Chapter 4

- 4.1 Normal approximation (Laplace's method)
- 4.2 Large-sample theory
- 4.3 Counter examples
  - includes examples of difficult posteriors for MCMC, too
- 4.4 Frequency evaluation*
- 4.5 Other statistical methods*

Normal approximation (Laplace approximation)

- Often posterior converges to normal distribution when $n \to \infty$
- If posterior is unimodal and close to symmetric
  - we can approximate $p(\theta|y)$ with normal distribution

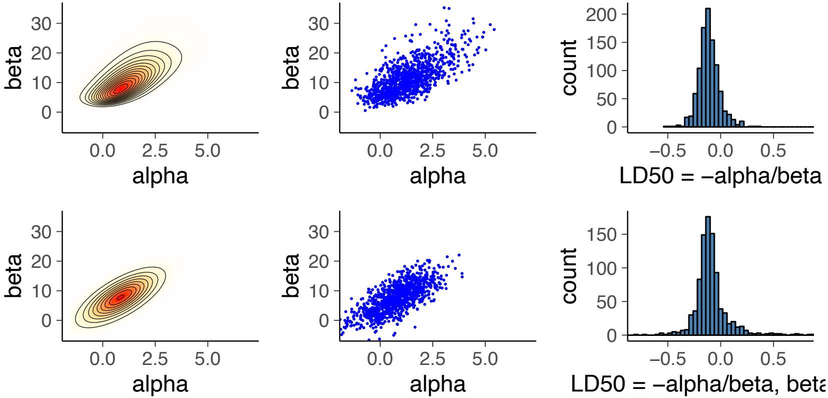
$$p(\theta|y) \approx \frac{1}{\sqrt{2\pi}\sigma_\theta} \exp\left(-\frac{1}{2\sigma_\theta^2}(\theta - \hat{\theta})^2\right)$$

  - Laplace used this (before Gauss) to approximate the posterior of binomial model to infer ratio of girls and boys born
  - A most strict proof by LeCam in 1950's

Normal approximation (Laplace approximation)

- Often posterior converges to normal distribution when $n \to \infty$
- If posterior is unimodal and close to symmetric
    - we can approximate $p(\theta|y)$ with normal distribution

$$p(\theta|y) \approx \frac{1}{\sqrt{2\pi}\sigma_\theta} \exp\left(-\frac{1}{2\sigma_\theta^2}(\theta - \hat{\theta})^2\right)$$



Taylor series

- We can approximate $p(\theta|y)$ with normal distribution

$$p(\theta|y) \approx \frac{1}{\sqrt{2\pi}\sigma_\theta} \exp\left(-\frac{1}{2\sigma_\theta^2}(\theta - \hat{\theta})^2\right)$$

  - i.e. log posterior $\log p(\theta|y)$ can be approximated with a quadratic function

$$\log p(\theta|y) \approx \alpha(\theta - \hat{\theta})^2 + C$$

- Univariate Taylor series expansion around $\theta = \hat{\theta}$

$$f(\theta) = f(\hat{\theta}) + f'(\hat{\theta})(\theta - \hat{\theta}) + \frac{f''(\hat{\theta})}{2!}(\theta - \hat{\theta})^2 + \frac{f^{(3)}(\hat{\theta})}{3!}(\theta - \hat{\theta})^3 + \ldots$$

  - if $\hat{\theta}$ is at mode, then $f'(\hat{\theta}) = 0$
  - often when $n \to \infty$, $\frac{f^{(3)}(\hat{\theta})}{3!}(\theta - \hat{\theta})^3 + \ldots$ is small

Multivariate Taylor series

- Multivariate series expansion

$$f(\theta) = f(\hat{\theta}) + \frac{df(\theta')}{d\theta'}_{\theta'=\hat{\theta}}(\theta - \hat{\theta}) + \frac{1}{2!}(\theta - \hat{\theta})^T \frac{d^2 f(\theta')}{d\theta'^2}_{\theta'=\hat{\theta}}(\theta - \hat{\theta}) + \ldots$$

