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 (Lecture 9) and prior sensitivity analysis (Lecture ?)
- 10.5 **How many simulation draws are needed?**
  - **see chapter notes for how many significant digits to report**
  - **this week focus on independent draws and importance sampling, next week necessary adjustments needed for Markov chain Monte Carlo**
- 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)) $\longrightarrow$ 0 (underflow)
    - sum(dnorm(qr, log=TRUE)) $\longrightarrow$ -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)) $\longrightarrow$ 0 (underflow)
    - sum(dnorm(qr, log=TRUE)) $\longrightarrow$ -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. $\log(\exp(800) + \exp(800)) \longrightarrow$ Inf

Numerical accuracy – log scale

- Log densities
  - use log densities to avoid over- and underflows in floating point presentation
    - prod(dnorm(qr)) $\longrightarrow$ 0 (underflow)
    - sum(dnorm(qr, log=TRUE)) $\longrightarrow$ -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. $\log(\exp(800) + \exp(800)) \longrightarrow$ 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)) $\longrightarrow$ 0 (underflow)
    - sum(dnorm(qr, log=TRUE)) $\longrightarrow$ -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. $\log(\exp(800) + \exp(800))$ $\longrightarrow$ 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
      $\log(a/b) = \log(a) - \log(b)$

Numerical accuracy – log scale

- Log densities
  - use log densities to avoid over- and underflows in floating point presentation
    - prod(dnorm(qr)) $\longrightarrow$ 0 (underflow)
    - sum(dnorm(qr, log=TRUE)) $\longrightarrow$ -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. $\log(\exp(800) + \exp(800)) \longrightarrow$ 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 $\log(a/b) = \log(a) - \log(b)$
  - convenience functions
    - `matrixStats::logSumExp(lx)` computes `log(sum(exp(lx)))` using the above rule
    - `log1p(x)` computes `log(1+x)` accurately also for $|x| \ll 1$
    - `expm1(x)` computes `exp(x) - 1` accurately also for $|x| \ll 1$

It's all about expectations

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

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

It's all about expectations

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

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

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

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

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

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

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

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

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 (lecture 3)

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

It's all about expectations

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

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

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 (lecture 3)

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

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

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

It's all about expectations

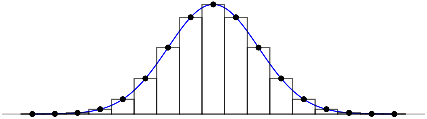
$$E_\theta[h(\theta)] = \int h(\theta)p(\theta|y)d\theta$$

- Conjugate priors and analytic solutions (Ch 1-5, Lec 2–3)
- Grid integration and other quadrature rules (Ch 3, 10, Lec 3–4)
- Independent Monte Carlo, rejection and importance sampling (Ch 10, Lec 4)
- Markov Chain Monte Carlo (Ch 11-12, Lec 5–6)
- Distributional approximations (Laplace, VB, EP) (Ch 4, 13)

Quadrature integration

- The simplest quadrature integration is grid integration

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

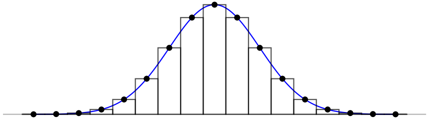


where $w^{(t)}$ is the normalized probability of a grid cell t, and $\alpha^{(t)}$ and $\beta^{(t)}$ are center locations of grid cells

Quadrature integration

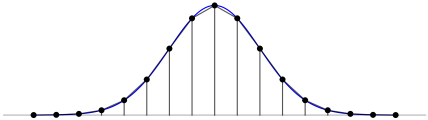
- The simplest quadrature integration is grid integration

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



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

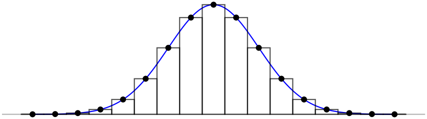
- In 1D further variations with better accuracy, e.g. trapezoid



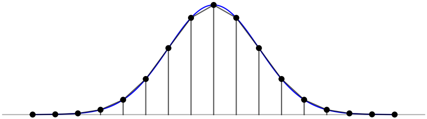
Quadrature integration

- 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

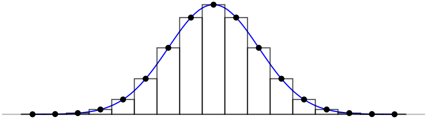


- Adaptive quadrature methods add evaluation points where needed, e.g., R function `integrate()`

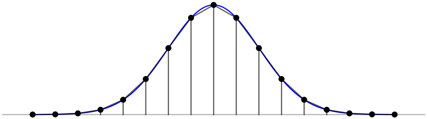
Quadrature integration

- 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

Grid sampling and curse of dimensionality

- In general the number of evaluations increase exponentially c^D
- 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

Grid sampling and curse of dimensionality

- In general the number of evaluations increase exponentially c^D
- 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, and 10 dimensions
  - $\rightarrow$ $50^{10} \approx$ 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

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

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
  - JAGS, Nimble, Tensorflow probability, PyMC, Pyro, BlackJAX Turing.jl, ...
  - Štrumbelj et al. (2024). Past, Present, and Future of Software for Bayesian Inference. *Statistical Science*, 39(1):46-61. https://doi.org/10.1214/23-STS907

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 for the same accuracy)

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
  - next week

How many simulation draws are needed?

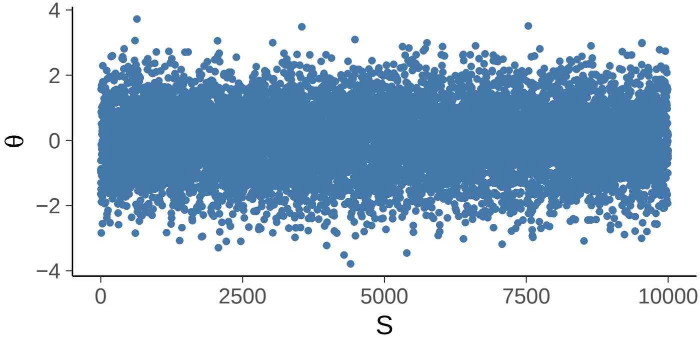
- Expectation of unknown quantity $\mathrm{E}(\theta) \approx \frac{1}{S} \sum_{s=1}^{S} \theta^{(s)}$
  - If S is big,
  - $\theta^{(s)}$ are independent,
  - $p(\theta)$ has finite variance,

  then the central limit theorem (CLT) states that the distribution of the expectation approaches normal distribution (see BDA3 Ch 4) with variance σ_θ^2/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,
  - $\theta^{(s)}$ are independent,
  - $p(\theta)$ has finite variance,

  then the central limit theorem (CLT) states that the distribution of the expectation approaches normal distribution (see BDA3 Ch 4) with variance σ_θ^2/S

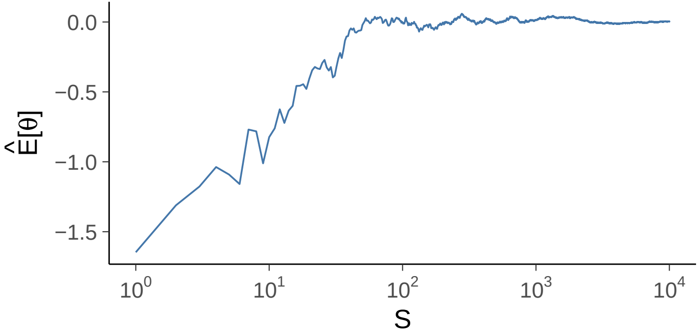



```
rnorm(n=10000, mean=0, sd=1)
```

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,
  - $\theta^{(s)}$ are independent,
  - $p(\theta)$ has finite variance,

  then the central limit theorem (CLT) states that the distribution of the expectation approaches normal distribution (see BDA3 Ch 4) with variance σ_θ^2/S

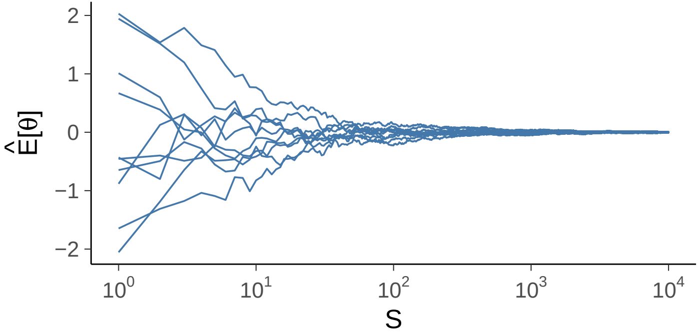



```
cummean(rnorm(n=10000, mean=0, sd=1))
```

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,
  - $\theta^{(s)}$ are independent,
  - $p(\theta)$ has finite variance,

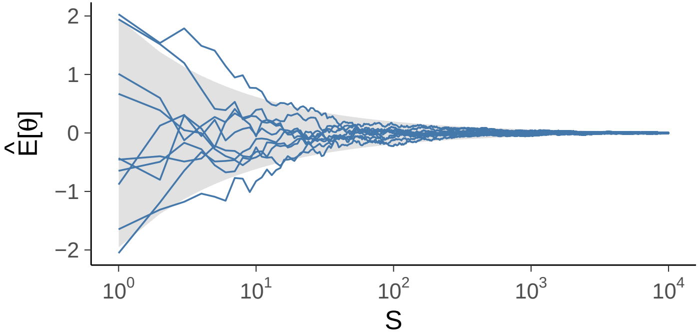
  then the central limit theorem (CLT) states that the distribution of the expectation approaches normal distribution (see BDA3 Ch 4) with variance σ_θ^2/S



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,
  - $\theta^{(s)}$ are independent,
  - $p(\theta)$ has finite variance,

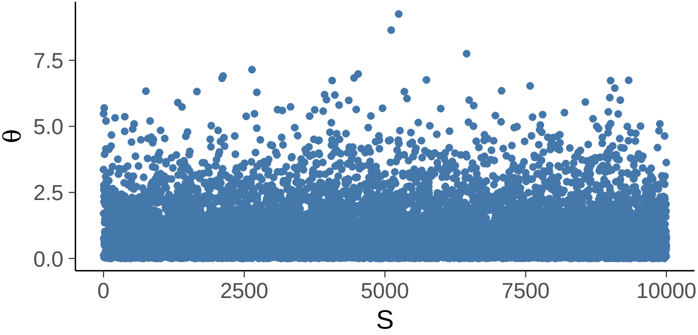
  then the central limit theorem (CLT) states that the distribution of the expectation approaches normal distribution (see BDA3 Ch 4) with variance σ_θ^2/S



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,
  - $\theta^{(s)}$ are independent,
  - $p(\theta)$ has finite variance,

  then the central limit theorem (CLT) states that the distribution of the expectation approaches normal distribution (see BDA3 Ch 4) with variance σ_θ^2/S

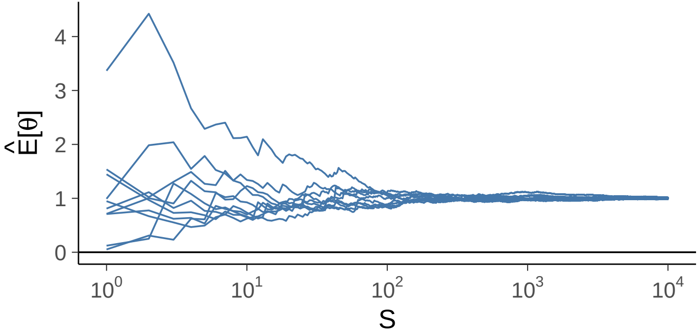


`rexp(n=10000, rate=1)`

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,
  - $\theta^{(s)}$ are independent,
  - $p(\theta)$ has finite variance,

  then the central limit theorem (CLT) states that the distribution of the expectation approaches normal distribution (see BDA3 Ch 4) with variance σ_θ^2/S

