

# Data Asymptotics

Dr. Jarad Niemi

STAT 544 - Iowa State University

February 8, 2024

## Normal approximation to the posterior

Suppose  $p(\theta|y)$  is unimodal and roughly symmetric, then a Taylor series expansion of the logarithm of the posterior around the posterior mode  $\hat{\theta}$  is

$$\log p(\theta|y) = \log p(\hat{\theta}|y) - \frac{1}{2}(\theta - \hat{\theta})^\top \left[ -\frac{d^2}{d\theta^2} \log p(\theta|y) \right]_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \dots$$

where the linear term in the expansion is zero because the derivative of the log-posterior density is zero at its mode.

Discarding the higher order terms, this expansion provides a normal approximation to the posterior, i.e.

$$p(\theta|y) \stackrel{d}{\approx} N(\hat{\theta}, J(\hat{\theta})^{-1})$$

where  $J(\hat{\theta})$  is the sum of the prior and observed information, i.e.

$$J(\hat{\theta}) = -\frac{d^2}{d\theta^2} \log p(\theta)|_{\theta=\hat{\theta}} - \frac{d^2}{d\theta^2} \log p(y|\theta)|_{\theta=\hat{\theta}}.$$

## Binomial probability

Let  $y \sim \text{Bin}(n, \theta)$  and  $\theta \sim \text{Be}(a, b)$ , then  $\theta|y \sim \text{Be}(a + y, b + n - y)$  and the posterior mode is

$$\hat{\theta} = \frac{y'}{n'} = \frac{a + y - 1}{a + b + n - 2}.$$

Thus

$$J(\hat{\theta}) = \frac{n'}{\hat{\theta}(1 - \hat{\theta})}.$$

Thus

$$p(\theta|y) \stackrel{d}{\approx} N\left(\hat{\theta}, \frac{\hat{\theta}(1 - \hat{\theta})}{n'}\right).$$

# Binomial probability

```
a <- b <- 1      # Prior
n <- 10; y <- 3   # Data (attempts, successes)

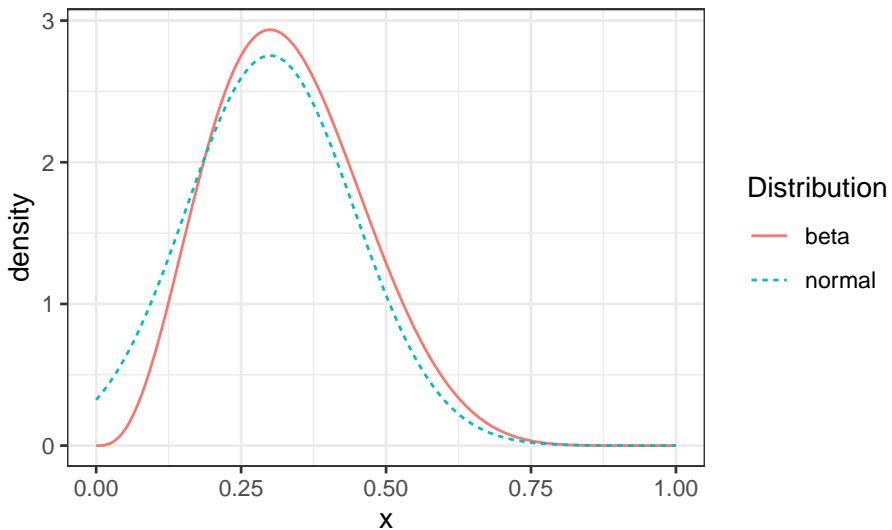
yp <- a + y - 1; np <- a + b + n - 2
theta_hat = yp / np # mode of a beta

d <- data.frame(x = seq(0, 1, length = 1001)) |>
  mutate(beta = dbeta(x,a+y,b+n-y),
         normal = dnorm(x, theta_hat, sqrt(theta_hat*(1-theta_hat)/np))) |>
  pivot_longer(beta:normal, names_to = "Distribution", values_to = "density")

ggplot(d, aes(x = x, y = density, color = Distribution, linetype = Distribution)) +
  geom_line()
```

<https://youtu.be/cRhD9FbSb34>

# Binomial probability



# Large-sample theory

Consider a model  $y_i \stackrel{iid}{\sim} p(y|\theta_0)$  for some true value  $\theta_0$ .