Normal approximation

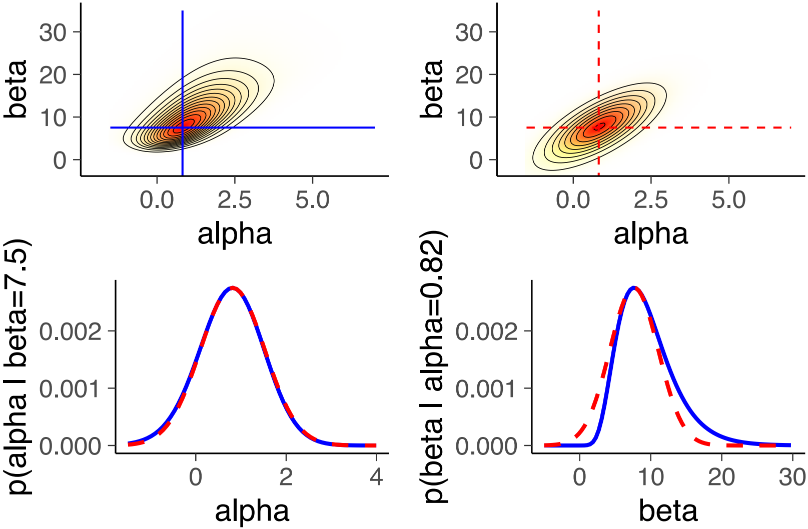
- Taylor series expansion of the log posterior around the posterior mode $\hat{\theta}$

$$\log p(\theta|y) = \log p(\hat{\theta}|y) + \frac{1}{2}(\theta - \hat{\theta})^T \left[\frac{d^2}{d\theta^2} \log p(\theta'|y)\right]_{\theta'=\hat{\theta}} (\theta - \hat{\theta}) + \dots$$

- Multivariate normal $\propto |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(\theta - \hat{\theta}^T)\Sigma^{-1}(\theta - \hat{\theta})\right)$

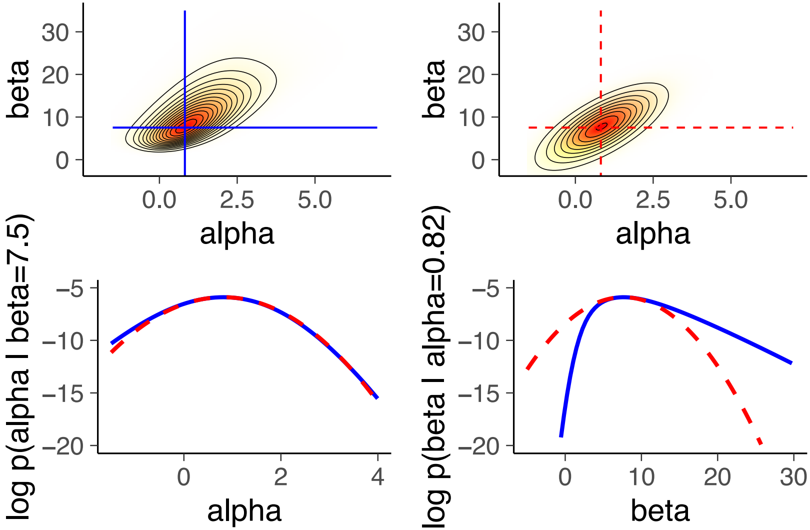
Normal approximation

- Multivariate normal $\propto |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(\theta - \hat{\theta}^T)\Sigma^{-1}(\theta - \hat{\theta})\right)$



Normal approximation

- Multivariate normal $\propto |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(\theta - \hat{\theta}^T)\Sigma^{-1}(\theta - \hat{\theta})\right)$



Normal approximation

- Taylor series expansion of the log posterior around the posterior mode $\hat{\theta}$

$$\log p(\theta|y) = \log p(\hat{\theta}|y) + \frac{1}{2}(\theta-\hat{\theta})^T \left[\frac{d^2}{d\theta^2}\log p(\theta'|y)\right]_{\theta'=\hat{\theta}} (\theta-\hat{\theta}) + \ldots$$

- Multivariate normal $\propto |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(\theta-\hat{\theta}^T)\Sigma^{-1}(\theta-\hat{\theta})\right)$
- Normal approximation

$$p(\theta|y) \approx \mathrm{N}(\hat{\theta}, [I(\hat{\theta})]^{-1})$$

where $I(\theta)$ is called *observed information*

$$I(\theta) = -\frac{d^2}{d\theta^2}\log p(\theta|y)$$

Normal approximation

- $I(\theta)$ is called *observed information*

$$I(\theta) = -\frac{d^2}{d\theta^2} \log p(\theta|y)$$

  - $I(\hat{\theta})$ is the second derivatives at the mode and thus describes the curvature at the mode
  - if the mode is inside the parameter space, $I(\hat{\theta})$ is positive
  - if θ is a vector, then $I(\theta)$ is a matrix

