# Data Asymptotics

Dr. Jarad Niemi

Iowa State University

February 1, 2016

# 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

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

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

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

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

where $\mathrm{J}(\hat{\theta})$ is the observed information, i.e.

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

# Binomial probability

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

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

Thus

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

Thus

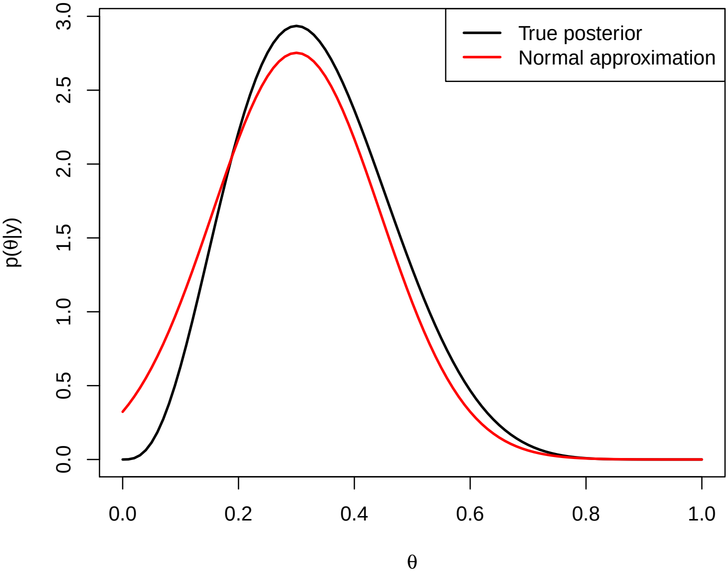
$$p(\theta|y) \overset{d}{\approx} N\left(\hat{\theta}, \frac{\hat{\theta}(1-\hat{\theta})}{n'}\right).$$

# Binomial probability

```
a = b = 1
n = 10
y = 3
par(mar=c(5,4,0.5,0)+.1)
curve(dbeta(x,a+y,b+n-y), lwd=2, xlab=expression(theta), ylab=expression(paste("p(", theta,"|y)")))

# Approximation
yp = a+y-1
np = a+b+n-2
theta_hat = yp/np
curve(dnorm(x,theta_hat, sqrt(theta_hat*(1-theta_hat)/np)), add=TRUE, col="red", lwd=2)
legend("topright",c("True posterior","Normal approximation"), col=c("black","red"), lwd=2)
```

# 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 $\theta_0$.

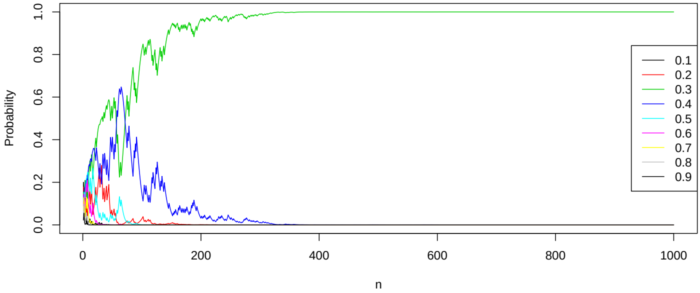
**Theorem**

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

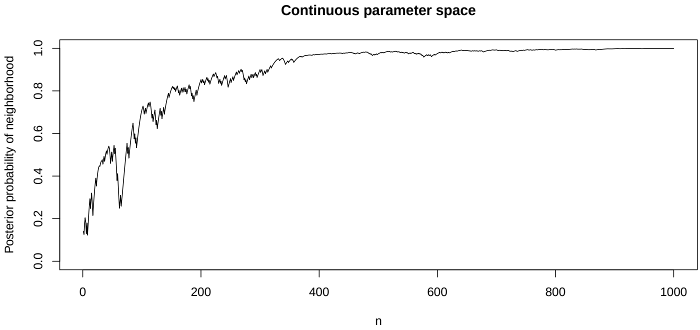
**Theorem**

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

```
library(smcUtils)
theta = seq(0.1,0.9, by=0.1); theta0 = 0.3
n = 1000
y = rbinom(n, 1, theta0)
p = matrix(NA, n,length(theta))
p[1,] = renormalize(dbinom(y[1],1,theta, log=TRUE), log=TRUE)
for (i in 2:n) {
  p[i,] = renormalize(dbinom(y[i],1,theta, log=TRUE)+log(p[i-1,]), log=TRUE)
}
plot(p[,1], ylim=c(0,1), type="l", xlab="n", ylab="Probability")
for (i in 1:length(theta)) lines(p[,i], col=i)
legend("right", legend=theta, col=1:9, lty=1)
```



```r
a = b = 1
e = 0.05
p = rep(NA,n)
for (i in 1:n) {
  yy = sum(y[1:i])
  zz = i-yy
  p[i] = diff(pbeta(theta0+c(-1,1)*e, a+yy, b+zz))
}
plot(p, type="l", ylim=c(0,1), ylab="Posterior probability of neighborhood",
     xlab="n", main="Continuous parameter space")
```