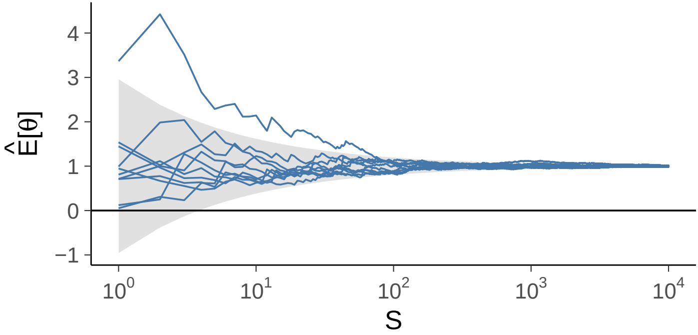



```
cummean(rexp(n=10000, rate=1))
```

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,
  - $\theta^{(s)}$ are independent,
  - $p(\theta)$ has finite variance,

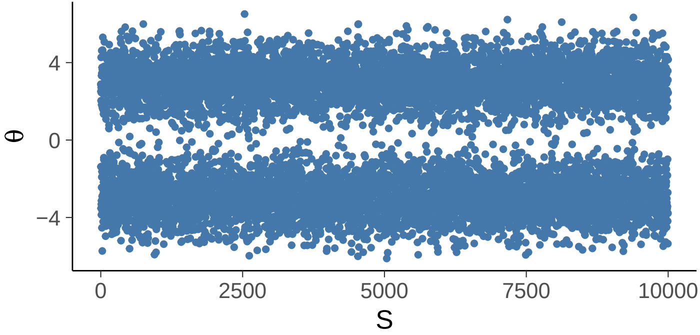
  then the central limit theorem (CLT) states that the distribution of the expectation approaches normal distribution (see BDA3 Ch 4) with variance σ_θ^2/S



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,
  - $\theta^{(s)}$ are independent,
  - $p(\theta)$ has finite variance,

  then the central limit theorem (CLT) states that the distribution of the expectation approaches normal distribution (see BDA3 Ch 4) with variance σ_θ^2/S

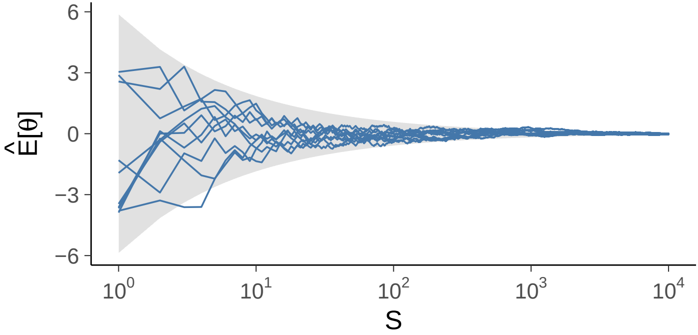



```
rnorm(n=10000, mean=sample(c(-3,3), 10000, replace=TRUE))
```

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,
  - $\theta^{(s)}$ are independent,
  - $p(\theta)$ has finite variance,

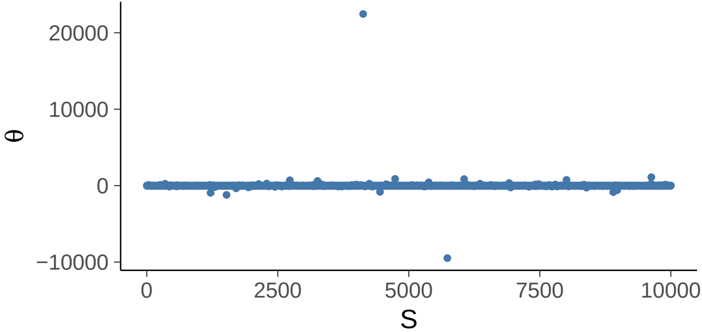
  then the central limit theorem (CLT) states that the distribution of the expectation approaches normal distribution (see BDA3 Ch 4) with variance σ_θ^2/S



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,
  - $\theta^{(s)}$ are independent,
  - $p(\theta)$ has finite variance,

  then the central limit theorem (CLT) states that the distribution of the expectation approaches normal distribution (see BDA3 Ch 4) with variance σ_θ^2/S

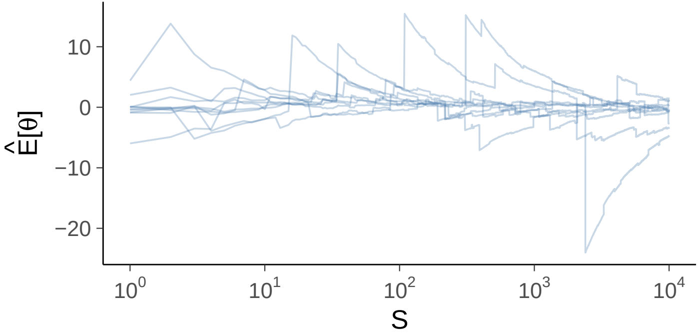



```
rt(n=10000, df=1)
```

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,
  - $\theta^{(s)}$ are independent,
  - $p(\theta)$ has finite variance,

  then the central limit theorem (CLT) states that the distribution of the expectation approaches normal distribution (see BDA3 Ch 4) with variance σ_θ^2/S




```
cummean(rt(n=10000, df=1))
```

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,
  - $\theta^{(s)}$ are independent, and
  - $p(\theta)$ has finite variance,

  then the central limit theorem (CLT) states that the distribution of the expectation approaches normal distribution (see BDA3 Ch 4) with variance σ_θ^2/S
  - this variance is independent on dimensionality of θ

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,
  - $\theta^{(s)}$ are independent, and
  - $p(\theta)$ has finite variance,

  then the central limit theorem (CLT) states that the distribution of the expectation approaches normal distribution (see BDA3 Ch 4) with variance σ_θ^2/S
  - this variance is independent on dimensionality of θ
  - See BDA3 Ch 4 for counter-examples for asymptotic normality

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,
  - $\theta^{(s)}$ are independent, and
  - $p(\theta)$ has finite variance,

  then the central limit theorem (CLT) states that the distribution of the expectation approaches normal distribution (see BDA3 Ch 4) with variance σ_θ^2/S
  - this variance is independent on dimensionality of θ
  - See BDA3 Ch 4 for counter-examples for asymptotic normality
  - $\sigma_\theta/\sqrt{S}$ is called Monte Carlo standard error (MCSE)

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,
  - $\theta^{(s)}$ are independent, and
  - $p(\theta)$ has finite variance,

  then the central limit theorem (CLT) states that the distribution of the expectation approaches normal distribution (see BDA3 Ch 4) with variance σ_θ^2/S
  - this variance is independent on dimensionality of θ
  - See BDA3 Ch 4 for counter-examples for asymptotic normality
  - $\sigma_\theta/\sqrt{S}$ is called Monte Carlo standard error (MCSE)
  - In practice, σ_θ will be estimated by $\sqrt{1/(S-1)\sum_{s=1}^{S}(\theta^{(s)} - \mathrm{E}(\theta))^2}$

Central limit theorem

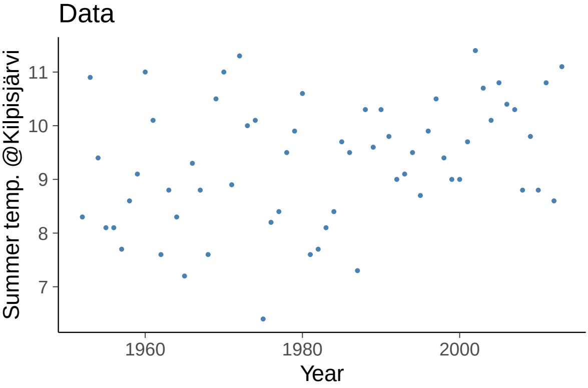
- Valid also when $p(\theta)$ discrete
  - the distribution of mean is discrete, but the comparison to continuous normal is done using cumalative distribution functions

Central limit theorem

- Valid also when $p(\theta)$ discrete
  - the distribution of mean is discrete, but the comparison to continuous normal is done using cumalative distribution functions
- 3Blue1Brown YouTube videos with nice visualisations
  - CLT with discrete distributions: *But what is the Central Limit Theorem?* https://www.youtube.com/watch?v=zeJD6dqJ5lo
  - CLT with continuous distributions: *Convolutions | Why X+Y in probability is a beautiful mess* https://www.youtube.com/watch?v=IaSGqQa5O-M

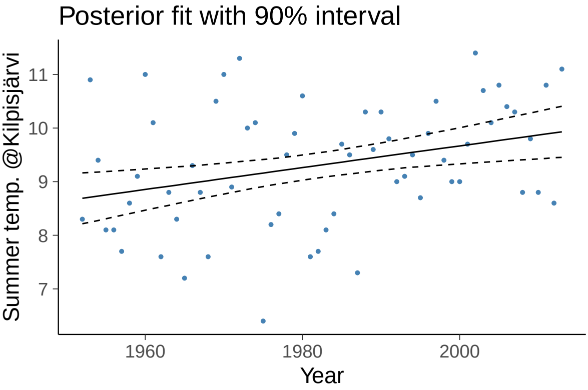
Example: Kilpisjärvi summer temperature

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

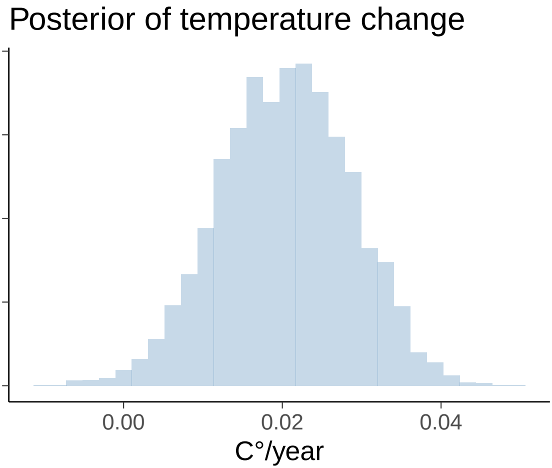


Example: Kilpisjärvi summer temperature

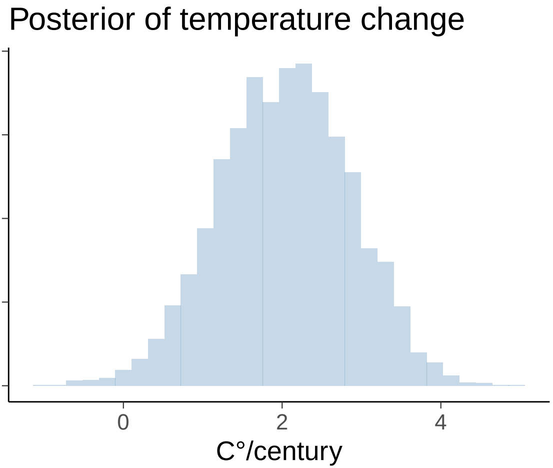
Average temperature in June, July, and August at Kilpisjärvi, Finland in 1952–2013



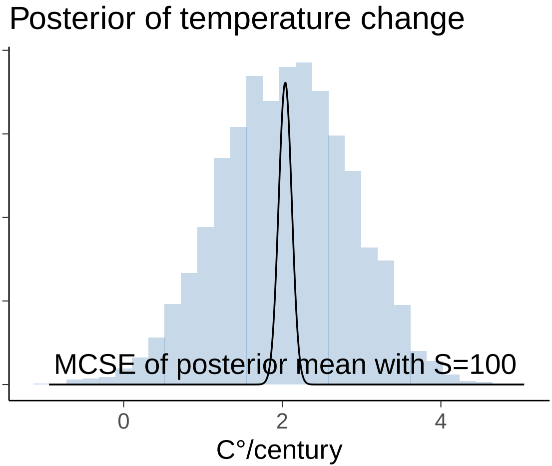
Example: Kilpisjärvi summer temperature



Example: Kilpisjärvi summer temperature



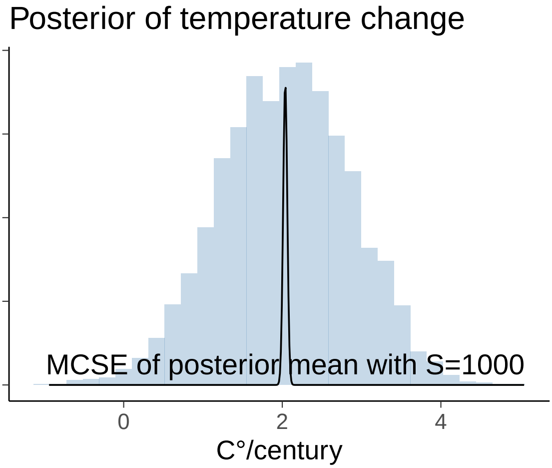
Example: Kilpisjärvi summer temperature



$\sigma_\theta \approx 0.83$, MCSE $= \sigma_\theta/\sqrt{S} \approx 0.083$,

in repeated sampling we may expect mean estimate to vary within (1.8, 2.1) (90% interval)

