Motivation for Bayesian Estimation

Example: Toss a coin n times.
Estimate $\theta = P(\text{heads})$.

$X_1, X_2, \ldots, X_n$ IID Bernoulli(θ).

$\hat{\theta}_{MLE} = \hat{\theta}_{MoM} = \bar{X}$.

Situation: Toss "normal looking" coin
$n = 3$ times.
Is $\bar{x}$ reasonable estimate of θ?
Suppose $\bar{x} = 0$ or 1?
Note: $P(\bar{X} = 0 \text{ or } 1) = \frac{1}{4}$
when $\theta = \frac{1}{2}$.

We have strong prior knowledge about θ:
Expect θ to be very close to 1/2.

How can we incorporate prior knowledge into our estimation procedure?

(Subjective) Bayesian Viewpoint

Unknown quantities (parameters)

⇓

(Uncertain)

should be viewed as random.

Uncertainty should be described by probability distributions.

$\pi(\theta)$ = prior

describes my prior belief about θ.

$f(\underset{\sim}{x}|\theta)$ = conditional density (pmf) for $\underset{\sim}{X}$ given θ

$= L(\theta|\underset{\sim}{x})$ = likelihood function.

"Experiment" consists of:

(1) Nature chooses $\theta \sim \pi$. (but doesn't tell us!)

(2) Scientist collects $\underset{\sim}{X} \sim f(\underset{\sim}{x}|\theta)$.

After collecting data $\underset{\sim}{x}$, our beliefs about θ are described by the posterior distn $\pi(\theta|\underset{\sim}{x})$. giving the conditional dist. of θ given $\underset{\sim}{X}$.

Formulas

Given (or dreamed up):

- $f(\underset{\sim}{x}|\theta)$ (family $\{P_\theta : \theta \in \Theta\}$)
- $\pi(\theta)$ (= prior = marginal for θ)

Joint density (or pmf)

$$f(\underset{\sim}{x}, \theta) = f(\underset{\sim}{x}|\theta)\,\pi(\theta)$$

$$\left[f_{X,Y}(x,y) = f_{X|Y}(x|y)\, f_Y(y) \right]$$

Marginal density (or pmf) of $\underset{\sim}{X}$

$$m(\underset{\sim}{x}) = \int_\Theta f(\underset{\sim}{x}|\theta)\,\pi(\theta)\,d\theta$$

$$\left(= \sum_\theta f(x|\theta)\,\pi(\theta) \right)$$

$$\left[f_X(x) = \int f_{X,Y}(x,y)\,dy \right]$$

Posterior density (or pmf) of θ $\left[f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} \right]$

$$\pi(\theta|\underset{\sim}{x}) = \frac{f(\underset{\sim}{x}|\theta)\,\pi(\theta)}{m(\underset{\sim}{x})} \propto f(\underset{\sim}{x}|\theta)\,\pi(\theta)$$

$$\propto \text{Likelihood} \times \text{Prior}$$

Notation: (Bayesian) Summary

$f(\underset{\sim}{x}|\theta)$

$\pi(\theta)$ [= prior = marginal for θ]

$f(\underset{\sim}{x},\theta) = f(\underset{\sim}{x}|\theta)\pi(\theta)$

$m(\underset{\sim}{x}) = \int f(\underset{\sim}{x}|\theta)\pi(\theta)\,d\theta$ [= marginal for x]

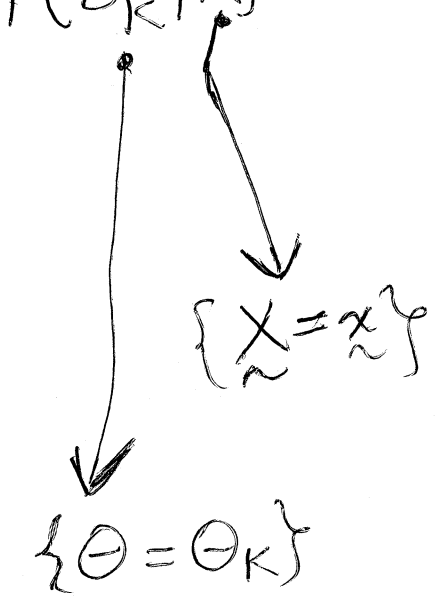
(or $\sum f(\underset{\sim}{x}|\theta)\pi(\theta)$)

$\pi(\theta|\underset{\sim}{x}) = \dfrac{f(\underset{\sim}{x}|\theta)\pi(\theta)}{m(\underset{\sim}{x})} \propto f(\underset{\sim}{x}|\theta)\pi(\theta)$

[= Likelihood x Prior]

If $\underset{\sim}{X}, \Theta$ are discrete, this is just Bayes Theorem:

$$P(B_K|A) = \frac{P(A|B_K)\,P(B_K)}{\sum_j P(A|B_j)\,P(B_j)}$$



Posterior = Likelihood × Prior, renormalized.

