Don't show digits which are just random noise
  - check what is the Monte Carlo standard error
- Show meaningful digits given the posterior uncertainty
- Example: The mean and 90% central posterior interval for temperature increase C°/century based on posterior draws
  - 2.050774 and [0.7472868 3.3017524] (NO!)
  - 2.1 and [0.7 3.3]
  - 2 and [1 3] (depends on the context)
- Example: The probability that temp increase is positive
  - 0.9960000 (NO!)
  - 1.00 (depends on the context)

How many digits to show in reports?

- Too many digits make reading of the results slower and give false impression of the accuracy
- Don't show digits which are just random noise
  - check what is the Monte Carlo standard error
- Show meaningful digits given the posterior uncertainty
- Example: The mean and 90% central posterior interval for temperature increase C°/century based on posterior draws
  - 2.050774 and [0.7472868 3.3017524] (NO!)
  - 2.1 and [0.7 3.3]
  - 2 and [1 3] (depends on the context)
- Example: The probability that temp increase is positive
  - 0.9960000 (NO!)
  - 1.00 (depends on the context)
  - With 4000 draws MCSE $\approx$ 0.002. We could report that probability is very likely larger than 0.99, or sample more to justify reporting three digits

How many digits to show in reports?

- Too many digits make reading of the results slower and give false impression of the accuracy
- Don't show digits which are just random noise
  - check what is the Monte Carlo standard error
- Show meaningful digits given the posterior uncertainty
- Example: The mean and 90% central posterior interval for temperature increase C°/century based on posterior draws
  - 2.050774 and [0.7472868 3.3017524] (NO!)
  - 2.1 and [0.7 3.3]
  - 2 and [1 3] (depends on the context)
- Example: The probability that temp increase is positive
  - 0.9960000 (NO!)
  - 1.00 (depends on the context)
  - With 4000 draws MCSE $\approx$ 0.002. We could report that probability is very likely larger than 0.99, or sample more to justify reporting three digits
  - For probabilities close to 0 or 1, consider also when the model assumption justify certain accuracy

How many digits to show in reports?

- Too many digits make reading of the results slower and give false impression of the accuracy
- Don't show digits which are just random noise
  - check what is the Monte Carlo standard error
- Show meaningful digits given the posterior uncertainty
- Example: The mean and 90% central posterior interval for temperature increase C°/century based on posterior draws
  - 2.050774 and [0.7472868 3.3017524] (NO!)
  - 2.1 and [0.7 3.3]
  - 2 and [1 3] (depends on the context)
- Example: The probability that temp increase is positive
  - 0.9960000 (NO!)
  - 1.00 (depends on the context)
  - With 4000 draws MCSE $\approx$ 0.002. We could report that probability is very likely larger than 0.99, or sample more to justify reporting three digits
  - For probabilities close to 0 or 1, consider also when the model assumption justify certain accuracy

See also https://avehtari.github.io/casestudies/Digits/digits.html

More data

- The analysis I just showed used data from 1952–2013

More data

- The analysis I just showed used data from 1952–2013
- With data data from 1952–2022
    - The probability that temp increase is positive:
      0.9995 ± 0.000353,
      which can be reported as more than 99.9% probability
    - With data from other locations we would be even more certain

How many simulation draws are needed?

- Less draws needed with
  - deterministic methods
  - marginalization (Rao-Blackwellization)
  - variance reduction methods, such, control variates

How many simulation draws are needed?

- Number of independent draws needed doesn't depend on the number of dimensions
  - but it may be difficult to obtain independent draws in high dimensional case

Direct simulation

- Produces independent draws
  - Using analytic transformations of uniform random numbers (e.g. appendix A)
  - factorization
  - numerical inverse-CDF
- Problem: restricted to limited set of models

Random number generators

- Good pseudo random number generators are sufficient for Bayesian inference
  - pseudo random generator uses deterministic algorithm to produce a sequence which is difficult to make difference from truly random sequence
  - modern software used for statistical analysis have good pseudo RNGs

Direct simulation: Example

- Box-Muller -method:
  If U_1 and U_2 are independent draws from distribution $\mathrm{U}(0, 1)$, and

$$X_1 = \sqrt{-2\log(U_1)}\cos(2\pi U_2)$$
$$X_2 = \sqrt{-2\log(U_1)}\sin(2\pi U_2)$$

  then X_1 and X_2 are independent draws from the distribution $\mathrm{N}(0, 1)$

Direct simulation: Example

- Box-Muller -method:
  If U_1 and U_2 are independent draws from distribution $\mathrm{U}(0, 1)$, and

$$X_1 = \sqrt{-2\log(U_1)}\cos(2\pi U_2)$$
$$X_2 = \sqrt{-2\log(U_1)}\sin(2\pi U_2)$$

  then X_1 and X_2 are independent draws from the distribution $\mathrm{N}(0, 1)$
    - not the fastest method due to trigonometric computations
    - for normal distribution more than ten different methods
    - e.g. R uses inverse-CDF

Grid sampling and curse of dimensionality

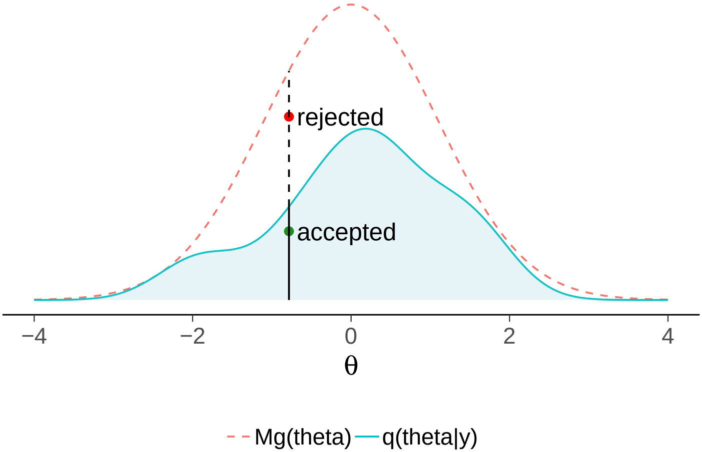
- 10 parameters
- if we don't know beforehand where the posterior mass is
  - need to choose wide box for the grid
  - need to have enough grid points to get some of them where essential mass is
- e.g. 50 or 1000 grid points per dimension
  - $\rightarrow$ $50^{10} \approx$ 1e17 grid points
  - $\rightarrow$ $1000^{10} \approx$ 1e30 grid points
- R and my current laptop can compute density of normal distribution about 50 million times per second
  - $\rightarrow$ evaluation in 1e17 grid points would take 60 years
  - $\rightarrow$ evaluation in 1e30 grid points would take 600 billion years