Normal approximation

- BDA3 Ch 4 has an example where it is easy to compute first and second derivatives and there is easy analytic solution to find where the first derivatives are zero

Normal approximation – example

- Normal distribution, unknown mean and variance
  - uniform prior $(\mu, \log \sigma)$
  - normal approximation for the posterior of $(\mu, \log \sigma)$

$$\log p(\mu, \log \sigma | y) = \text{constant} - n \log \sigma - \frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y} - \mu)^2]$$

first derivatives

$$\begin{aligned} \frac{d}{d\mu} \log p(\mu, \log \sigma | y) &= \frac{n(\bar{y} - \mu)}{\sigma^2}, \\ \frac{d}{d(\log \sigma)} \log p(\mu, \log \sigma | y) &= -n + \frac{(n-1)s^2 + n(\bar{y} - \mu)^2}{\sigma^2}, \end{aligned}$$

from which it is easy to compute the mode

$$(\hat{\mu}, \log \hat{\sigma}) = \left(\bar{y}, \frac{1}{2} \log \left(\frac{n-1}{n} s^2 \right) \right)$$

Normal approximation – example

- Normal distribution, unknown mean and variance
  first derivatives

$$
\begin{aligned}
\frac{d}{d\mu} \log p(\mu, \log\sigma|y) &= \frac{n(\bar{y}-\mu)}{\sigma^2}, \\
\frac{d}{d(\log\sigma)} \log p(\mu, \log\sigma|y) &= -n + \frac{(n-1)s^2 + n(\bar{y}-\mu)^2}{\sigma^2}
\end{aligned}
$$

  second derivatives

$$
\begin{aligned}
\frac{d^2}{d\mu^2} \log p(\mu, \log\sigma|y) &= -\frac{n}{\sigma^2}, \\
\frac{d^2}{d\mu d(\log\sigma)} \log p(\mu, \log\sigma|y) &= -2n\frac{\bar{y}-\mu}{\sigma^2}, \\
\frac{d^2}{d(\log\sigma)^2} \log p(\mu, \log\sigma|y) &= -\frac{2}{\sigma^2}((n-1)s^2 + n(\bar{y}-\mu)^2)
\end{aligned}
$$

Normal approximation – example

- Normal distribution, unknown mean and variance
  second derivatives

$$
\begin{aligned}
\frac{d^2}{d\mu^2}\log p(\mu,\log\sigma|y) &= -\frac{n}{\sigma^2},\\
\frac{d^2}{d\mu(\log\sigma)}\log p(\mu,\log\sigma|y) &= -2n\frac{\bar{y}-\mu}{\sigma^2},\\
\frac{d^2}{d(\log\sigma)^2}\log p(\mu,\log\sigma|y) &= -\frac{2}{\sigma^2}((n-1)s^2+n(\bar{y}-\mu)^2)
\end{aligned}
$$

  matrix of the second derivatives at $(\hat{\mu},\log\hat{\sigma})$

$$
\begin{pmatrix} -n/\hat{\sigma}^2 & 0 \\ 0 & -2n \end{pmatrix}
$$

Normal approximation – example

- Normal distribution, unknown mean and variance
  posterior mode

$$(\hat{\mu}, \log \hat{\sigma}) = \left(\bar{y}, \frac{1}{2} \log \left(\frac{n-1}{n} s^2\right)\right)$$

  matrix of the second derivatives at $(\hat{\mu}, \log \hat{\sigma})$

$$\begin{pmatrix} -n/\hat{\sigma}^2 & 0 \\ 0 & -2n \end{pmatrix}$$

  normal approximation

$$p(\mu, \log \sigma | y) \approx \mathrm{N}\left(\begin{pmatrix} \mu \\ \log \sigma \end{pmatrix} \middle| \begin{pmatrix} \bar{y} \\ \log \hat{\sigma} \end{pmatrix}, \begin{pmatrix} \hat{\sigma}^2/n & 0 \\ 0 & 1/(2n) \end{pmatrix}\right)$$