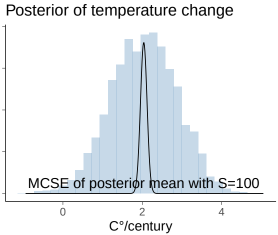
Example: Kilpisjärvi summer temperature

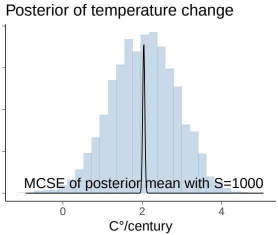


$\sigma_\theta \approx 0.83$, MCSE ≈ 0.026,

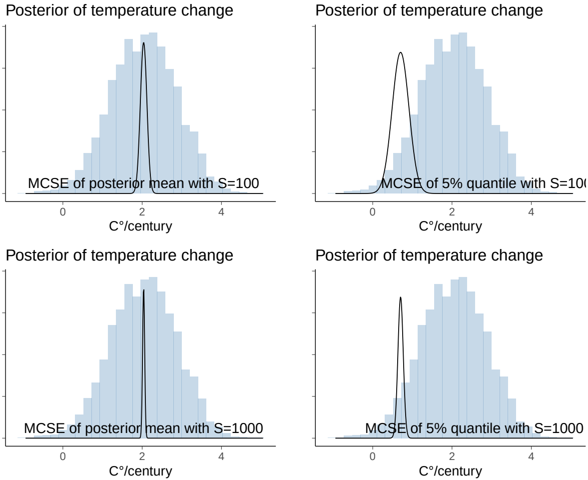
in repeated sampling we may expect mean estimate to vary within (1.9, 2.0) (90% interval)

Example: Kilpisjärvi summer temperature

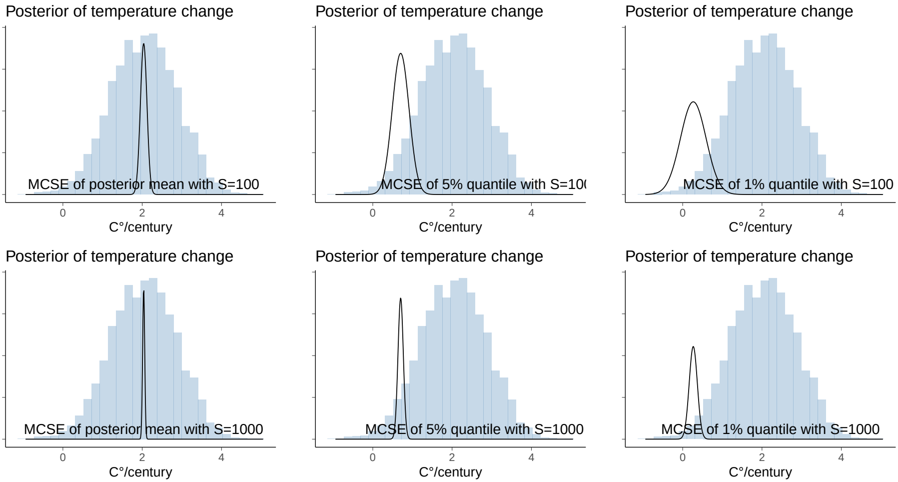




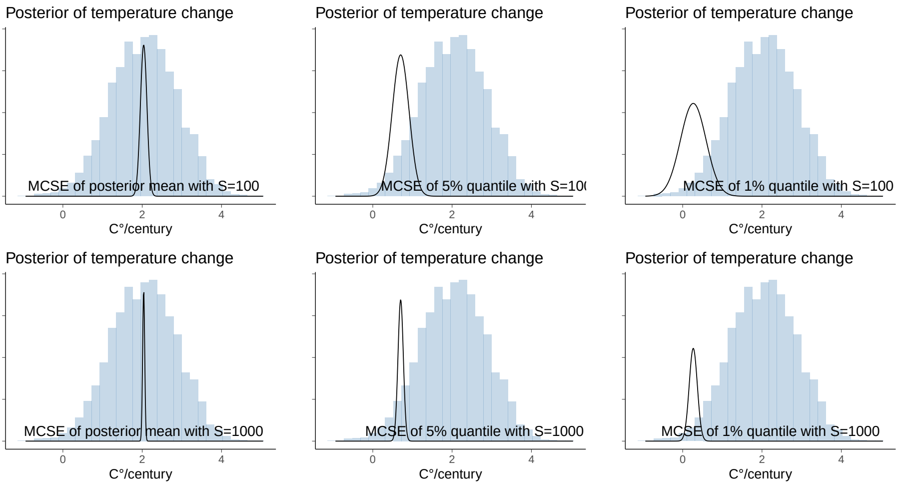
Example: Kilpisjärvi summer temperature



Example: Kilpisjärvi summer temperature



Example: Kilpisjärvi summer temperature



Tail quantiles are more difficult to estimate

See Vehtari, Gelman, Simpson, Carpenter, & Bürkner (2021) for quantile MCSE computation.

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 $p(\theta \in A)$
    - use beta CDF, or normal approximation
    - $\rightarrow$ $\mathrm{var}(I(\cdot)) = p(1-p)S$ (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 $p(\theta \in A)$
    - use beta CDF, or normal approximation

    $\rightarrow$ $\mathrm{var}(I(\cdot)) = p(1-p)S$ (Appendix A, p. 579)

    $\rightarrow$ standard deviation of p is $\sqrt{p(1-p)/S}$
  - if $S = 100$ and we observe $\frac{1}{S} \sum_l I(\theta^{(s)} \in A) = 0.05$,
    then $\sqrt{p(1-p)/S} \approx 0.02$
    i.e. accuracy is about 4% units
    or from quantiles of beta distribution the range is $(0.02, 0.1)$

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 $p(\theta \in A)$
    - use beta CDF, or normal approximation

    $\rightarrow$ $\mathrm{var}(I(\cdot)) = p(1-p)S$ (Appendix A, p. 579)

    $\rightarrow$ standard deviation of p is $\sqrt{p(1-p)/S}$
  - if $S = 100$ and we observe $\frac{1}{S}\sum_l I(\theta^{(s)} \in A) = 0.05$,
    then $\sqrt{p(1-p)/S} \approx 0.02$
    i.e. accuracy is about 4% units
    or from quantiles of beta distribution the range is $(0.02, 0.1)$
  - $S = 2000$ draws needed for 1% unit accuracy

How many simulation draws are needed?

- Posterior probability

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

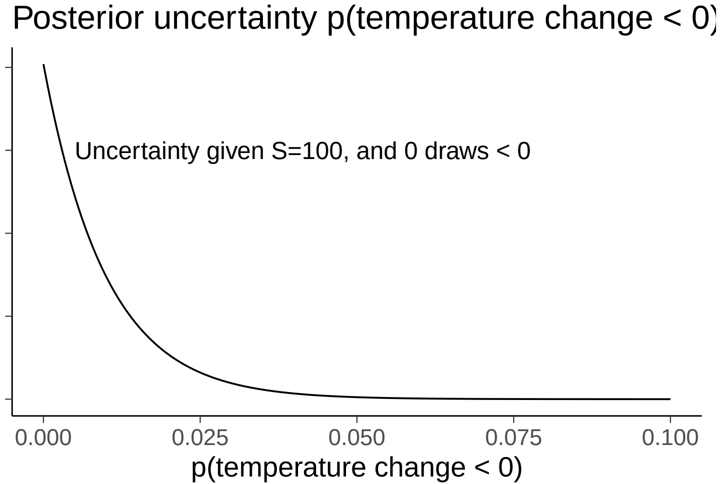
where $I(\theta^{(s)} \in A) = 1$ if $\theta^{(s)} \in A$

  - $I(\cdot)$ is binomially distributed as $p(\theta \in A)$
    - use beta CDF, or normal approximation

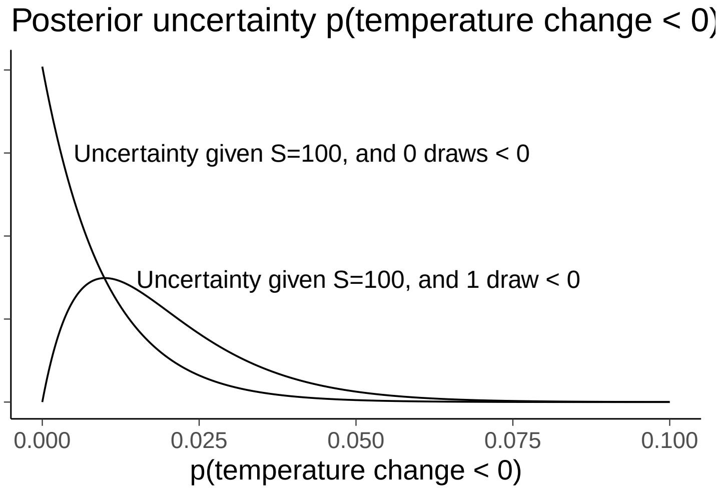
    $\rightarrow$ $\mathrm{var}(I(\cdot)) = p(1-p)S$ (Appendix A, p. 579)

    $\rightarrow$ standard deviation of p is $\sqrt{p(1-p)/S}$
  - if $S = 100$ and we observe $\frac{1}{S}\sum_l I(\theta^{(s)} \in A) = 0.05$,
    then $\sqrt{p(1-p)/S} \approx 0.02$
    i.e. accuracy is about 4% units
    or from quantiles of beta distribution the range is $(0.02, 0.1)$
  - $S = 2000$ draws needed for 1% unit accuracy
- To estimate small probabilities, a large number of draws is needed
  - to be able to estimate small p, need to get draws with $\theta^{(l)} \in A$, which in expectation requires $S \gg 1/p$