Indirect sampling

- Rejection sampling
- Importance sampling
- Markov chain Monte Carlo (next week)

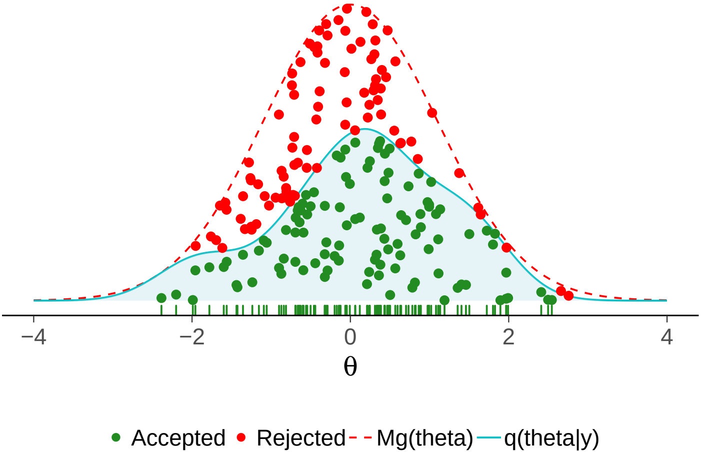
Rejection sampling

- Proposal forms envelope over the target distribution $q(\theta|y)/Mg(\theta) \leq 1$
- Draw from the proposal and accept with probability $q(\theta|y)/Mg(\theta)$



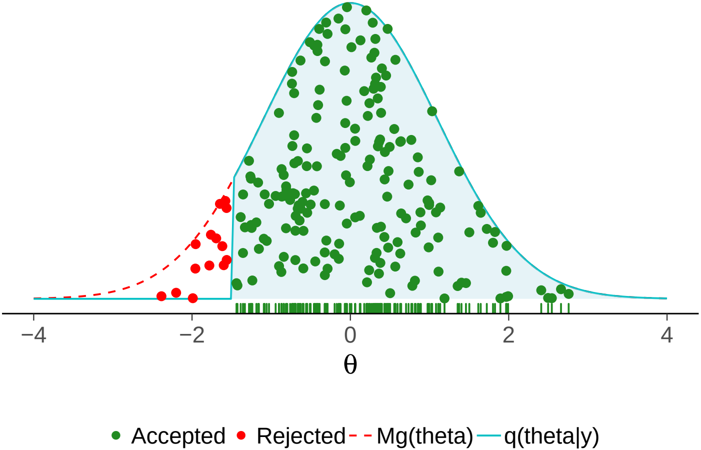
Rejection sampling

- Proposal forms envelope over the target distribution $q(\theta|y)/Mg(\theta) \leq 1$
- Draw from the proposal and accept with probability $q(\theta|y)/Mg(\theta)$



Rejection sampling

- Proposal forms envelope over the target distribution $q(\theta|y)/Mg(\theta) \leq 1$
- Draw from the proposal and accept with probability $q(\theta|y)/Mg(\theta)$
- Common for truncated distributions

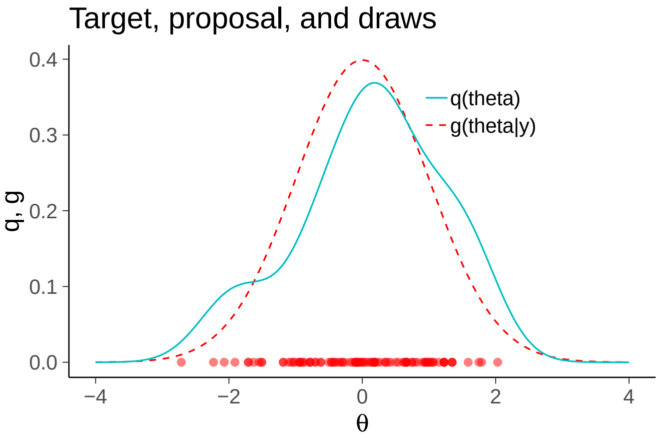


Rejection sampling

- The number of accepted draws is the effective sample size
  - with bad proposal distribution may require a lot of trials
  - selection of good proposal gets very difficult when the number of dimensions increase
  - reliable diagnostics and thus can be a useful part

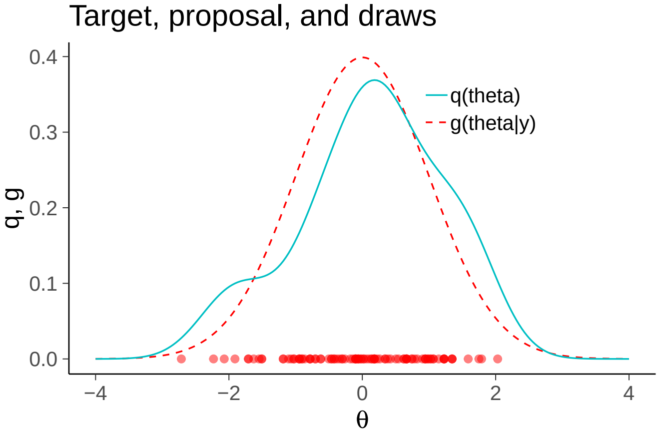
Importance sampling

- Proposal does not need to have a higher value everywhere



Importance sampling

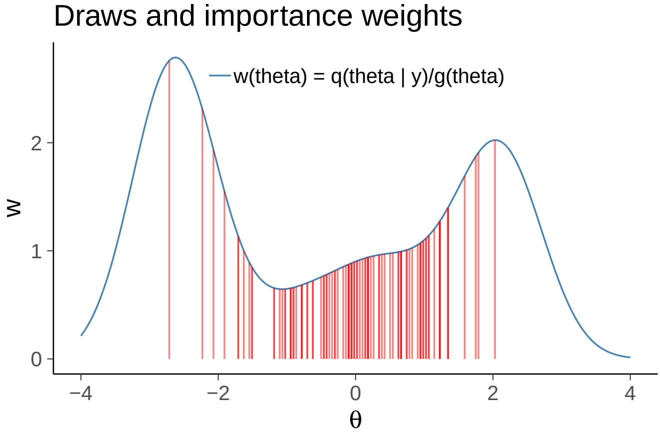
- Proposal does not need to have a higher value everywhere



$$E[f(\theta)] \approx \frac{\sum_s w_s f(\theta^{(s)})}{\sum_s w_s}, \qquad \text{where} \quad w_s = \frac{q(\theta^{(s)})}{g(\theta^{(s)})}$$