Normal approximation – numerically

- Normal approximation can be computed numerically
  - iterative optimization to find a mode (may use gradients)
  - autodiff or finite-difference for gradients and Hessian
  - e.g. in R, demo4_1.R:

```
bioassayfun <- function(w, df) {
  z <- w[1] + w[2]*df$x
  -sum(df$y*(z) - df$n*log1p(exp(z)))
}

theta0 <- c(0,0)
optimres <- optim(w0, bioassayfun, gr=NULL, df1, hessian=T)
thetahat <- optimres$par
Sigma <- solve(optimres$hessian)
```

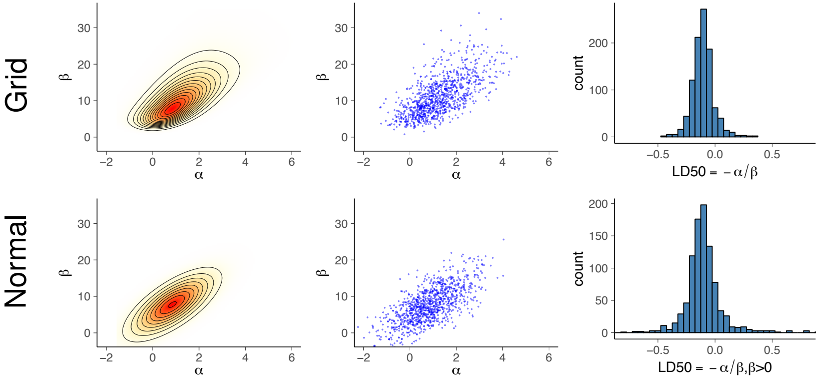
Normal approximation – numerically

- Normal approximation can be computed numerically
  - iterative optimization to find a mode (may use gradients)
  - autodiff or finite-difference for gradients and Hessian
- RStanARM has an option `algorithm='optimizing'`
  - uses L-BFGS quasi-Newton optimization algorithm for finding the mode
  - uses autodiff for gradients
  - uses finite differences of gradients to compute Hessian
    - second order autodiff coming to Stan

Normal approximation

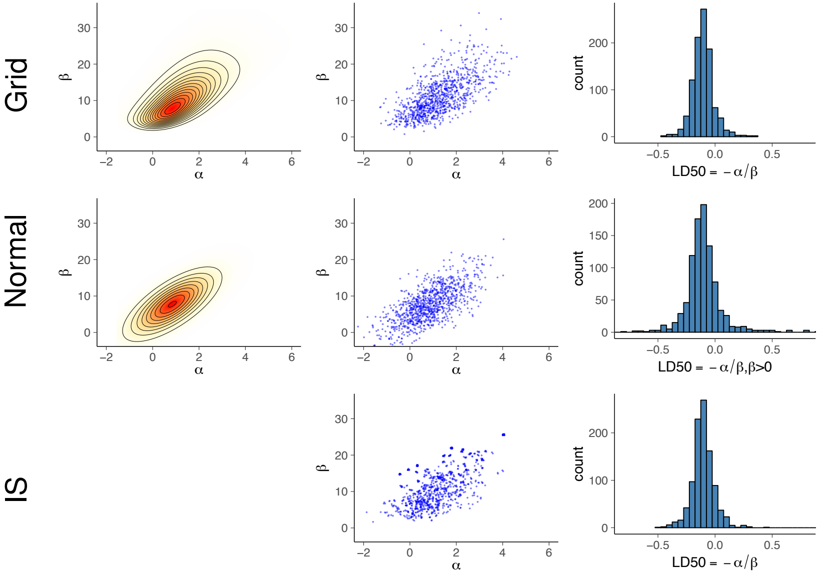
- Optimization and computation of Hessian requires usually much less density evaluations than MCMC
- In some cases accuracy is sufficient
- In some cases accuracy for a conditional distribution is sufficient (Ch 13)
  - e.g. Gaussian latent variable models, such as Gaussian processes (Ch 21)
  - Rasmussen & Williams: Gaussian Processes for Machine Learning
- Accuracy can be improved by importance sampling (Ch 10)