# 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. $\hat{\theta} \stackrel{p}{\to} \theta_0$, if

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

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

# Binomial example

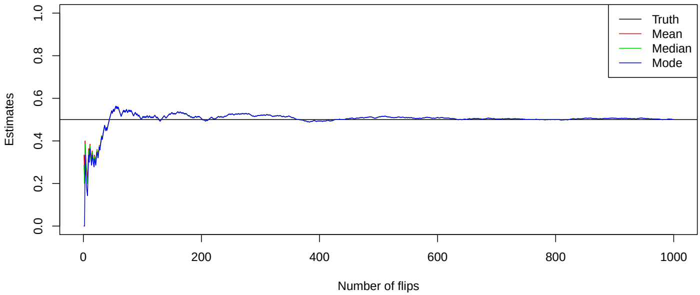
Consider $y \sim Bin(n, \theta)$ with true value $\theta = \theta_0$ and prior $\theta \sim Be(a, b)$. Then $\theta|y \sim Be(a + y, b + n - y)$.

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

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

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

```r
a = b = 1
n = 1000
theta0 = 0.5
y = rbinom(n, 1, theta0)
yy = cumsum(y)
nn = 1:n
plot(0,0, type="n", xlim=c(0,n), ylim=c(0,1), xlab="Number of flips", ylab="Estimates")
abline(h=theta0)
lines((a+yy)/(a+b+nn), col=2)
lines((a+yy-1/3)/(a+b+nn-2/3), col=3)
lines((a+yy-1)/(a+b+nn-2), col=4)
legend("topright",c("Truth","Mean","Median","Mode"), col=1:4, lty=1)
```

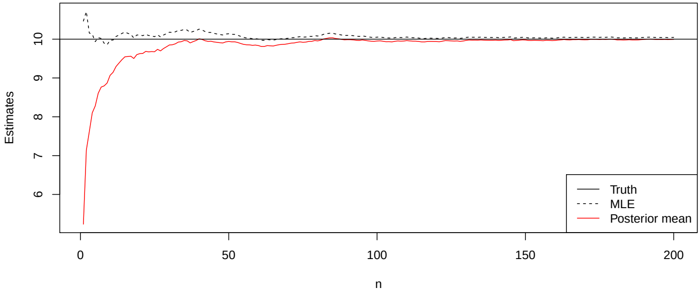


# Normal example

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

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

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



# 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.

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

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

where $\mathrm{I}(\theta_0)$ is the expected Fisher information and thus, by the LLN, the second term converges to $-n\mathrm{I}(\theta_0)$.

# Asymptotic normality

Since for large $n$

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

and since $\hat{\theta} \to \theta_0$, we have

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

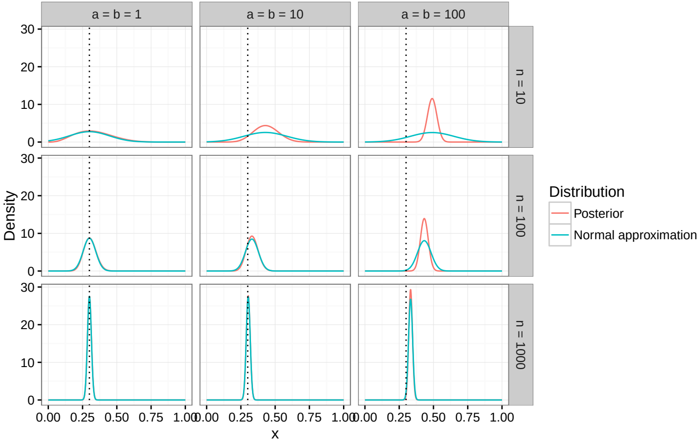
and thus, as $n \to \infty$

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

Thus, the posterior distribution is asymptotically normal.

# Binomial example

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



# 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 parameter space

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

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

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 $p(y|\theta)$.