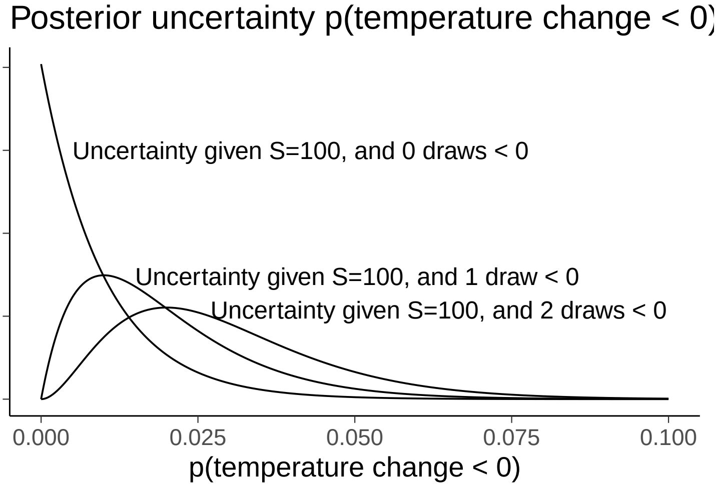
Example: Kilpisjärvi summer temperature



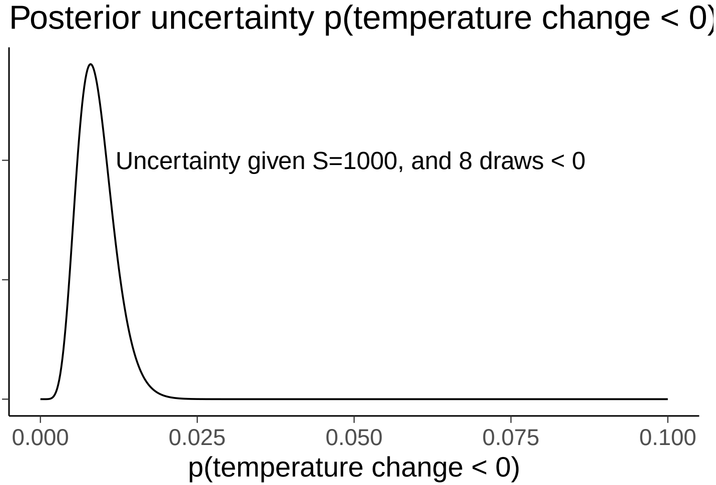
Example: Kilpisjärvi summer temperature



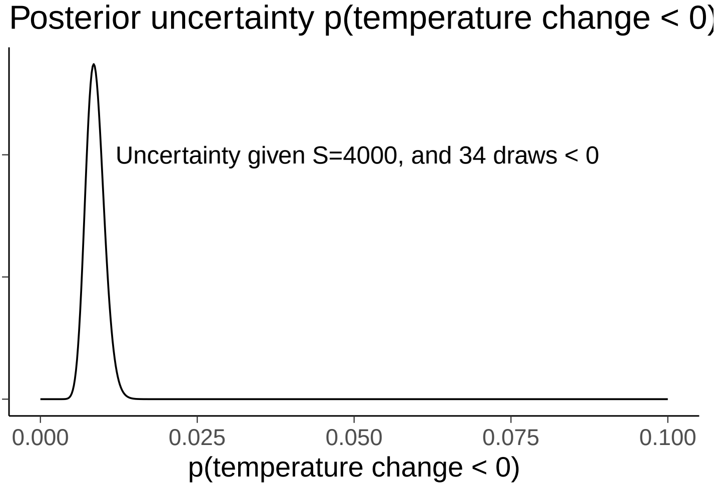
Example: Kilpisjärvi summer temperature



Example: Kilpisjärvi summer temperature



Example: Kilpisjärvi summer temperature



From probabilities to quantiles

- Probability: $p(\theta < A) \approx \frac{1}{S} \sum_l I(\theta^{(s)} < A)$
- 5%-quantile: Find A so that $p(\theta < A) = 0.05$

From probabilities to quantiles

- Probability: $p(\theta < A) \approx \frac{1}{S} \sum_l I(\theta^{(s)} < A)$
- 5%-quantile: Find A so that $p(\theta < A) = 0.05$
- If $S = 1000$ and uncertainty interval for 5% probability is $(0.04, 0.06)$ (see earlier slide), we can find uncertainty interval (A^-, A^+), so that $p(\theta < A^-) = 0.04$, and $p(\theta < A^+) = 0.06$

From probabilities to quantiles

- Probability: $p(\theta < A) \approx \frac{1}{S} \sum_l I(\theta^{(s)} < A)$
- 5%-quantile: Find A so that $p(\theta < A) = 0.05$
- If $S = 1000$ and uncertainty interval for 5% probability is $(0.04, 0.06)$ (see earlier slide), we can find uncertainty interval (A^-, A^+), so that $p(\theta < A^-) = 0.04$, and $p(\theta < A^+) = 0.06$
  - we can summarise this interval by transforming it to MCSE
  - see examples in https://avehtari.github.io/casestudies/Digits/digits.html
  - if interested, see algorithm details in Vehtari, Gelman, Simpson, Carpenter, & Bürkner (2021), doi.org/10.1214/20-BA1221.

posterior package

Posterior mean and 5% and 95% quantiles:

```
draws |>
  subset_draws("beta100") |>
  summarize_draws(mean,
                  ~quantile(.x, probs = c(0.05, 0.95)))
```

posterior package

Posterior mean and 5% and 95% quantiles:

```
draws |>
  subset_draws("beta100") |>
  summarize_draws(mean,
                  ~quantile(.x, probs = c(0.05, 0.95)))
```

The corresponding MCSE estimates:

```
draws |>
  subset_draws("beta100") |>
  summarize_draws(mcse_mean,
                  ~mcse_quantile(.x, probs = c(0.05, 0.95)))
```

posterior package

Posterior mean and 5% and 95% quantiles:

```
draws |>
  subset_draws("beta100") |>
  summarize_draws(mean,
                  ~quantile(.x, probs = c(0.05, 0.95)))
```

The corresponding MCSE estimates:

```
draws |>
  subset_draws("beta100") |>
  summarize_draws(mcse_mean,
                  ~mcse_quantile(.x, probs = c(0.05, 0.95)))
```

These `_mcse` functions are for MCMC draws, but if the number of draws is big (≥ 1000), then these are accurate enough for independent MC draws, too

posterior package

Posterior probability and the corresponding MCSE estimate:

```
draws |>
  mutate_variables(beta0p = beta100>0) |>
  subset_draws("beta0p") |>
  summarize_draws(mean,
                  mcse = mcse_mean)
```

How many digits to report?

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

How many digits to report?

- 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 report?

- 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 report?

- 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

How many digits to report?

- 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!)

How many digits to report?

- 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 report?

- 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)

How many digits to report?

- 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

How many digits to report?

- 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 report?

- 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 report?

- 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 report?

- 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 report?

- 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://users.aalto.fi/~ave/casestudies/Digits/digits.html

More data

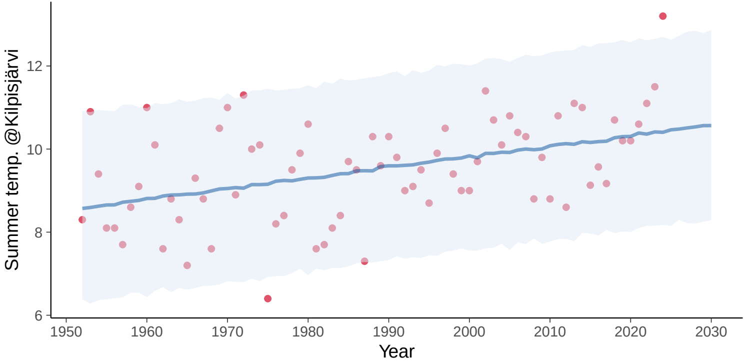
- The analysis I just showed used data from 1952–2013

More data

- The analysis I just showed used data from 1952–2013
- With data data from 1952–2024
  - The probability that temp increase is positive: 0.99975 ± 0.00025 (90% interval),
    which can be reported as more than 99.95% probability
  - With data from other locations we would be even more certain

More data

- The analysis I just showed used data from 1952–2013
- With data data from 1952–2024
  - The probability that temp increase is positive: 0.99975 ± 0.00025 (90% interval),
    which can be reported as more than 99.95% probability
  - With data from other locations we would be even more certain
- Summer 2023 was the second hottest in the recorded history
- Summer 2024 was the hottest in the recorded history



How many simulation draws are needed?

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

How many simulation draws are needed?

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

How many simulation draws are needed?

- Fewer draws needed with
  - deterministic methods
  - marginalization (Rao-Blackwellization)
  - variance reduction methods, such, control variates
- 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
- Some algorithms are less efficient
  - Compute MCSE using *effective sample size (ESS)* instead of the number of draws S
  - Usually ESS$< S$

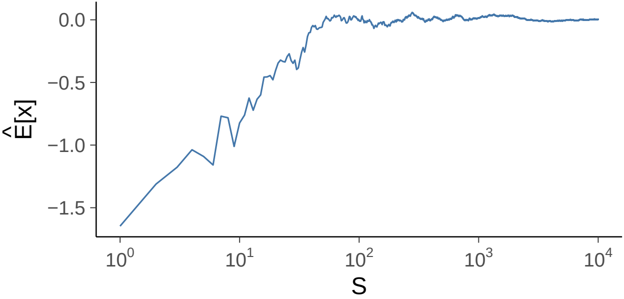
How many simulation draws are needed?

- Fewer draws needed with
  - deterministic methods
  - marginalization (Rao-Blackwellization)
  - variance reduction methods, such, control variates
- 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
- Some algorithms are less efficient
  - Compute MCSE using *effective sample size (ESS)* instead of the number of draws S
  - Usually ESS$< S$
- How to check if a distribution has finite mean and variance?
  - Pareto-$\hat{k}$ diagnostic

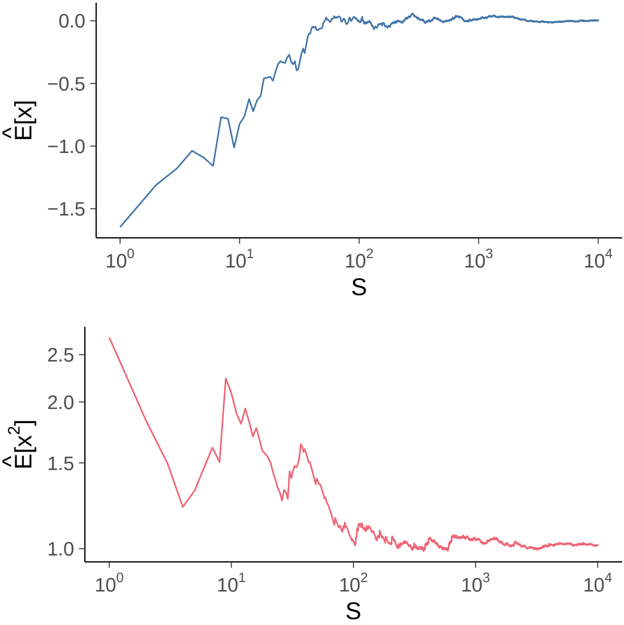
Simple example: $x \sim N, \quad t_4, \quad t_2, \quad t_1, \quad t_{1/2}$