Conjugate Priors (Bernoulli trials)

$\theta \sim \pi$

conditional on θ,

$\underbrace{X_1, X_2, \ldots, X_n}_{\underset{\sim}{X}}$ are IID Bernoulli(θ).

$$\pi(\theta \mid \underset{\sim}{X}) \propto (\text{Likelihood}) \times (\text{Prior})$$
$$\propto \theta^t (1-\theta)^{n-t} \pi(\theta)$$

where $t = \sum X_i$

If $\pi(\theta) = \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$

$\hookrightarrow \pi \sim \text{Beta}(\alpha, \beta)$,

then $\pi(\theta \mid \underset{\sim}{X}) \sim \text{Beta}(\alpha+t, \beta+n-t)$

(Beta family is "closed under sampling". This is defining property of conjugate priors.)

Proof:

$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$\propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$L(\theta) = \theta^{t}(1-\theta)^{n-t}$$

($L(\theta|\underset{\sim}{x})$ versus $L(\theta|T)$, $T=\Sigma x_i$ ↗ will be proportional)

$$\pi(\theta|\underset{\sim}{x}) \propto [\theta^{t}(1-\theta)^{n-t}][\theta^{\alpha-1}(1-\theta)^{\beta-1}]$$

$$\propto \theta^{(\alpha+t)-1}(1-\theta)^{(\beta+n-t)-1}$$

$$\propto \text{Beta}(\alpha+t, \beta+n-t) \text{ distn.}$$

Parameters of prior	Params of posterior
(α, β)	$(\alpha+t, \beta+n-t)$

Parameters have been "updated".

Summarize distn. by giving mean and variance

Prior (also $\text{Mode}(\theta) = (\alpha-1)/(\alpha+\beta-2)$)

$$E\theta = \frac{\alpha}{\alpha+\beta}, \quad \text{Var}(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Posterior

$$E(\theta|\underset{\sim}{X}) = \frac{\alpha+t}{(\alpha+t)+(\beta+n-t)} = \frac{\alpha+t}{\alpha+\beta+n}$$

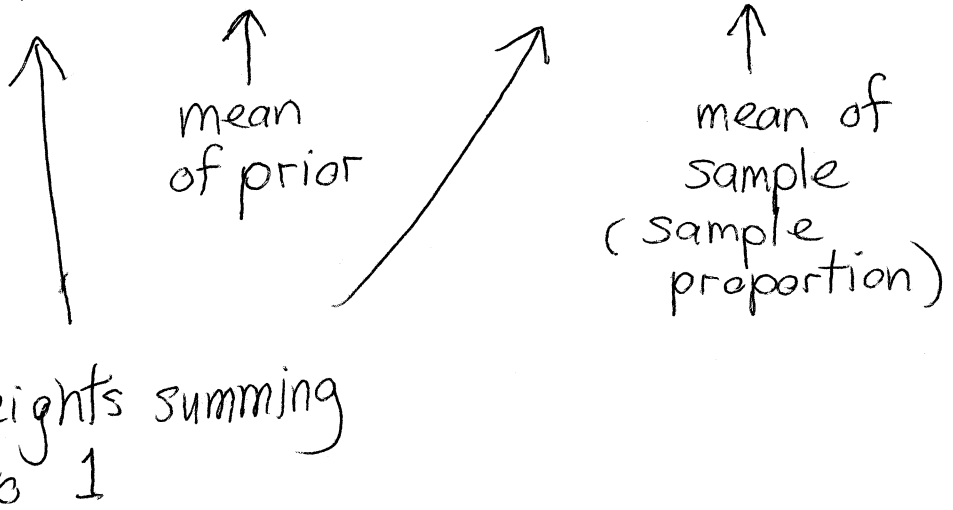
(Note: $\hat{\theta}_{\text{Bayes}} = E(\theta|\underset{\sim}{X})$ is usual Bayesian point estimator.)

$$\text{Var}(\theta|\underset{\sim}{X}) = \frac{(\alpha+t)(\beta+n-t)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)}$$

$$= \frac{\mu(1-\mu)}{\alpha+\beta+n+1} \quad \text{where } \mu = \frac{\alpha+t}{\alpha+\beta+n}$$

Observations

$$E(\theta|\underset{\sim}{X}) = \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\left(\frac{\alpha}{\alpha+\beta}\right) + \left(\frac{n}{\alpha+\beta+n}\right)\left(\frac{t}{n}\right)$$



Compromise between "prior" and "data".
Weighting depends on n.