- Does the posterior distribution converge to  $\theta_0$ ?
- Does a point estimator (mode) converge to  $\theta_0$ ?
- What is the limiting posterior distribution?

# Convergence of the posterior distribution

Consider a model  $y_i \stackrel{iid}{\sim} p(y|\theta_0)$  for some true value  $\theta_0$ .

## Theorem

*If the parameter space  $\Theta$  is discrete and  $Pr(\theta = \theta_0) > 0$ , then  $Pr(\theta = \theta_0|y) \rightarrow 1$  as  $n \rightarrow \infty$ .*

## Theorem

*If the parameter space  $\Theta$  is continuous and  $A$  is a neighborhood around  $\theta_0$  with  $Pr(\theta \in A) > 0$ , then  $Pr(\theta \in A|y) \rightarrow 1$  as  $n \rightarrow \infty$ .*

```

n <- 1000
theta0 <- 0.3
d <- data.frame(
  n      = 1:n,
  y      = rbinom(n, 1, prob = 0.3))

dt <- expand_grid(d, theta = seq(0.1, 0.9, by = 0.1)) |>
  mutate(
    log_prob = dbinom(y, 1, prob = theta, log = TRUE),
  ) |>
  group_by(theta) |>
  arrange(n) |>
  mutate(
    log_prob = cumsum(log_prob)
  ) |>
  group_by(n) |>
  mutate(
    log_prob = log_prob - max(log_prob),
    prob      = exp(log_prob),
    prob      = prob / sum(prob),
    theta     = factor(theta)
  )

ggplot(dt, aes(x = n, y = prob,
               color = theta, group = theta)) +
  geom_line() +
  labs(
    x = "Sample size",
    y = "Posterior probability",
    title = "Posterior convergence for discrete distribution"
  )

```



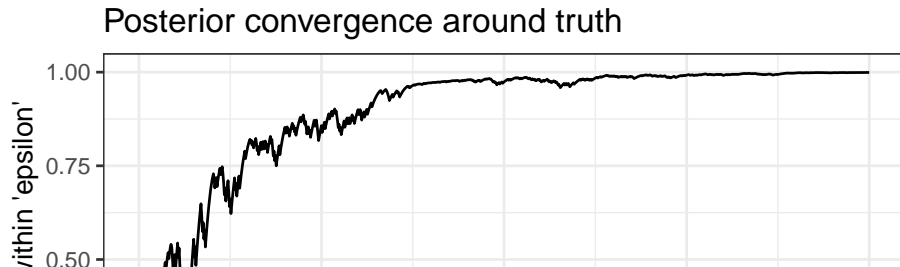
```

a <- b <- 1 # prior
e <- 0.05   # window half-width

# Calculate  $P(\theta_0 - e < \theta < \theta_0 + e \mid y)$ 
dc <- d |> mutate(y = cumsum(y),
                  prob = pbeta(theta0 + e, a + y, b + n - y) -
                        pbeta(theta0 - e, a + y, b + n - y))

# Plot calculated probability as a function of sample size
ggplot(dc, aes(x = n, y = prob)) +
  geom_line() +
  labs(
    x = "Sample size",
    y = "Probability within 'epsilon'",
    title = "Posterior convergence around truth"
  ) +
  ylim(0,1)

```



# Consistency of Bayesian point estimates

Suppose  $y_i \stackrel{iid}{\sim} p(y|\theta_0)$  where  $\theta_0$  is a particular value for  $\theta$ .

Recall that an estimator is consistent, i.e.  $\hat{\theta} \xrightarrow{p} \theta_0$ , if

$$\lim_{n \rightarrow \infty} P(|\hat{\theta} - \theta_0| < \epsilon) = 1.$$

Recall, under regularity conditions that  $\hat{\theta}_{MLE} \xrightarrow{p} \theta_0$ . If Bayesian estimators converge to the MLE, then they have the same properties.