Importance sampling

- Proposal does not need to have a higher value everywhere



$$E[f(\theta)] \approx \frac{\sum_s w_s f(\theta^{(s)})}{\sum_s w_s}, \qquad \text{where} \quad w_s = \frac{q(\theta^{(s)})}{g(\theta^{(s)})}$$

Some uses of importance sampling

In general selection of good proposal gets more difficult when the number of dimensions increase, but there are many use case

Some uses of importance sampling

In general selection of good proposal gets more difficult when the number of dimensions increase, but there are many use case

- Fast leave-one-out cross-validation
- Fast bootstrapping
- Fast prior and likelihood sensitivity analysis
- Conformal Bayesian computation
- Particle filtering
- Improving distributional approximations (e.g Laplace, VI)

IS finite variance and central limit theorem

- If $h(\theta)w$ and w have finite variance $\rightarrow$ CLT
  - variance goes down as $1/S$
  - Effective sample size (ESS) takes into account the variability in the weights

IS finite variance and central limit theorem

- If $h(\theta)w$ and w have finite variance $\rightarrow$ CLT
  - variance goes down as $1/S$
  - Effective sample size (ESS) takes into account the variability in the weights
- We would like to have finite variance and CLT
  - sometimes these can be guaranteed by construction, e.g., by choosing $g(\theta)$ so that $w(\theta)$ is bounded
  - generally not trivial

IS finite variance and central limit theorem

- If $h(\theta)w$ and w have finite variance $\rightarrow$ CLT
  - variance goes down as $1/S$
  - Effective sample size (ESS) takes into account the variability in the weights
- We would like to have finite variance and CLT
  - sometimes these can be guaranteed by construction, e.g., by choosing $g(\theta)$ so that $w(\theta)$ is bounded
  - generally not trivial
- If variance is infinite, but mean is finite
  $\rightarrow$ *generalized CLT and asymptotic consistency*

IS finite variance and central limit theorem

- If $h(\theta)w$ and w have finite variance $\rightarrow$ CLT
  - variance goes down as $1/S$
  - Effective sample size (ESS) takes into account the variability in the weights
- We would like to have finite variance and CLT
  - sometimes these can be guaranteed by construction, e.g., by choosing $g(\theta)$ so that $w(\theta)$ is bounded
  - generally not trivial
- If variance is infinite, but mean is finite
  $\rightarrow$ *generalized CLT and asymptotic consistency*
- Pre-asymptotic and asymptotic behavior can be really different!

Importance re-sampling

- Using the weighted draws is good

$$\mathrm{E}[f(\theta)] \approx \frac{\sum_s w_s f(\theta^{(s)})}{\sum_s w_s}$$

Importance re-sampling

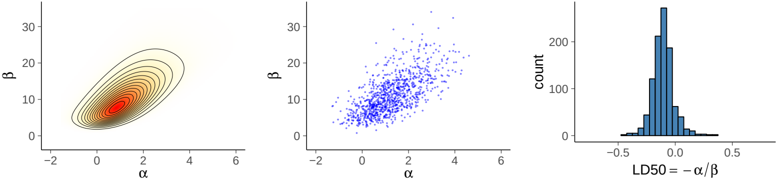
- Using the weighted draws is good

$$\mathrm{E}[f(\theta)] \approx \frac{\sum_s w_s f(\theta^{(s)})}{\sum_s w_s}$$

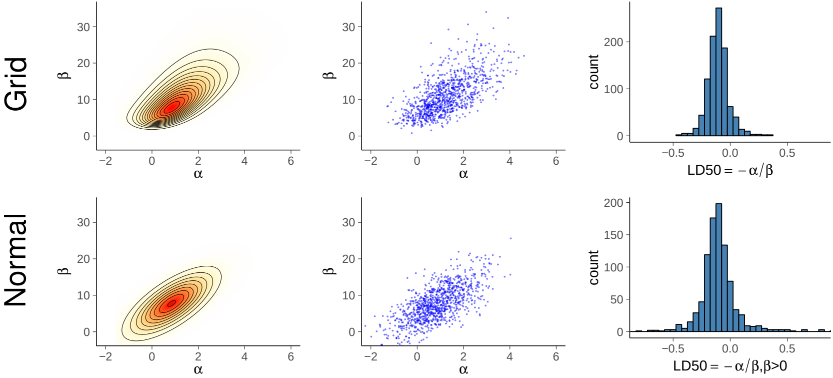
- But it can be convenient to obtain draws with equal weights
  - resample the draws according to the weights
  - some original draws may be included more than once
  - loses some information, but now the weights are equal