Example: Importance sampling in Bioassay



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



Grid sd(LD50) ≈ 0.1, IS sd(LD50) ≈ 0.1

Normal approximation

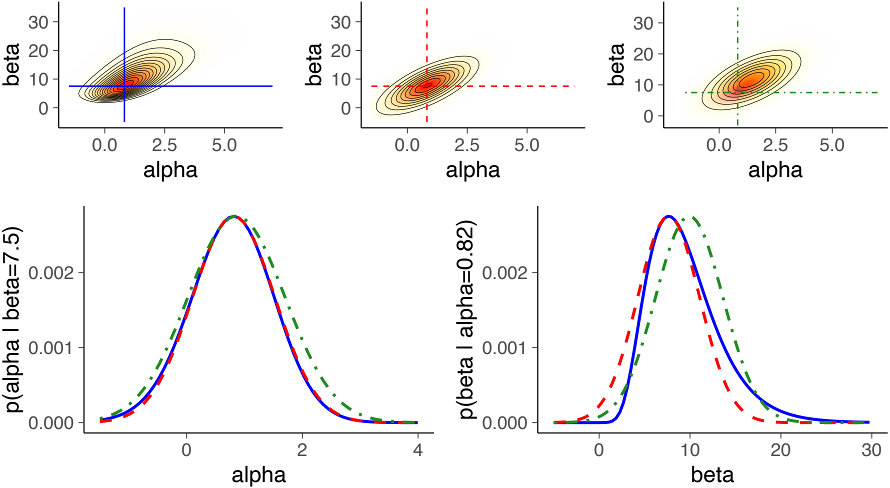
- Accuracy can be improved by importance sampling
- Pareto-k diagnostic of importance sampling weights can be used for diagnostic
  - in Bioassay example $k = 0.57$, which is ok
- RStanARM has an option `algorithm='optimizing'`
  - since version 2.19.2 (2019-10-03)
    + Pareto-k diagnostic
    + importance resampling (IR)

Other distributional approximations*

- Higher order derivatives at the mode can be used
- Split-normal and split-*t* by Geweke use additional scaling along different principal axes
- Other distributions can be used (e.g. *t*-distribution)
- Instead of mode and Hessian at mode, e.g.
  - variational inference (Ch 13)
    - CS-E4820 - Machine Learning: Advanced Probabilistic Methods
    - Stan has an experimental ADVI algorithm
  - expectation propagation (Ch 13)
  - speed of these is usually between optimization and MCMC

Distributional approximations

Exact, Normal at mode, Normal with variational inference



Grid sd(LD50) $\approx$ 0.090,
Normal sd(LD50) $\approx$ .75, Normal + IR sd(LD50) $\approx$ 0.096 (Pareto-k = 0.57)
VI sd(LD50) $\approx$ 0.13, VI + IR sd(LD50) $\approx$ 0.095 (Pareto-k = 0.17)

Large sample theory

- Asymptotic normality
  - as n the number of observations y_i increases the posterior converges to normal distribution

Large sample theory

- Assume "true" underlying data distribution $f(y)$
  - observations $y_1, \ldots, y_n$ are independent samples from the joint distribution $f(y)$
  - "true" data distribution $f(y)$ is not always well defined
  - in the following we proceed as if there were true underlying data distribution
  - for the theory the exact form of $f(y)$ is not important as long at it has certain regularity conditions

Large sample theory

- Consistency
  - if true distribution is included in the parametric family, so that $f(y) = p(y|\theta_0)$ for some θ_0, then posterior converges to a point θ_0, when $n \to \infty$
  - a point doesn't have uncertainty
  - same result as for maximum likelihood estimate
- If true distribution is not included in the parametric family, then there is no true θ_0
  - true θ_0 is replaced with θ_0 which minimizes the Kullback-Leibler divergence from $f(y)$

$$H(\theta_0) = \int f(y_i) \log \left(\frac{f(y_i)}{p(y_i|\theta_0)} \right) dy_i$$

  - this point doesn't have uncertainty, but it's a wrong point!
  - same result as for maximum likelihood estimate

Large sample theory – counter examples

- Under- and non-identifiability
  - a model is under-identifiable, if the model has parameters or parameter combinations for which there is no information in the data
  - then there is no single point θ_0 where posterior would converge
  - e.g. if the model is

$$y \sim \mathrm{N}(a + b + cx, \sigma)$$

    - posterior would converge to a line with prior determining the density along the line
  - e.g. if we never observe u and v at the same time and the model is

$$\begin{pmatrix} u \\ v \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

    then correlation ρ is non-identifiable
    - e.g. u and v could be length and weight of a student; if only one of them is measured for each student, then ρ is non-identifiable
- Problem also for other inference methods like MCMC