## Binomial example

Consider  $y \sim \text{Bin}(n, \theta)$  with true value  $\theta = \theta_0$  and prior  $\theta \sim \text{Be}(a, b)$ . Then  $\theta|y \sim \text{Be}(a + y, b + n - y)$ .

Recall that  $\hat{\theta}_{MLE} = y/n$ . The following estimators are all consistent

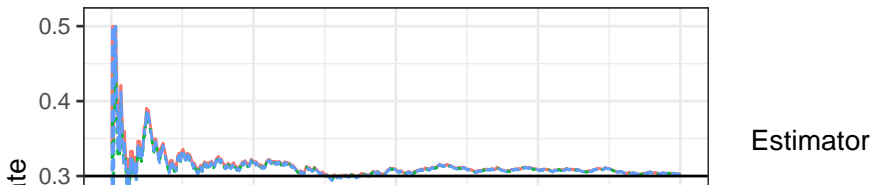
- Posterior mean:  $\frac{a+y}{a+b+n}$
- Posterior median:  $\approx \frac{a+y-1/3}{a+b+n-2/3}$  for  $\alpha, \beta > 1$
- Posterior mode:  $\frac{a+y-1}{a+b+n-2}$

since as  $n \rightarrow \infty$ , these all converge to  $\hat{\theta}_{MLE} = y/n$ .

```
# Calculate posterior mean, median, and mode
dbc <- dc |>
  mutate(
    mean    = (a + y) / (a + b + n),
    median  = qbeta(0.5, a + y, a + b + n - y),
    mode    = (a + y - 1) / (a + b + n - 2)
  ) |>
  pivot_longer(mean:mode, names_to = "Estimator", values_to = "estimate")

# Plot estimates vs sample size
ggplot(dbc, aes(x = n, y = estimate,
               color = Estimator, linetype = Estimator, group = Estimator)) +
  geom_line() +
  geom_hline(yintercept = theta0) +
  labs(
    x = "Sample size",
    y = "Estimate",
    title = "Binomial: Bayesian Estimator Convergence"
  )
)
```

## Binomial: Bayesian Estimator Convergence



## Normal example

Consider  $Y_i \stackrel{iid}{\sim} N(\theta, 1)$  with known and prior  $\theta \sim N(c, 1)$ . Then

$$\theta|y \sim N\left(\frac{1}{n+1}c + \frac{n}{n+1}\bar{y}, \frac{1}{n+1}\right)$$

Recall that  $\hat{\theta}_{MLE} = \bar{y}$ . Since the posterior mean converges to the MLE, then the posterior mean (as well as the median and mode) are consistent.

```
Error in v * c: non-numeric argument to binary operator
Error in 'geom_line()':
! Problem while computing aesthetics.
i Error occurred in the 1st layer.
Caused by error:
! object 'yhat' not found
```

# Asymptotic normality

Consider the Taylor series expansion of the log posterior

$$\log p(\theta|y) = \log p(\hat{\theta}|y) - \frac{1}{2}(\theta - \hat{\theta})^\top \left[ -\frac{d^2}{d\theta^2} \log p(\theta|y) \right]_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + R$$

where the linear term is zero because the derivative at the posterior mode  $\hat{\theta}$  is zero and  $R$  represents all higher order terms.

With iid observations, the coefficient for the quadratic term can be written as

$$-\frac{d^2}{d\theta^2} [\log p(\theta|y)]_{\theta=\hat{\theta}} = -\frac{d^2}{d\theta^2} \log p(\theta)_{\theta=\hat{\theta}} - \sum_{i=1}^n \frac{d^2}{d\theta^2} [\log p(y_i|\theta)]_{\theta=\hat{\theta}}$$

where

$$E_y \left[ -\frac{d^2}{d\theta^2} [\log p(y_i|\theta)]_{\theta=\hat{\theta}} \right] = I(\theta_0)$$

where  $I(\theta_0)$  is the expected Fisher information and thus, by the LLN, the second term converges to  $nI(\theta_0)$ .