For large n, $E(\theta|\underset{\sim}{X}) \approx \frac{t}{n}$,

$$Var(\theta|\underset{\sim}{X}) < \frac{1}{4n} \cdot \text{(small)}$$

Posterior is tightly concentrated around a mean which is close to t/n.
Holds regardless of α, β unless $\alpha+\beta$ large!

(The data "swamps" the prior.)

Let $\hat{p} = t/n$.

For large n, $E(\theta|\underset{\sim}{X}) \approx \hat{p}$

$$Var(\theta|\underset{\sim}{X}) \approx \frac{\hat{p}(1-\hat{p})}{n}.$$

Thus Bayesian reporting posterior mean and variance will be roughly agreeing with "frequentist" reporting $\hat{p}$ and its estimated variance.

Example: "Normal" coin.
Strong prior belief $\theta \approx 1/2$.
Represent by
$\pi \sim Beta(\alpha, \beta)$
with $\alpha = \beta$ large.
π tightly peaked around $1/2$.

Example: Bernoulli situation
with "vague" prior knowledge.
Might use flat prior: $\alpha = \beta = 1$.

Do this Earlier

These properties hold for any "reasonable" prior.

Example of Unreasonable Prior (Pigheaded)

$\pi(\theta) = 0$ for $\theta \in (a,b) \subset (0,1)$.

(Support of posterior) $\subset$ (Support of prior)

$$\pi(\theta|\underset{\sim}{x}) \propto \pi(\theta) f(\underset{\sim}{x}|\theta)$$

Another example

Point mass prior! Mass 1 at $\theta = 1/2$.

Reasonable prior has density $\pi(\theta)$ which is

smooth

support on all of $(0,1)$.

Weakness of Beta Priors
(more generally, conjugate priors)

Use of arbitrary priors
requires numerical integration

Mixtures of Beta priors (conjugate priors)

If $\pi(\theta) = p_1 f_1(\theta) + p_2 f_2(\theta)$

$f_1 \sim \text{Beta}(\alpha_1, \beta_1)$

$f_2 \sim \text{Beta}(\alpha_2, \beta_2)$

$p_1 + p_2 = 1$

then $\pi(\theta | \underset{\sim}{x}) = p_1^* f_1^*(\theta) + p_2^* f_2^*(\theta)$

$f_1^* \sim \text{Beta}(\alpha_1 + t, \beta_1 + n - t)$

$f_2^* \sim \text{Beta}(\alpha_2 + t, \beta_2 + n - t)$

$p_1^* + p_2^* = 1$

p_1^*, p_2^* depend on $\underset{\sim}{x}$. Easy to compute.

Example:

Suppose $\lambda \sim \pi$ and, conditional on λ,
$X_1, \ldots, X_n$ are iid Poisson(λ).

What is the family of conjugate priors in this situation?

Examine the likelihood function and see what priors fit well with it: multiplying likelihood times prior should produce something in the same family (ignoring constants).

$$L(\lambda \,|\, \boldsymbol{x}) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \frac{\lambda^{\sum x_i} e^{-n\lambda}}{\prod x_i!} \propto \lambda^t e^{-n\lambda} \quad \text{where } t = \sum_{i=1}^{n} x_i$$

which has the general form $\lambda^{\text{“}a\text{”}} e^{-\text{“}b\text{”}\lambda}$, which is the kernel of a gamma density (as a function of λ).

Textbook parameterization of Gamma density:

$$f(x \,|\, \alpha, \beta) = \frac{x^{\alpha-1} e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)} \propto x^{\alpha-1} e^{-x/\beta}$$

Another common parameterization (call it Gamma(a, b)) substitutes $b = 1/\beta$:

$$f(x \,|\, a, b) = \frac{b^a x^{a-1} e^{-bx}}{\Gamma(a)} \propto x^{a-1} e^{-bx}\,.$$

(Substitute $x \mapsto \lambda$ in both pdf's.)

Suppose the prior π is Gamma(a, b). (Using Gamma(a, b) leads to a somewhat simpler updating rule.) Then

$$\begin{aligned} \pi(\lambda \,|\, \boldsymbol{x}) &\propto L(\lambda \,|\, \boldsymbol{x}) \pi(\lambda) \propto \lambda^t e^{-n\lambda} \cdot \lambda^{a-1} e^{-b\lambda} \\ &= \lambda^{(a+t)-1} e^{-(b+n)\lambda} \propto \text{Gamma}(a+t, b+n) \text{ pdf} \end{aligned}$$

so that the posterior distn is Gamma($a+t, b+n$). The Gamma family is closed under sampling and forms a conjugate family.