- N has all moments finite
- t_ν has less than ν fractional moments

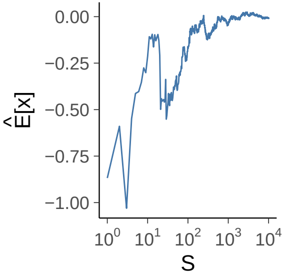
Simple example: $x \sim N$



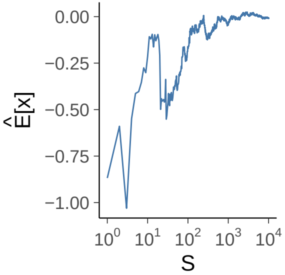
Simple example: $x \sim N$

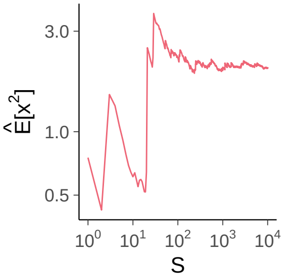


Simple example: $x \sim t_4, \quad t_2, \quad t_1, \quad t_{1/2}$

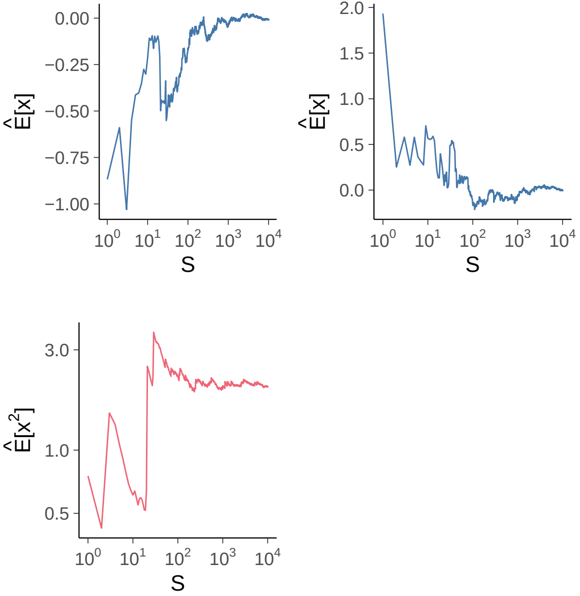


Simple example: $x \sim t_4, \quad t_2, \quad t_1, \quad t_{1/2}$

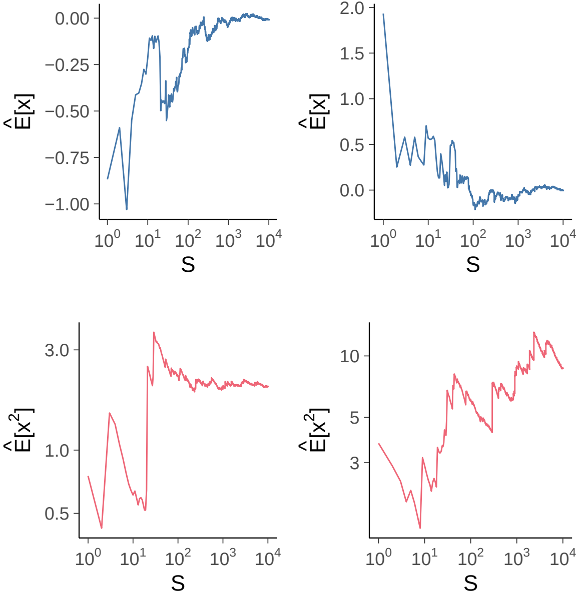




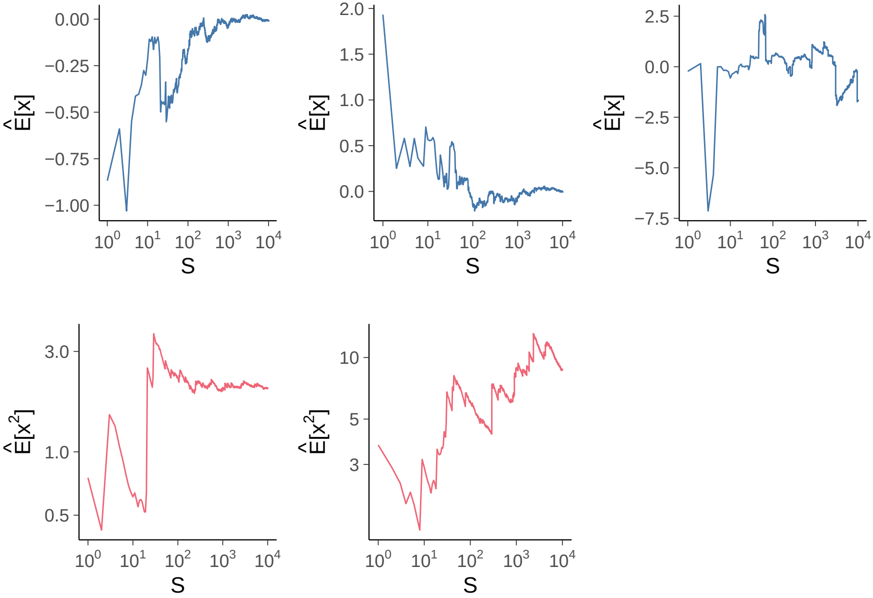
Simple example: $x \sim t_4, \quad t_2, \quad t_1, \quad t_{1/2}$



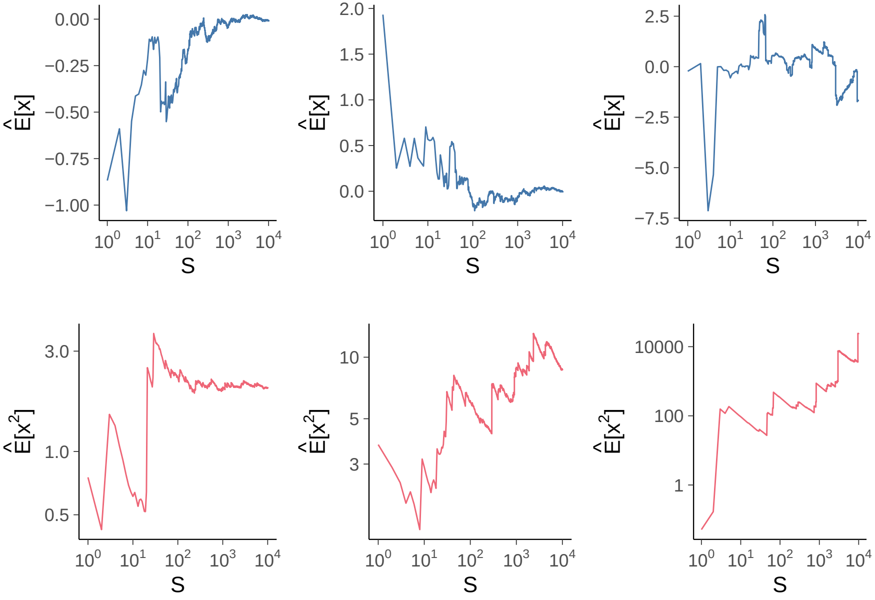
Simple example: $x \sim t_4, \quad t_2, \quad t_1, \quad t_{1/2}$



Simple example: $x \sim t_4, \quad t_2, \quad t_1, \quad t_{1/2}$

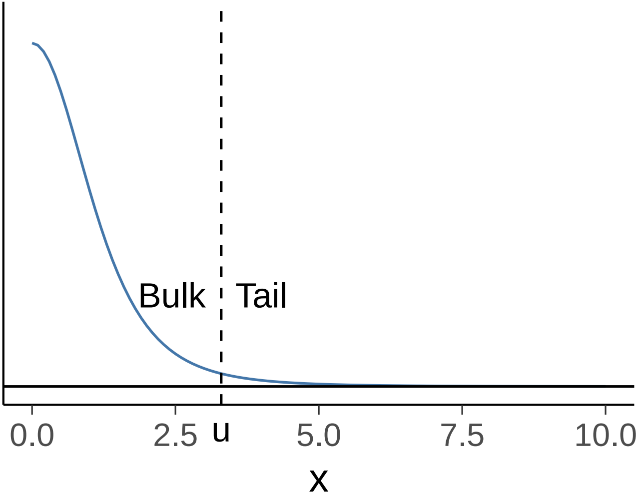


Simple example: $x \sim t_4, \quad t_2, \quad t_1, \quad t_{1/2}$



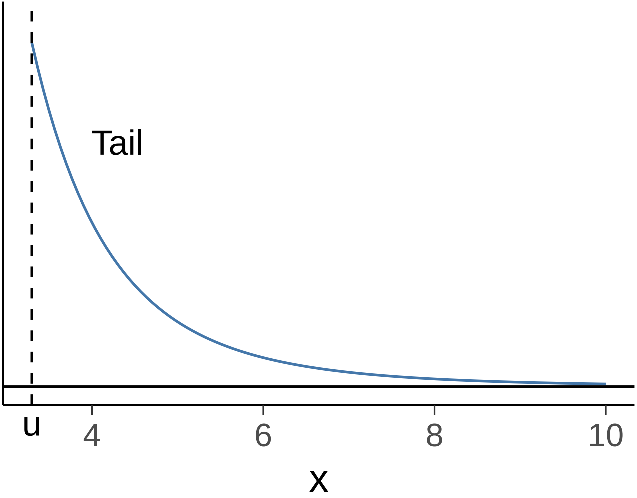
Pareto-$\hat{k}$ diagnostic

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



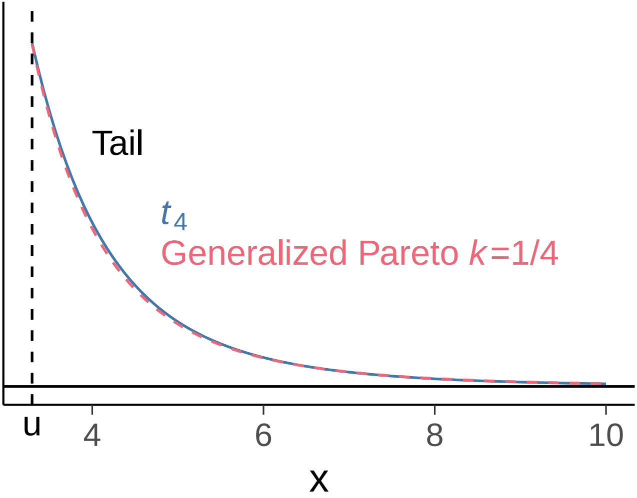
Pareto-$\hat{k}$ diagnostic

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



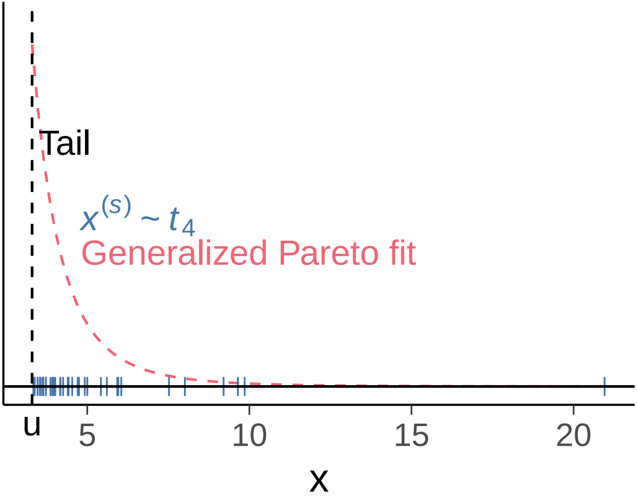
Pareto-$\hat{k}$ diagnostic

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



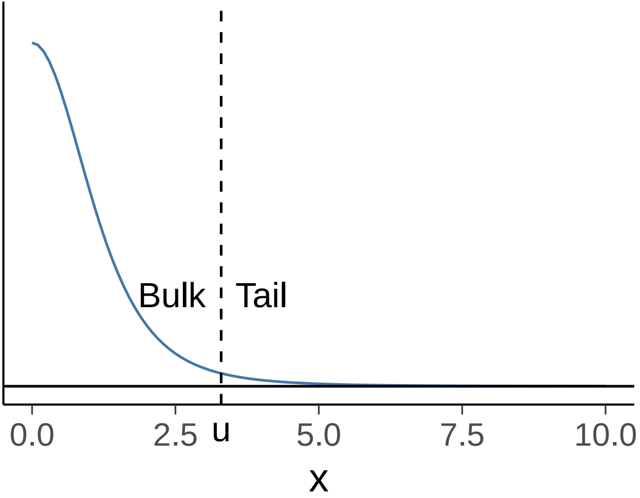
Pareto-$\hat{k}$ diagnostic

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

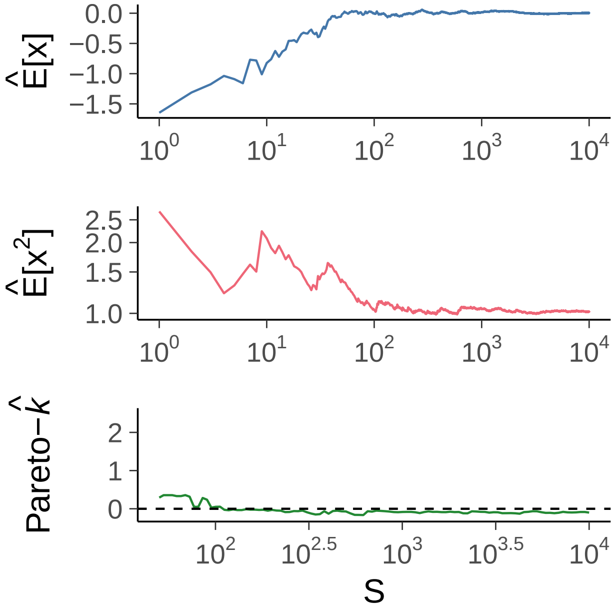


Pareto-$\hat{k}$ diagnostic

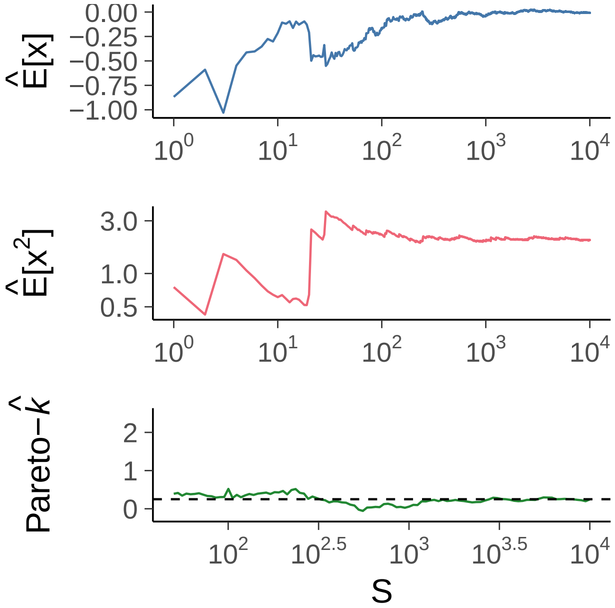
GPD has a shape parameter k,
and $1/k$ finite fractional moments