# Bernstein-von Mises Theorem

For large  $n$ , we have

$$\log p(\theta|y) \approx \log p(\hat{\theta}|y) - \frac{1}{2}(\theta - \hat{\theta})^\top [nI(\theta_0)] (\theta - \hat{\theta})$$

where  $\hat{\theta}$  is the posterior mode.

If  $\hat{\theta} \rightarrow \theta_0$  as  $n \rightarrow \infty$ ,  $I(\hat{\theta}) \rightarrow I(\theta_0)$  as  $n \rightarrow \infty$  and we have

$$p(\theta|y) \propto \exp \left( -\frac{1}{2}(\theta - \hat{\theta})^\top [nI(\hat{\theta})] (\theta - \hat{\theta}) \right).$$

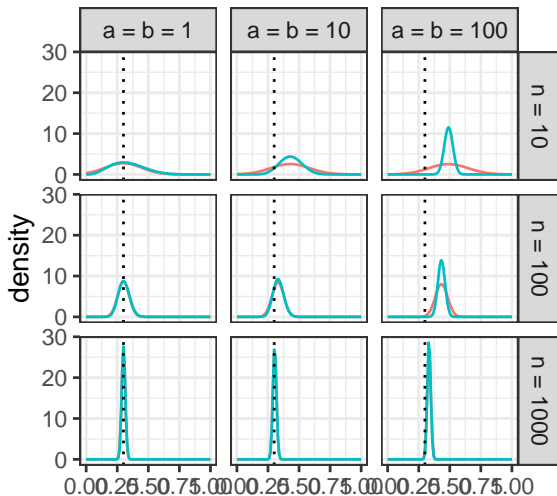
Thus, as  $n \rightarrow \infty$

$$\theta|y \xrightarrow{d} N \left( \hat{\theta}, \frac{1}{n}I(\hat{\theta})^{-1} \right)$$

Thus, the posterior distribution is asymptotically normal.

# Binomial example

Suppose  $y \sim \text{Bin}(n, \theta)$  and  $\theta \sim \text{Be}(a, b)$ .



## Distribution

- Normal approximation
- True posterior



# What can go wrong?

- Not unique to Bayesian statistics
  - Unidentified parameters
  - Number of parameters increase with sample size
  - Aliasing
  - Unbounded likelihoods
  - Tails of the distribution
  - True sampling distribution is not  $p(y|\theta)$
- Unique to Bayesian statistics
  - Improper posterior
  - Prior distributions that exclude the point of convergence
  - Convergence to the edge of the (prior) parameter space

# Truncated priors

Suppose

$$Y \sim \text{Bin}(n, \theta)$$

and the true value for  $\theta$  is

$$\theta_0 = 0.3.$$

Your belief is that there is no way  $\theta$  is less than 0.5 and thus you assign a truncated beta distribution for a prior, i.e.

$$\theta \sim \text{Be}(a, b)\text{I}(\theta > 0.5).$$

The posterior is then

$$\theta|y \sim \text{Be}(a + y, b + n - y)\text{I}(\theta > 0.5).$$

The following occurs:

- the posterior will not converge to a neighborhood around  $\theta_0$ ,
- no Bayesian estimators will converge to  $\theta_0$ , and
- the posterior will not converge to a normal distribution.

## True sampling distribution is not $p(y|\theta)$

Suppose that  $f(y)$ , the true sampling distribution, does not correspond to  $p(y|\theta)$  for any  $\theta = \theta_0$ .

Then the posterior  $p(\theta|y)$  converges to a  $\theta_0$  that is the smallest in Kullback-Leibler divergence to the true  $f(y)$  where

$$KL(f(y)||p(y|\theta)) = E \left[ \log \left( \frac{f(y)}{p(y|\theta)} \right) \right] = \int \log \left( \frac{f(y)}{p(y|\theta)} \right) f(y) dy.$$

That is, we do about the best that we can given that we have assumed the wrong sampling distribution  $p(y|\theta)$ .