Example: Importance sampling in Bioassay

Grid

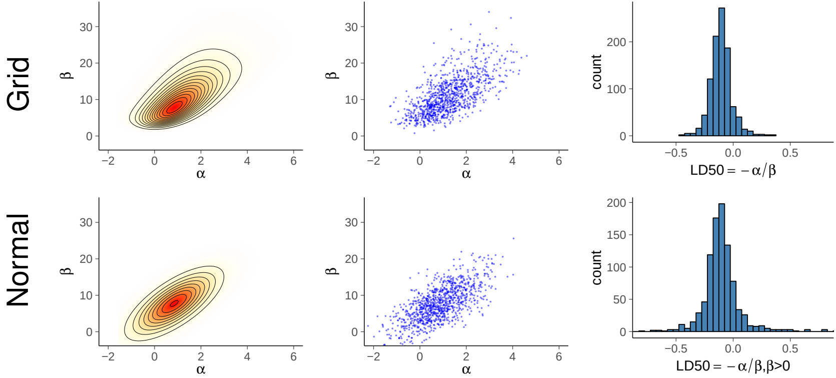


Example: Importance sampling in Bioassay



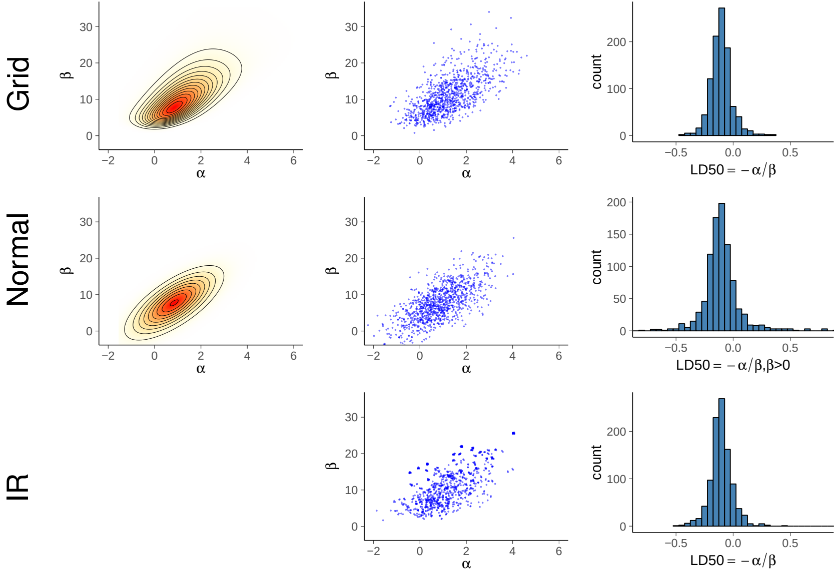
Normal approximation is discussed more in BDA3 Ch 4

Example: Importance sampling in Bioassay

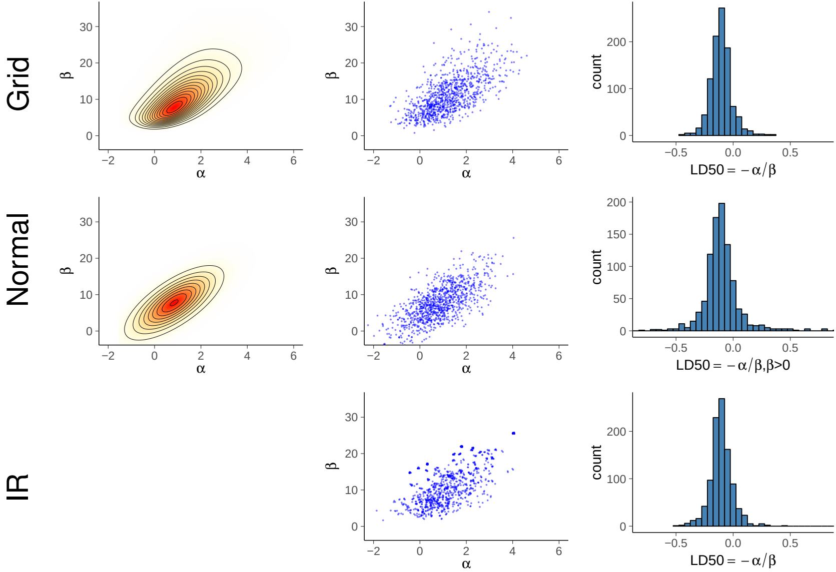


Normal approximation is discussed more in BDA3 Ch 4
But the normal approximation is not that good here:
Grid sd(LD50) $\approx$ 0.1, Normal sd(LD50) $\approx$ .75!