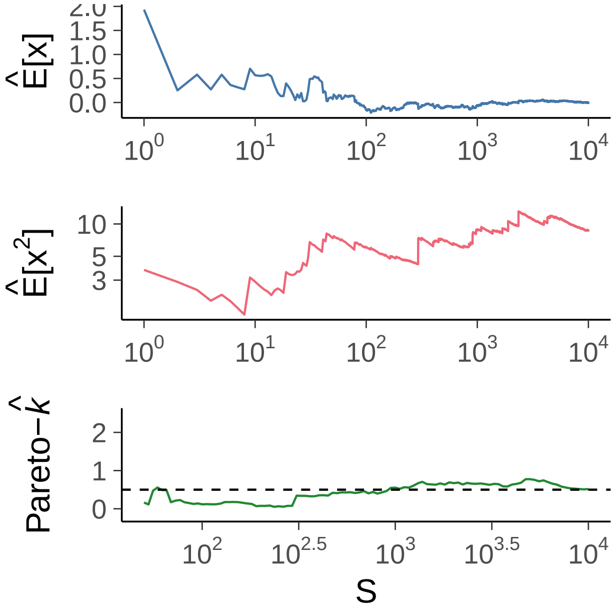
Pareto-$\hat{k}$ diagnostic: $x \sim N$



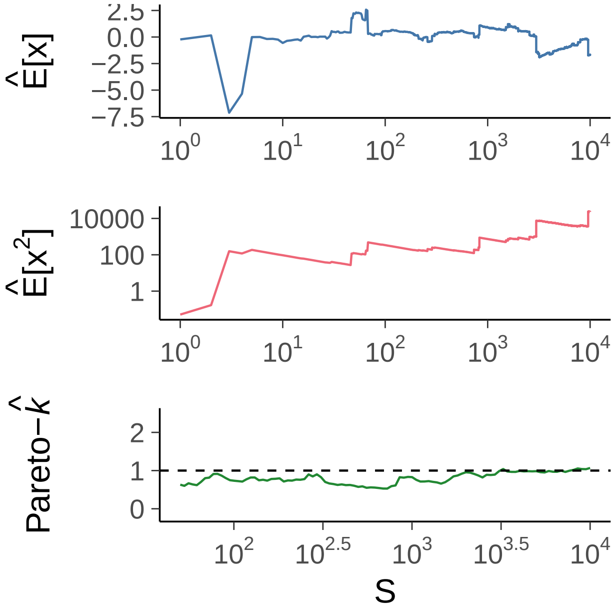
Pareto-$\hat{k}$ diagnostic: $x \sim t_4$



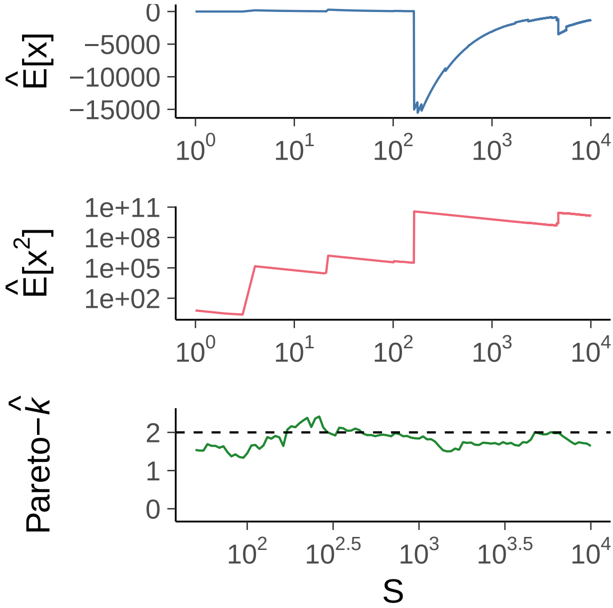
Pareto-$\hat{k}$ diagnostic: $x \sim t_2$



Pareto-$\hat{k}$ diagnostic: $x \sim t_1$



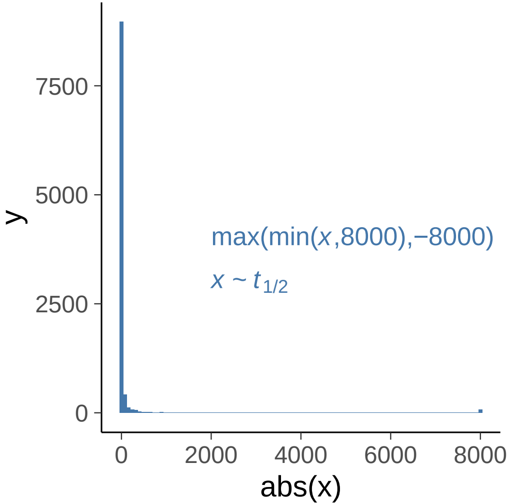
Pareto-$\hat{k}$ diagnostic: $x \sim t_{1/2}$



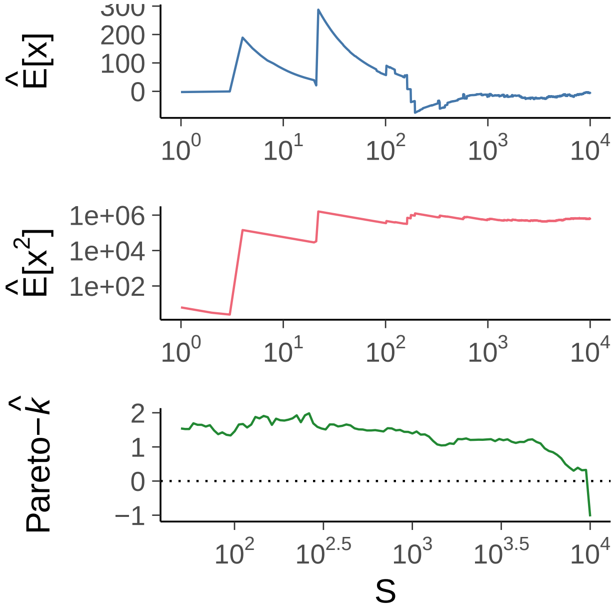
Pareto-$\hat{k}$ diagnostic is pre-asymptotic diagnostic

Thick tailed but truncated distribution

We can make estimates only based on what we have observed.



Pareto-$\hat{k}$ diagnostic: thick-tailed bounded distribution



Thick-tailed bounded distributions in practice

- Thick-tailed distributions are common in importance sampling and variational divergence estimation

Pareto-$\hat{k}$ in `posterior` package

```
> drt |> summarise_draws(mean, sd, mcse_mean)
```

```
variable    mean      sd mcse_mean
xn         0.007    0.99      0.01
xt3        0.004    1.66      0.02
xt2_5      0.002    2.01      0.02
xt2       -0.008    3.00      0.03
xt1_5     -0.067    8.14      0.08
xt1       -1.57   122.        1.21
```

Pareto-$\hat{k}$ in posterior package

```
> drt |> summarise_draws(mean, sd, mcse_mean, pareto_khat)

variable    mean      sd mcse_mean pareto_khat
xn         0.007    0.99      0.01       -0.02
xt3        0.004    1.66      0.02        0.36
xt2_5      0.002    2.01      0.02        0.43
xt2       -0.008    3.00      0.03        0.53
xt1_5     -0.067    8.14      0.08        0.72
xt1       -1.57   122.        1.21        1.08
```

How to use Pareto-$\hat{k}$ diagnostic

- To check posterior of any quantity of interest
  - if high $\hat{k}$, maybe use some other summary than mean, e.g., quantiles

See more in Vehtari, Simpson, Gelman, Yao, and Gabry (2024). Pareto smoothed importance sampling. *JMLR*, 25(72):1-58.

How to use Pareto-$\hat{k}$ diagnostic

- To check posterior of any quantity of interest
  - if high $\hat{k}$, maybe use some other summary than mean, e.g., quantiles
- Especially useful inside algorithms that rely on expectations
  - other summaries can't be used
  - automated diagnostic as in PSIS-LOO (Lecture 9) and `priorsense` (Lecture ?)

See more in Vehtari, Simpson, Gelman, Yao, and Gabry (2024). Pareto smoothed importance sampling. *JMLR*, 25(72):1-58.

How to use Pareto-$\hat{k}$ diagnostic

- To check posterior of any quantity of interest
  - if high $\hat{k}$, maybe use some other summary than mean, e.g., quantiles
- Especially useful inside algorithms that rely on expectations
  - other summaries can't be used
  - automated diagnostic as in PSIS-LOO (Lecture 9) and `priorsense` (Lecture ?)
- $\hat{k}$ estimate has it's own variation given finite sample size
  - e.g. if close to 0.5 more draws help to improve to decide whether $k < 0.5$

See more in Vehtari, Simpson, Gelman, Yao, and Gabry (2024). Pareto smoothed importance sampling. *JMLR*, 25(72):1-58.

How to use Pareto-$\hat{k}$ diagnostic

- To check posterior of any quantity of interest
  - if high $\hat{k}$, maybe use some other summary than mean, e.g., quantiles
- Especially useful inside algorithms that rely on expectations
  - other summaries can't be used
  - automated diagnostic as in PSIS-LOO (Lecture 9) and `priorsense` (Lecture ?)
- $\hat{k}$ estimate has it's own variation given finite sample size
  - e.g. if close to 0.5 more draws help to improve to decide whether $k < 0.5$
- Pareto-smoothing improves the mean estimate
  - reliable mean and MCSE estimates when Pareto-$k < 0.7$
  - required minimum sample size and convergence rate estimates for different values of k
  - more on lecture 9

See more in Vehtari, Simpson, Gelman, Yao, and Gabry (2024). Pareto smoothed importance sampling. *JMLR*, 25(72):1-58.

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

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

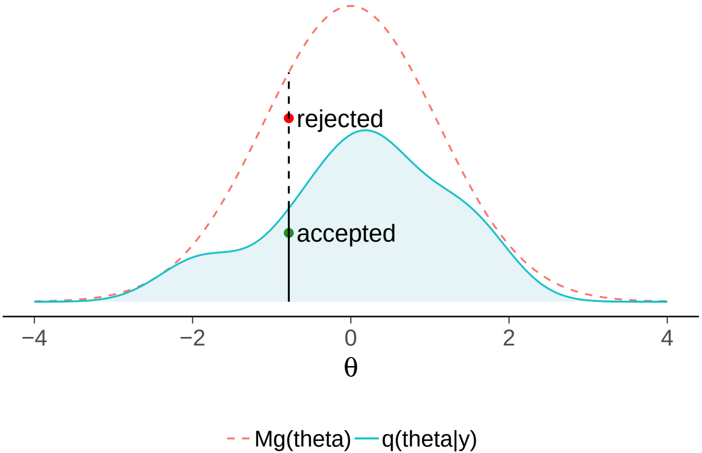
  then X_1 and X_2 are independent draws from the distribution $\mathrm{N}(0, 1)$
  - not the fastest method due to trigonometric computations
  - for normal distribution more than ten different methods
  - e.g. R uses inverse-CDF