Large sample theory – counter examples

- If the number of parameter increases as the number of observation increases
  - in some models number of parameters depends on the number of observations
  - e.g. time series models $y_i \sim \mathrm{N}(\theta_i, \sigma^2)$ and θ_i has prior in time
  - posterior of θ_i does not converge to a point, if additional observations do not bring enough information

Large sample theory – counter examples

- Aliasing (FI: valetoisto)
  - special case of under-identifiability where likelihood repeats in separate points
  - e.g. mixture of normals

$$p(y_i|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \lambda) = \lambda \, \mathrm{N}(\mu_1, \sigma_1^2) + (1 - \lambda) \, \mathrm{N}(\mu_2, \sigma_2^2)$$

    if (μ_1, μ_2) are switched, (σ_1^2, σ_2^2) are switched and replace λ with $(1 - \lambda)$, model is equivalent; posterior would usually have two modes which are mirror images of each other and the posterior does not converge to a single point

- For MCMC makes the convergence diagnostics more difficult, as it is difficult to identify aliasing from other multimodality

Large sample theory – counter examples

- Unbounded (FI: rajoittamaton) likelihood
  - if likelihood is unbounded it is possible that there is no mode in the posterior
  - e.g. previous normal mixture model; assume λ to be known (and not 0 or 1); if we set $\mu_1 = y_i$ for any i and $\sigma_1^2 \to 0$, then likelihood $\to \infty$
  - if prior for σ_1^2 does not go to zero when $\sigma_1^2 \to 0$, then the posterior is unbounded
  - when $n \to \infty$ the number of likelihood modes increases
- Problem for any inference method including MCMC
  - can be avoided with good priors
  - note that a prior close to a prior allowing unbounded posterior may produce almost unbounded posterior

Large sample theory – counter examples

- Improper posterior
  - asymptotic results assume that probability sums to 1
  - e.g. Binomial model, with Beta$(0, 0)$ prior and observation $y = n$
    - posterior $p(\theta|n, 0) = \theta^{n-1}(1 - \theta)^{-1}$
    - when $\theta \to 1$, then $p(\theta|n, 0) \to \infty$
- Problem for any inference method including MCMC
  - can be avoided with proper priors
  - note that prior close to a improper prior may produce almost improper posterior

Large sample theory – counter examples

- Prior distribution does not include the convergence point
  - if in discrete case $p(\theta_0) = 0$ or in continuous case $p(\theta) = 0$ in the neighborhood of θ_0, then the convergence results based on the dominance of the likelihood do not hold
- Should have a positive prior probability/density where needed

Large sample theory – counter examples

- Convergence point at the edge of the parameter space
  - if θ_0 is on the edge of the parameter space, Taylor series expansion has to be truncated, and normal approximation does not necessarily hold
  - e.g. $y_i \sim \mathrm{N}(\theta, 1)$ with a restriction $\theta \geq 0$ and assume that $\theta_0 = 0$
    - posterior of θ is left truncated normal distribution with $\mu = \bar{y}$
    - in the limit $n \to \infty$ posterior is half normal distribution
- Can be easy or difficult for MCMC

Large sample theory – counter examples

- Tails of the distribution
  - normal approximation may be accurate for the most of the posterior mass, but still be inaccurate for the tails
  - e.g. parameter which is constrained to be positive; given a finite n, normal approximation assumes non-zero probability for negative values

Frequency evaluations

- Bayesian theory has epistemic and aleatory probabilities
- Frequency evaluations focus on frequency properties given aleatoric repetition of an observation and modeling
  - Consistency
  - Asymptotic unbiasedness
    - not that important in Bayesian inference, small and decreasing error more important
  - Asymptotic efficiency
    - no other point estimate with smaller squared error

Frequency evaluations

- Bayesian theory has epistemic and aleatory probabilities
- Frequency evaluations focus on frequency properties given aleatoric repetition of an observation and modeling
  - Consistency
  - Asymptotic unbiasedness
    - not that important in Bayesian inference, small and decreasing error more important
  - Asymptotic efficiency
    - no other point estimate with smaller squared error
  - Calibration
    - α%-posterior interval has the true value in α% cases
    - α%-predictive interval has the true future values in α% cases
    - approximate calibration with shorter intervals for likely true values more important than exact calibration with bad intervals for all possible values.