Example: Importance sampling in Bioassay

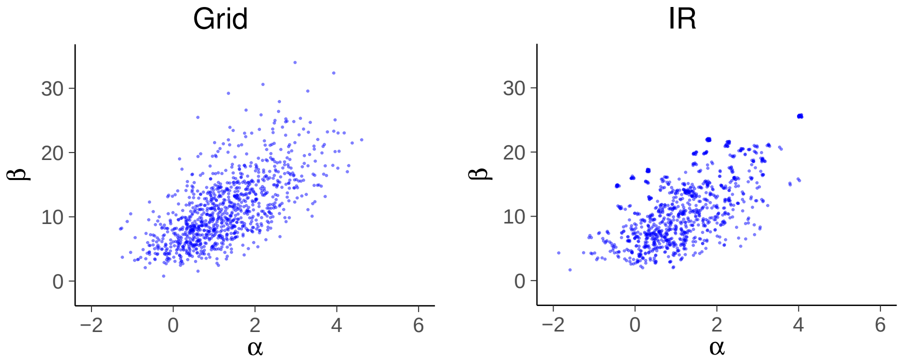


Example: Importance sampling in Bioassay

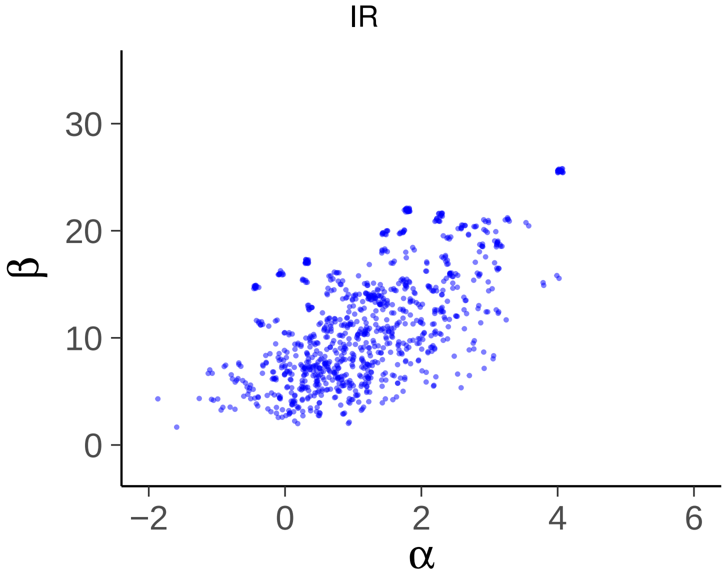


Grid sd(LD50) $\approx$ 0.1, IR sd(LD50) $\approx$ 0.1

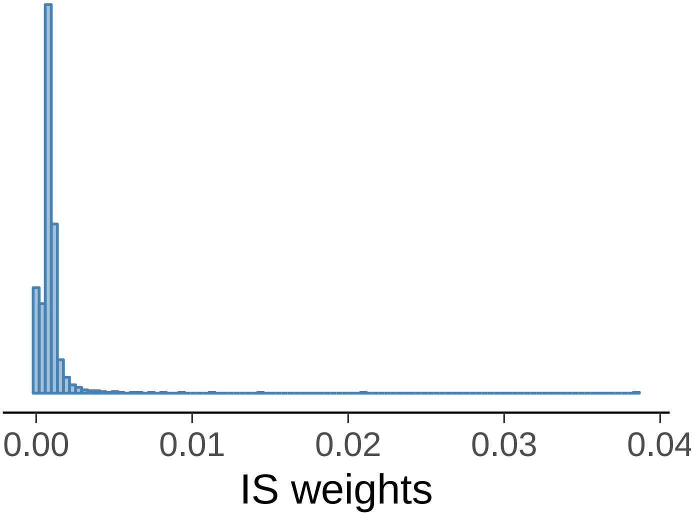
Example: Importance sampling in Bioassay



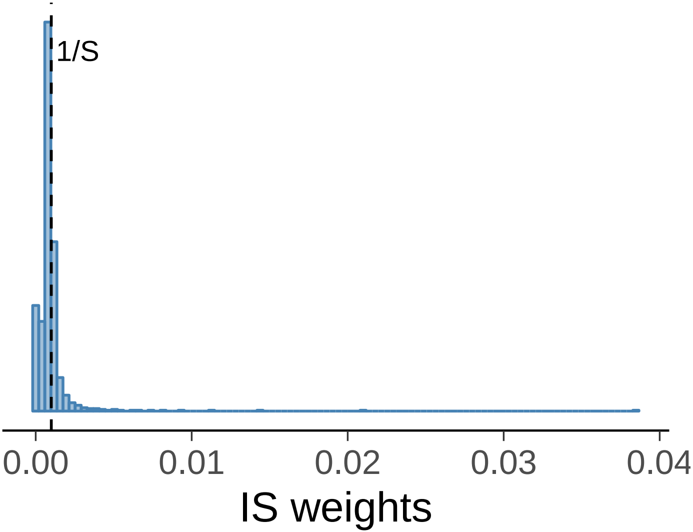
Example: Importance sampling in Bioassay



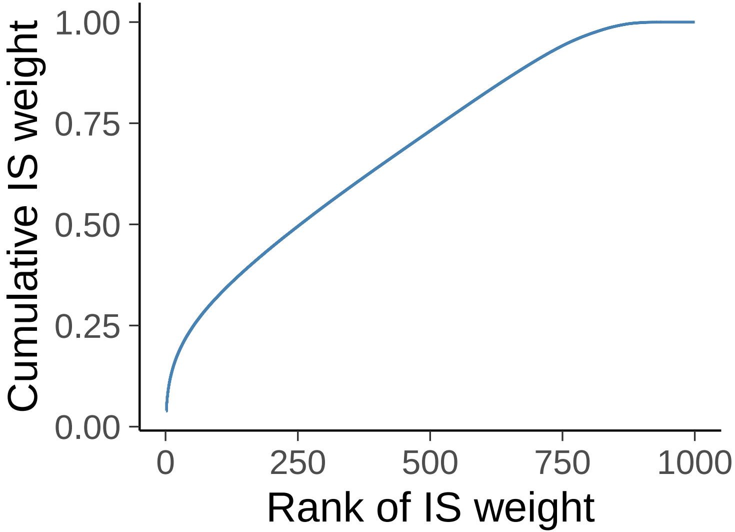
Example: Importance sampling in Bioassay



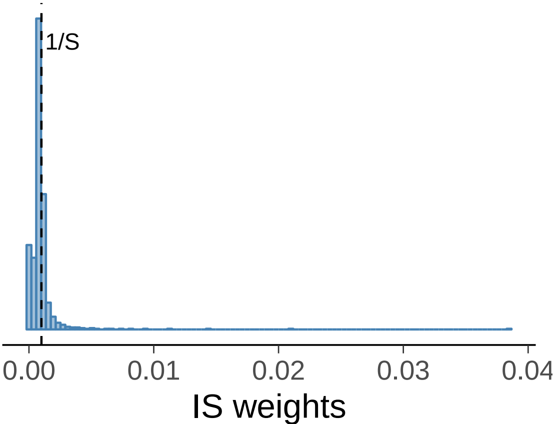
Example: Importance sampling in Bioassay



Example: Importance sampling in Bioassay

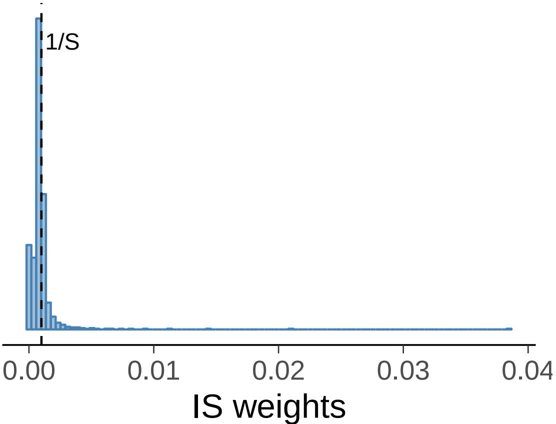


Example: Importance sampling in Bioassay



$$S_{\mathrm{eff}} = \frac{1}{\sum_{s=1}^{S}(\tilde{w}(\theta^s))^2}, \quad \text{where } \tilde{w}(\theta^s) = w(\theta^s) / \sum_{s'=1}^{S} w(\theta^{s'})$$

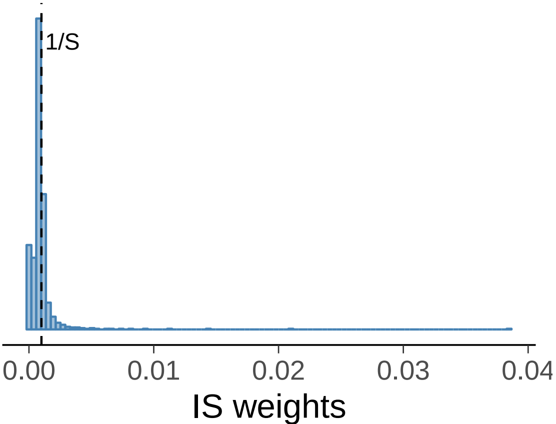
Example: Importance sampling in Bioassay



$$S_{\text{eff}} = \frac{1}{\sum_{s=1}^{S} (\tilde{w}(\theta^s))^2}, \quad \text{where } \tilde{w}(\theta^s) = w(\theta^s) / \sum_{s'=1}^{S} w(\theta^{s'})$$