Indirect sampling

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

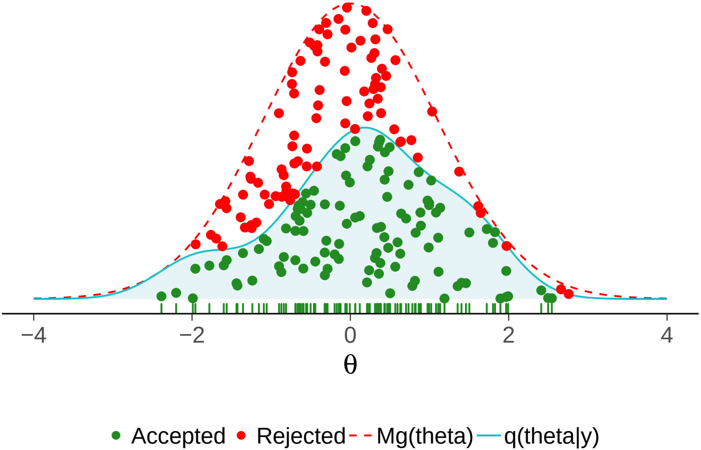
Rejection sampling

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



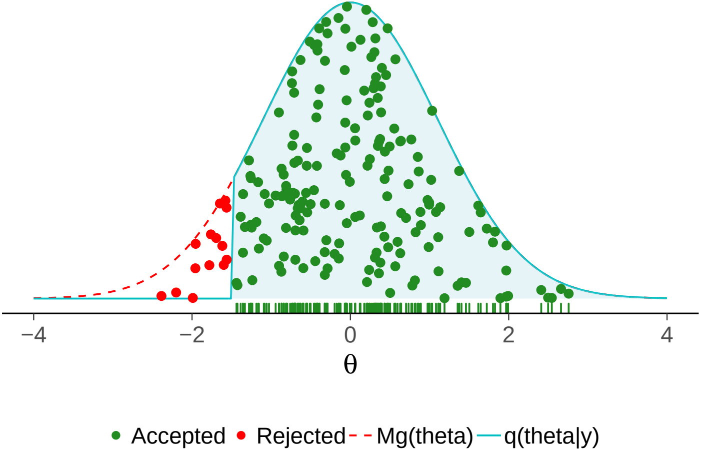
Rejection sampling

- 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

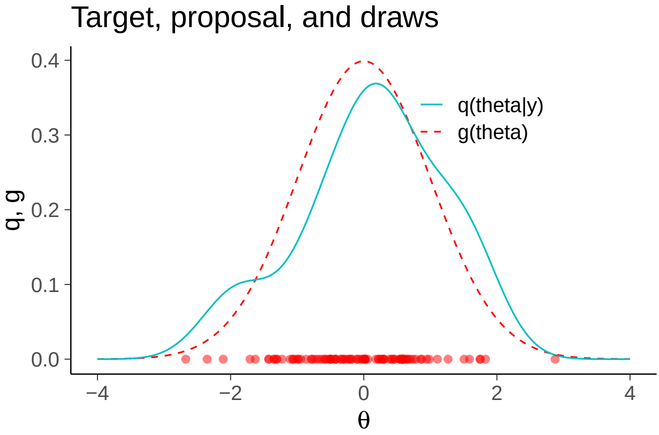


Rejection sampling

- The effective sample size (ESS) is the number of accepted draws
  - 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

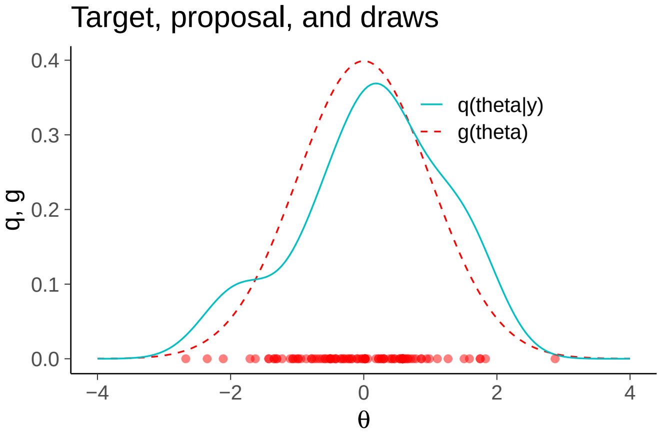
Importance sampling

- Proposal does not need to have a higher value everywhere



Importance sampling

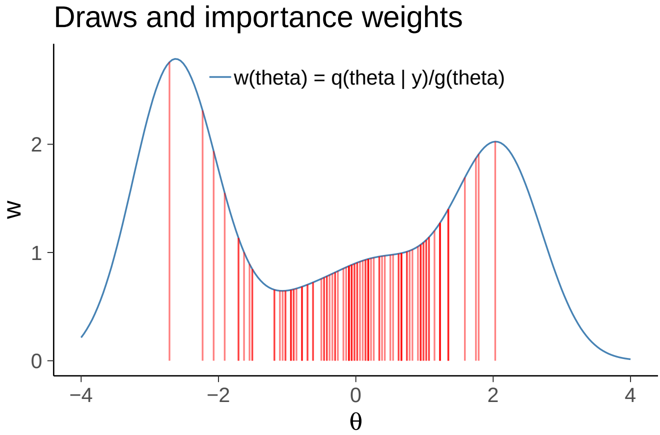
- Proposal does not need to have a higher value everywhere



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

Importance sampling

- Proposal does not need to have a higher value everywhere



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

Some uses of importance sampling

In general selection of good proposal gets more difficult when the number of dimensions increase, but there are many special use case which scale well (e.g. I've used IS up to 10k dimensions)

Some uses of importance sampling

In general selection of good proposal gets more difficult when the number of dimensions increase, but there are many special use case which scale well (e.g. I've used IS up to 10k dimensions)

- Fast leave-one-out cross-validation (`loo`)
- Fast bootstrapping
- Fast prior and likelihood sensitivity analysis (`priorsense`)
- Conformal Bayesian computation
- Particle filtering
- Improving distributional approximations (e.g Laplace, Pathfinder, 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
- Pre-asymptotic and asymptotic behavior can be really different!

Importance re-sampling

- Using the weighted draws is good

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

Importance re-sampling

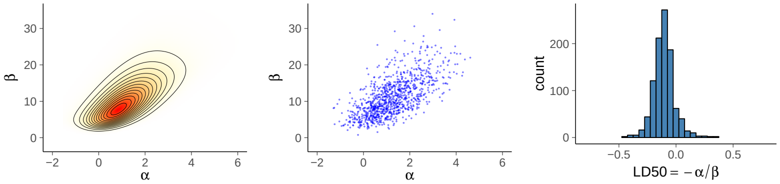
- Using the weighted draws is good

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

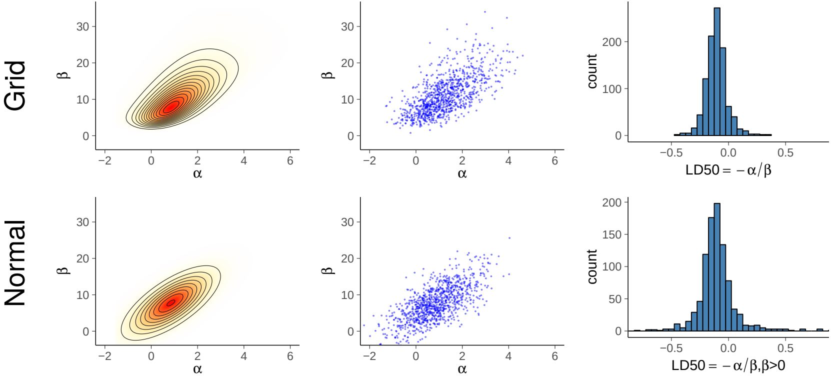
- But it can be convenient to obtain draws with equal weights
  - resample the draws according to the weights
  - some original draws may be included more than once
  - loses some information, but now the weights are equal

Example: Importance sampling in Bioassay

Grid

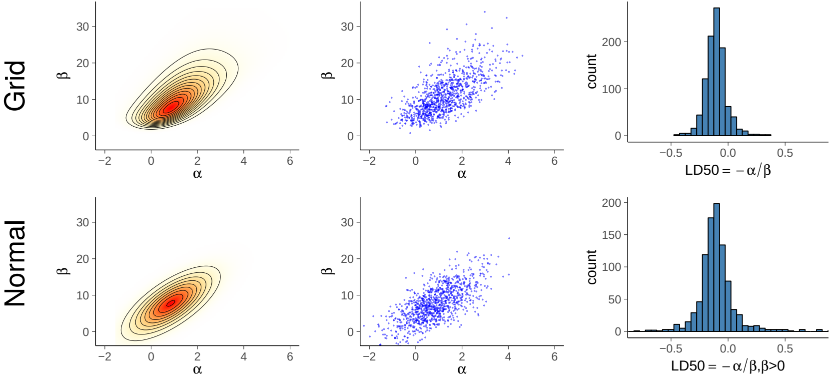


Example: Importance sampling in Bioassay



Normal approximation is discussed more in BDA3 Ch 4

Example: Importance sampling in Bioassay

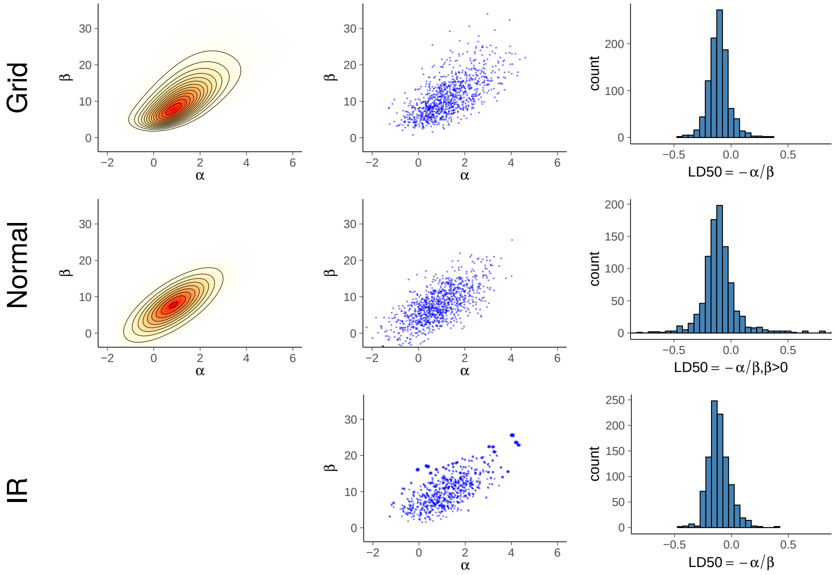


Normal approximation is discussed more in BDA3 Ch 4

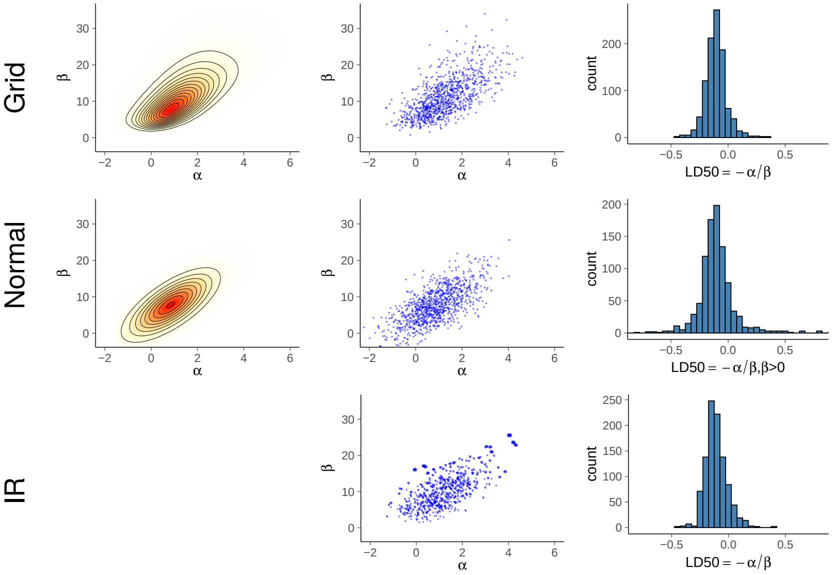
But the normal approximation is not that good here:

Grid sd(LD50) $\approx$ 0.1, Normal sd(LD50) $\approx$ .75!