BDA3 1st (2013) and 2nd (2014) printing have an error for $\tilde{w}(\theta^s)$. The equation should not have the multiplier S (the normalized weights should sum to one). Online version is correct. Errata for the book http://www.stat.columbia.edu/~gelman/book/errata_bda3.txt

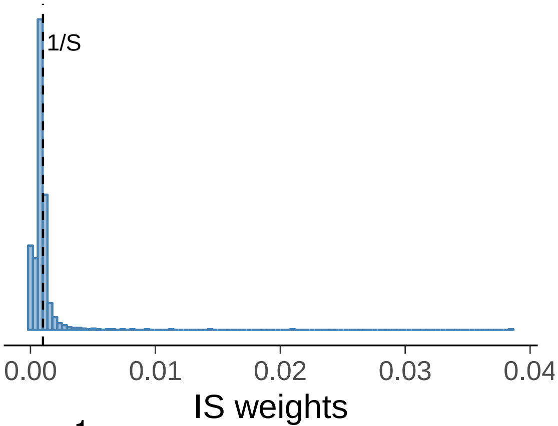
Example: Importance sampling in Bioassay



$$S_{\text{eff}} = \frac{1}{\sum_{s=1}^{S} (\tilde{w}(\theta^s))^2}, \quad \text{where } \tilde{w}(\theta^s) = w(\theta^s) / \sum_{s'=1}^{S} w(\theta^{s'})$$

$S_{\text{eff}} \approx 396$

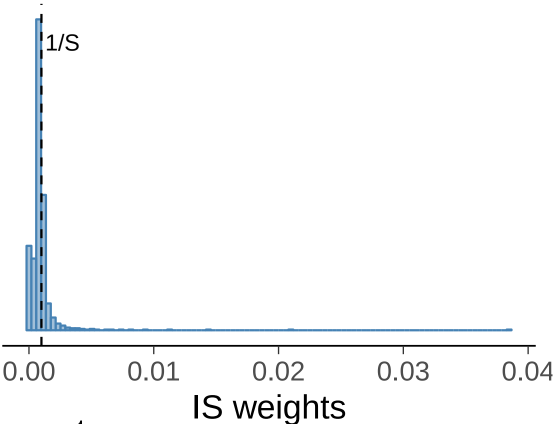
Example: Importance sampling in Bioassay



$$S_{\mathrm{eff}} = \frac{1}{\sum_{s=1}^{S}(\tilde{w}(\theta^s))^2},$$ is based on variance of $\tilde{w}(\theta^s)$

$S_{\mathrm{eff}} \approx 396$

Example: Importance sampling in Bioassay

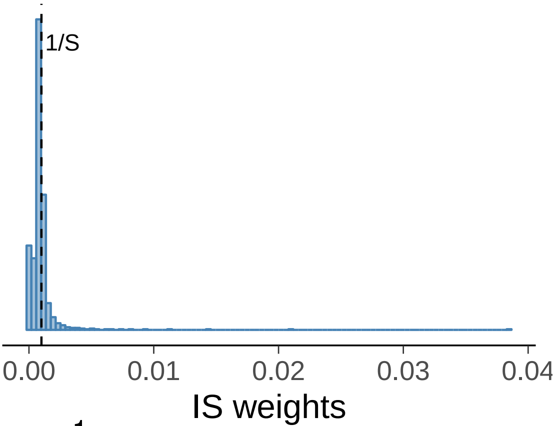


$$S_{\text{eff}} = \frac{1}{\sum_{s=1}^{S} (\tilde{w}(\theta^s))^2}, \quad \text{is based on variance of } \tilde{w}(\theta^s)$$

$$S_{\text{eff}} \approx 396$$

If all $\tilde{w}(\theta^s) = 1/S$, then $S_{\text{eff}} = 1/(SS^{-2}) = S$

Example: Importance sampling in Bioassay

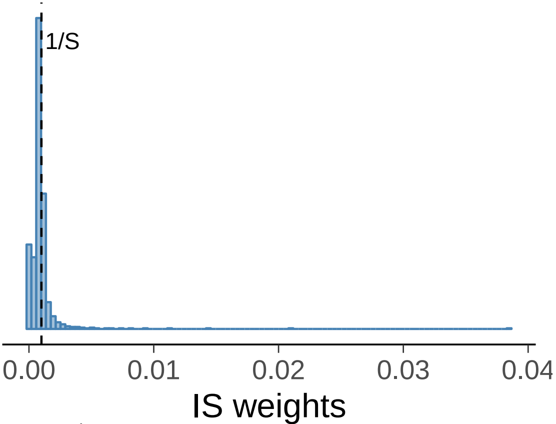


$$S_{\mathrm{eff}} = \frac{1}{\sum_{s=1}^{S} (\tilde{w}(\theta^s))^2}, \quad \text{is based on variance of } \tilde{w}(\theta^s)$$

$$S_{\mathrm{eff}} \approx 396$$

If all $\tilde{w}(\theta^s) = 1/S$, then $S_{\mathrm{eff}} = 1/(SS^{-2}) = S$
If one $\tilde{w}(\theta^s) = 1$, and others 0, then $S_{\mathrm{eff}} = 1/1 = 1$

Example: Importance sampling in Bioassay

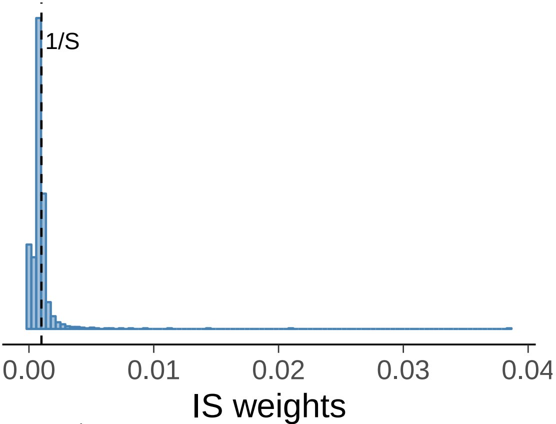


$$S_{\text{eff}} = \frac{1}{\sum_{s=1}^{S}(\tilde{w}(\theta^s))^2}, \quad \text{is based on variance of } \tilde{w}(\theta^s)$$

$S_{\text{eff}} \approx 396$

Pareto-k diagnostic preferably < 0.7:

Example: Importance sampling in Bioassay



$$S_{\mathrm{eff}} = \frac{1}{\sum_{s=1}^{S} (\tilde{w}(\theta^s))^2}, \quad \text{is based on variance of } \tilde{w}(\theta^s)$$

$S_{\mathrm{eff}} \approx 396$

Pareto-k diagnostic preferably < 0.7: $\hat{k} \approx 0.65$

Pareto-$\hat{k}$ diagnostic

- Based on extreme value analysis and generalized central limit theorem

See more in Vehtari, Simpson, Gelman, Yao, and Gabry (2022). Pareto smoothed importance sampling. arXiv:1507.02646.

Pareto-$\hat{k}$ diagnostic

- Based on extreme value analysis and generalized central limit theorem
  - we can estimate tail of a distribution with a Pareto distribution

See more in Vehtari, Simpson, Gelman, Yao, and Gabry (2022). Pareto smoothed importance sampling. arXiv:1507.02646.

Pareto-$\hat{k}$ diagnostic

- Based on extreme value analysis and generalized central limit theorem
  - we can estimate tail of a distribution with a Pareto distribution
  - shape parameter k tells the number of *fractional moments* as $1/k$
  - estimate $\hat{k}$ from finite data

See more in Vehtari, Simpson, Gelman, Yao, and Gabry (2022). Pareto smoothed importance sampling. arXiv:1507.02646.

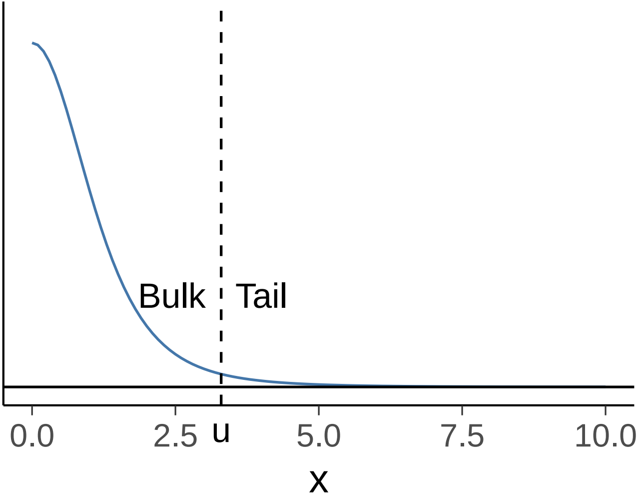
Pareto-$\hat{k}$ diagnostic

- Based on extreme value analysis and generalized central limit theorem
  - we can estimate tail of a distribution with a Pareto distribution
  - shape parameter k tells the number of *fractional moments* as $1/k$
  - estimate $\hat{k}$ from finite data
  - the statistical behavior of distribution of mean can be predicted by generalized CLT
    - minimum sample size and convergence rate given $\hat{k}$

See more in Vehtari, Simpson, Gelman, Yao, and Gabry (2022). Pareto smoothed importance sampling. arXiv:1507.02646.

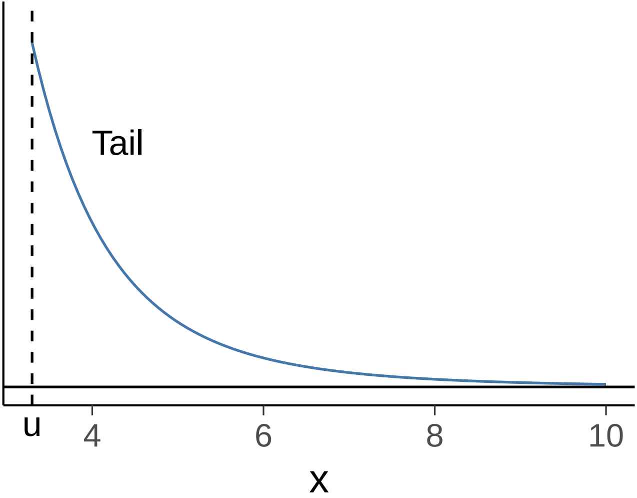
Pareto-$\hat{k}$ diagnostic

Pickands (1975): many distributions have tail ($x > u$) that is well approximated with Generalized Pareto distribution (GPD)



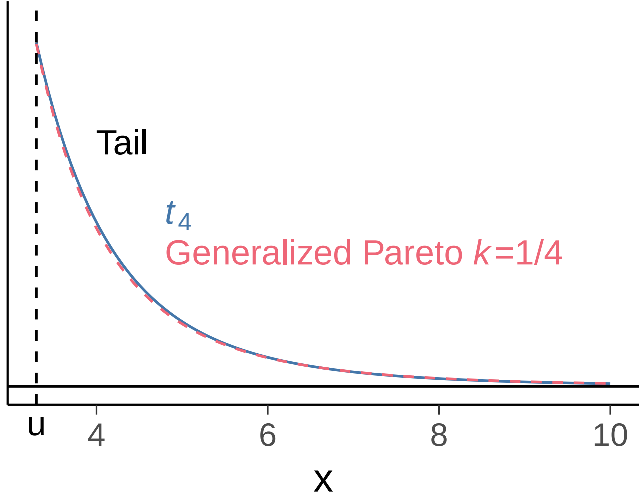
Pareto-$\hat{k}$ diagnostic

Pickands (1975): many distributions have tail ($x > u$) that is well approximated with Generalized Pareto distribution (GPD)