Example: Importance sampling in Bioassay

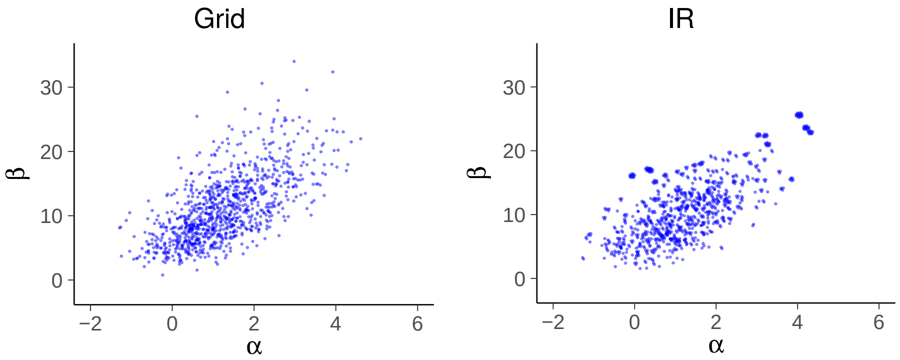


Example: Importance sampling in Bioassay

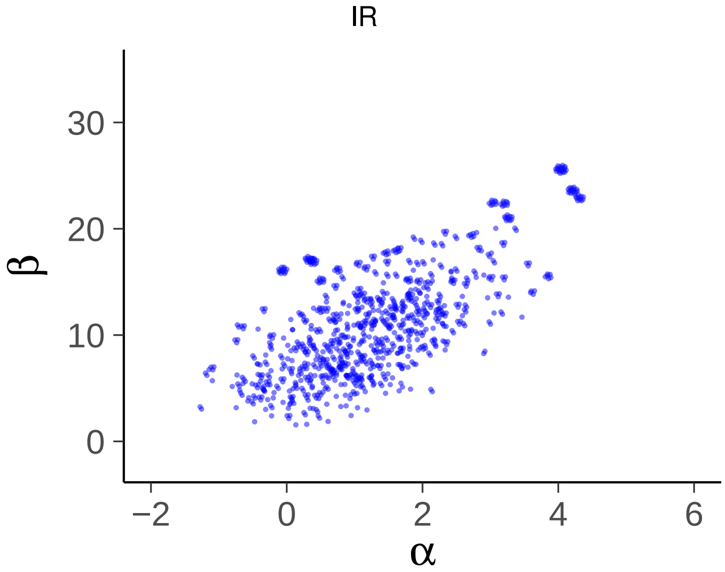


Grid sd(LD50) $\approx$ 0.1, IR sd(LD50) $\approx$ 0.1

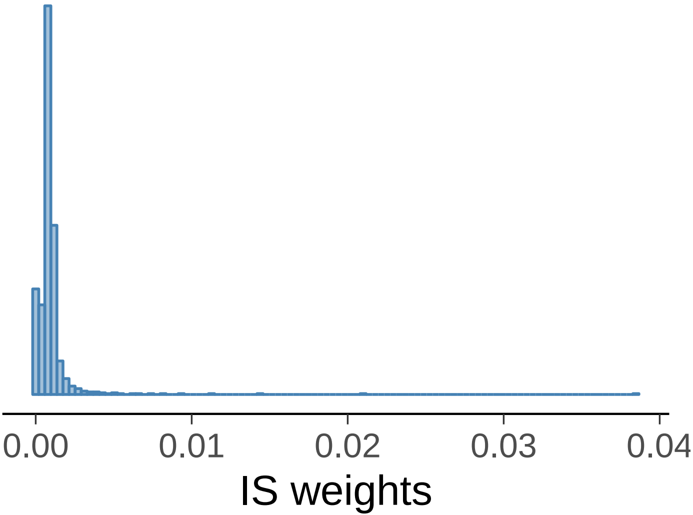
Example: Importance sampling in Bioassay



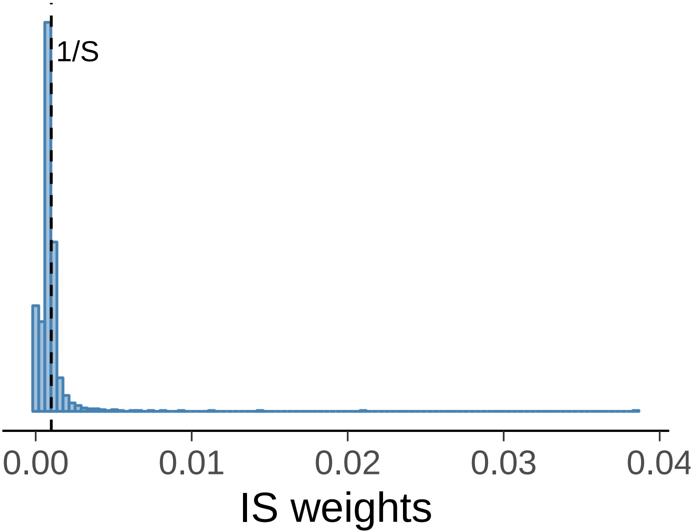
Example: Importance sampling in Bioassay



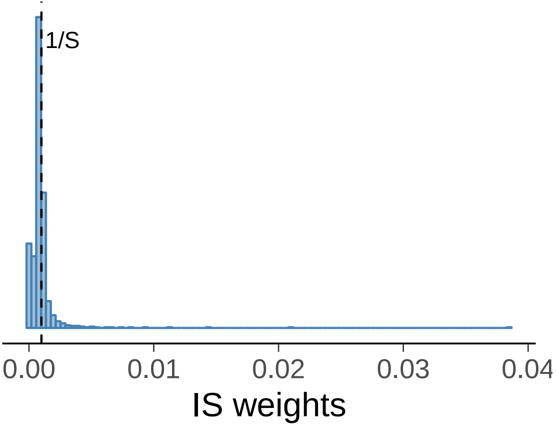
Example: Importance sampling in Bioassay



Example: Importance sampling in Bioassay

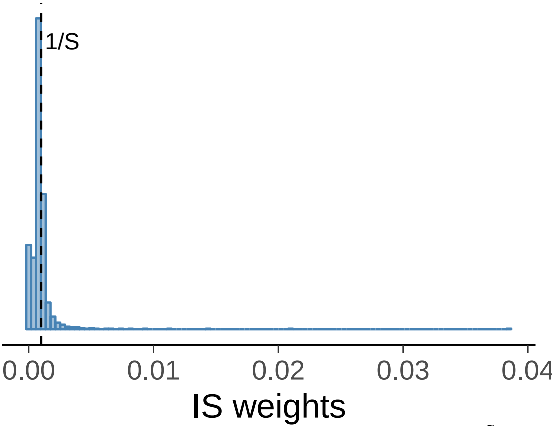


Example: Importance sampling in Bioassay



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

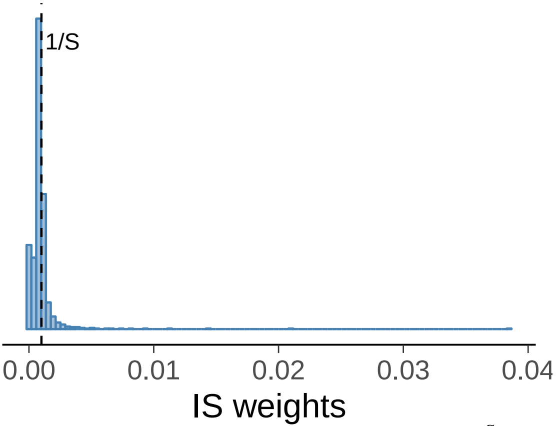
Example: Importance sampling in Bioassay



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

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

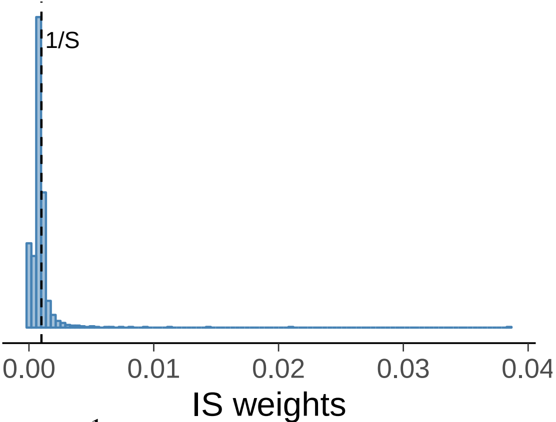
Example: Importance sampling in Bioassay



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

$$\text{ESS} \approx 396, \quad (\text{ESS} < S = 1000)$$

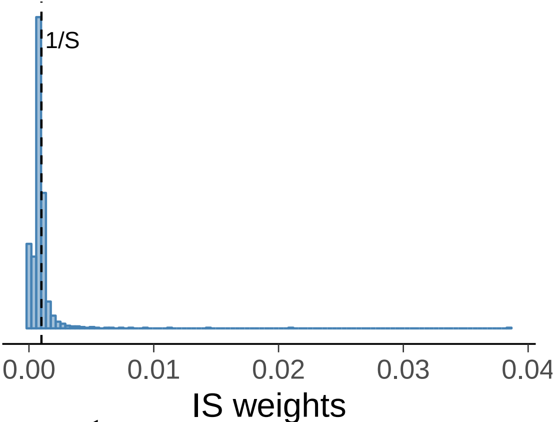
Example: Importance sampling in Bioassay



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

ESS ≈ 396

Example: Importance sampling in Bioassay

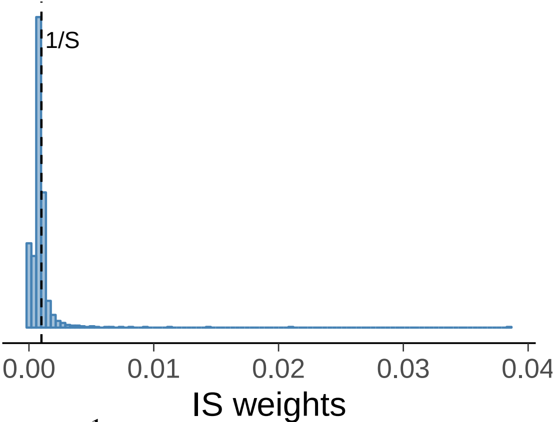


$$\text{ESS} = \frac{1}{\sum_{s=1}^{S}(\tilde{w}(\theta^s))^2},$$ is based on variance of $\tilde{w}(\theta^s)$

$$\text{ESS} \approx 396$$

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

Example: Importance sampling in Bioassay



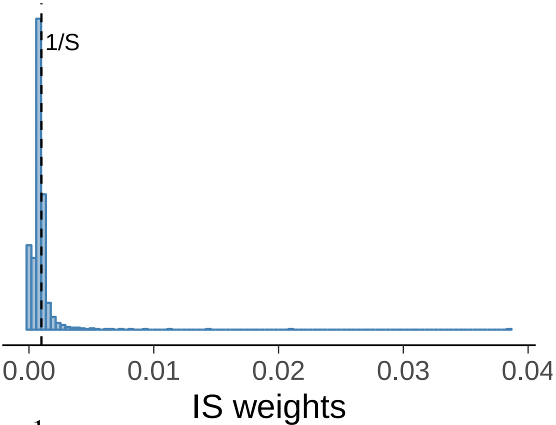
$$\text{ESS} = \frac{1}{\sum_{s=1}^{S}(\tilde{w}(\theta^s))^2}, \quad \text{is based on variance of } \tilde{w}(\theta^s)$$

$$\text{ESS} \approx 396$$

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

If one $\tilde{w}(\theta^s) = 1$, and others 0, then ESS $= 1/1 = 1$

Example: Importance sampling in Bioassay

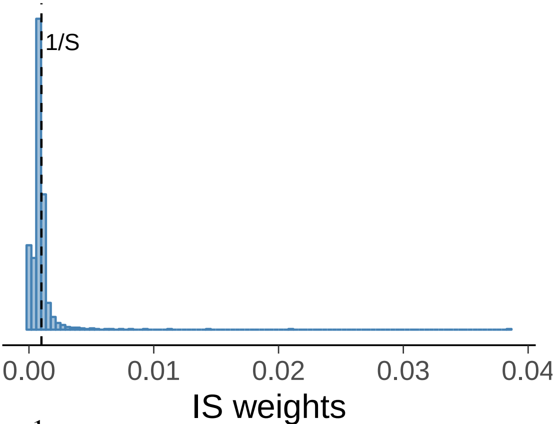


$$\text{ESS} = \frac{1}{\sum_{s=1}^{S} (\tilde{w}(\theta^s))^2},$$ is based on variance of $\tilde{w}(\theta^s)$

ESS ≈ 396

Pareto-$\hat{k} \approx 0.65$, CLT does not hold

Example: Importance sampling in Bioassay



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

ESS ≈ 396

Pareto-$\hat{k} \approx 0.65$, CLT does not hold

with Pareto-smoothing the estimate would be fine if $\hat{k} < 0.7$

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

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

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)

Curse of dimensionality

- Number of grid points increases exponentially
- Concentration of the measure, that is, 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