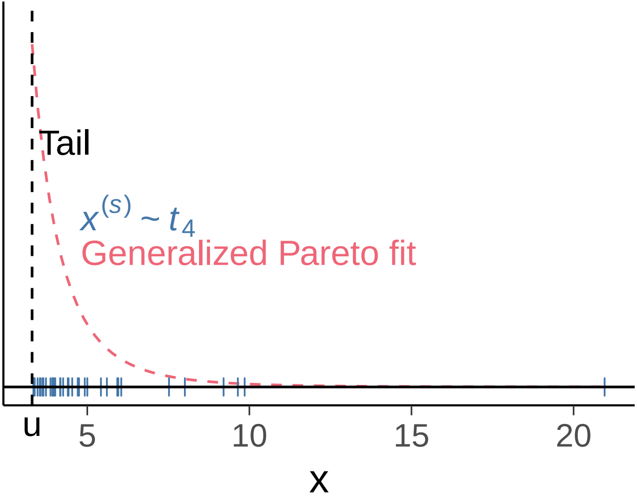
Pareto-$\hat{k}$ diagnostic

Pickands (1975): many distributions have tail ($x > u$) that is well approximated with Generalized Pareto distribution (GPD)



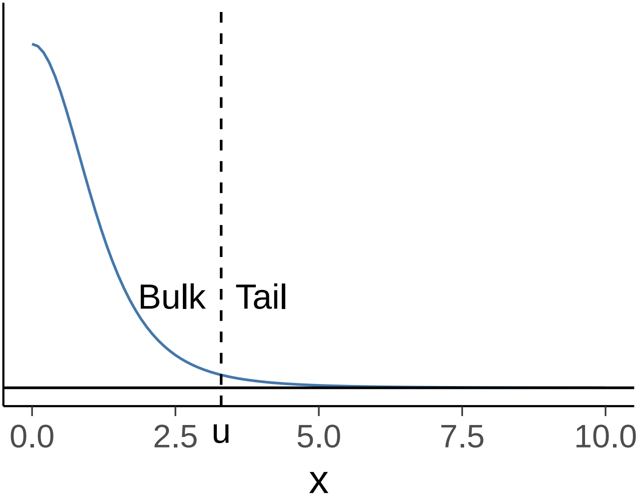
Pareto-$\hat{k}$ diagnostic

Pickands (1975): many distributions have tail ($x > u$) that is well approximated with Generalized Pareto distribution (GPD)

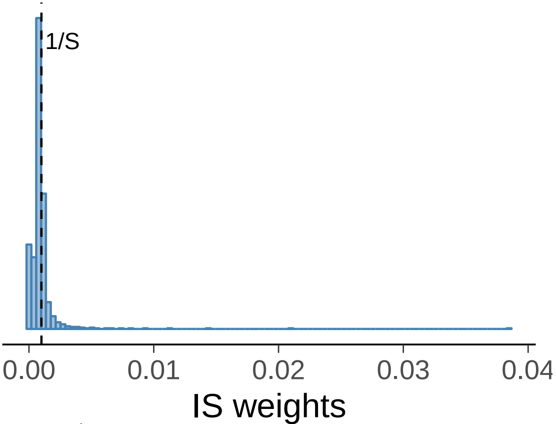


Pareto-$\hat{k}$ diagnostic

GPD has a shape parameter k,
and $1/k$ finite fractional moments



Example: Importance sampling in Bioassay



$$S_{\text{eff}} = \frac{1}{\sum_{s=1}^{S}(\tilde{w}(\theta^s))^2}, \quad \text{is based on variance of } \tilde{w}(\theta^s)$$

$S_{\text{eff}} \approx 396$

Pareto-k diagnostic preferably < 0.7: $\hat{k} \approx 0.65$

Pareto-$\hat{k}$ and convergence rate

- CLT says that to half the MCSE, need 4 times bigger S

Pareto-$\hat{k}$ and convergence rate

- CLT says that to half the MCSE, need 4 times bigger S
- If Pareto-$\hat{k} \approx 0.7$, to half the MCSE, need 10 times bigger S

Pareto-$\hat{k}$ and convergence rate

- CLT says that to half the MCSE, need 4 times bigger S
- If Pareto-$\hat{k} \approx 0.7$, to half the MCSE, need 10 times bigger S
- If Pareto-$\hat{k} > 1$, to half the MCSE, nothing helps

Pareto smoothed importance sampling (PSIS)

- Replace the largest observed ratios with expected ordered statistics of the fitted Pareto distribution
  - corresponds to modeling of the tail, and as usual, modeling reduces the noise

Estimating Pareto-$\hat{k}$

- Fast empirical profile Bayes quadrature estimate by Zhang and Stephens (2009)
  - excellent accuracy compared to exact Bayesian inference
  - see more in Vehtari, Simpson, Gelman, Yao & Gabry (2022)

Pareto-$\hat{k}$ diagnostic use cases

- Importance sampling
  - leave-one-out cross-validation (Vehtari et al., 2016, 2017; Bürkner at al, 2020)
  - Bayesian stacking (Yao et al., 2018, 2021, 2022)
  - leave-future-out cross-validation (Bürkner et al., 2020)
  - Bayesian bootstrap (Paananen et al, 2021, online appendix)
  - prior and likelihood sensitivity analysis (Kallioinen et al., 2021)
  - improving distributional approximations (Yao et al., 2018; Zhang et al., 2021; Dhaka et al., 2021)
  - implicitly adaptive importance sampling (Paananen et al., 2021)
- Stochastic optimization (Dhaka et al., 2020)
- Divergences and gradients in VI (Dhaka et al., 2021)
- MCMC (Paananen et al., 2021)

Importance sampling leave-one-out cross-validation

- Later in the course you will learn how $p(\theta|y)$ can be used as a proposal distribution for $p(\theta|y_{-i})$
  - which allows fast computation of leave-one-out cross-validation

$$p(y_i|y_{-i}) = \int p(y_i|\theta)p(\theta|y_{-i})d\theta$$

Curse of dimensionality

- Number of grid points increases exponentially
- Concentration of the measure, i.e., where is the most of the mass?

Markov chain Monte Carlo (MCMC)

- Pros
  - Markov chain goes where most of the posterior mass is
  - Certain MCMC methods scale well to high dimensions
- Cons
  - Draws are dependent (affects how many draws are needed)
  - Convergence in practical time is not guaranteed
- MCMC methods in this course
  - Gibbs: “iterative conditional sampling”
  - Metropolis: “random walk in joint distribution”
  - Dynamic Hamiltonian Monte Carlo: “state-of-the-art” used in Stan