Example: $X_1, X_2, \ldots, X_n$ iid $N(\theta, \sigma^2)$ (σ^2: known)

conjugate prior for θ?

Updating rule to obtain posterior?

$$L(\theta|X) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i-\theta)^2\right)$$

$$= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right)$$

($(2\pi\sigma^2)^{-n/2}$ → drop)

$$\propto \exp\left(-\frac{1}{2\sigma^2}\sum x_i^2 + \frac{\theta}{\sigma^2}\sum x_i - \frac{n\theta^2}{2\sigma^2}\right)$$

($-\frac{1}{2\sigma^2}\sum x_i^2$ → drop)

$$\propto \exp\left(-\frac{1}{2}\left(\frac{n}{\sigma^2}\theta^2 - \frac{2n\bar{x}}{\sigma^2}\theta\right)\right)$$

which has the general form

$$L(\theta|X) \propto \exp\left(-\frac{1}{2}(a\theta^2 - b\theta)\right), \quad a > 0.$$

Choose prior of the same form:

$$\pi(\theta) \propto \exp\left(-\frac{1}{2}(c\theta^2 - d\theta)\right), \quad c > 0.$$

Then $\pi(\theta|X) \propto L(\theta|X)\,\pi(\theta)$

$$\propto \exp\left(-\frac{1}{2}\left[(a+c)\theta^2 - (b+d)\theta\right]\right).$$

Easy updating rule!

What are these priors? Normal distns!

Lemma: $\pi(\theta) \propto \exp\left(-\frac{1}{2}(a\theta^2 - b\theta)\right), a > 0$

is the $N\left(+\frac{b}{2a}, \frac{1}{a}\right)$ pdf.

Proof: $a\theta^2 - b\theta = a\left(\theta^2 - \frac{b}{a}\theta\right)$

$$= a\left(\left(\theta - \frac{b}{2a}\right)^2 - \left(\frac{b}{2a}\right)^2\right)$$

$$= \frac{\left(\theta - \frac{b}{2a}\right)^2}{1/a} - a\left(\frac{b}{2a}\right)^2$$

Suppose Prior $\theta \sim N(\mu, \tau^2)$

Then $\pi(\theta) \propto \exp\left(-\frac{1}{2\tau^2}(\theta - \mu)^2\right)$

$$\propto \exp\left(-\frac{1}{2}\left(\frac{1}{\tau^2}\theta^2 - \frac{2\mu}{\tau^2}\theta\right)\right)$$

$\pi(\theta|x) \propto \pi(\theta) L(\theta)$

$$\propto \exp\left(-\frac{1}{2}\left[\left(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right)\theta^2 - \left(\frac{2\mu}{\tau^2} + \frac{2n\bar{x}}{\sigma^2}\right)\theta\right]\right)$$

$$\propto N\left(\frac{\frac{\mu}{\tau^2} + \frac{n\bar{x}}{\sigma^2}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}} , \frac{1}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}} \right)$$

Bayesian Estimation and Sufficient Statistics

Suppose $T(\boldsymbol{X})$ is a sufficient statistic for θ.

Fact: Posterior distributions depend on the data $\boldsymbol{X}$ only through the sufficient statistic $T(\boldsymbol{X})$.

Corollary: $E(\theta \,|\, \boldsymbol{X})$ and $\text{Var}(\theta \,|\, \boldsymbol{X})$ (and any other posterior quantity) depend on the data $\boldsymbol{X}$ only through $T(\boldsymbol{X})$.

Proof: By the FC, $\exists\, g, h$ such that $f(\boldsymbol{x} \,|\, \theta) = g(T(\boldsymbol{x}), \theta) h(\boldsymbol{x})$ for all $\boldsymbol{x}$ and θ. Thus

$$\begin{aligned} \pi(\theta \,|\, \boldsymbol{x}) &\propto f(\boldsymbol{x} \,|\, \theta)\pi(\theta) = g(T(\boldsymbol{x}), \theta) h(\boldsymbol{x}) \pi(\theta) \\ &\propto g(T(\boldsymbol{x}), \theta)\pi(\theta) \end{aligned}$$

so that

$$\pi(\theta \,|\, \boldsymbol{x}) = \frac{g(T(\boldsymbol{x}), \theta)\pi(\theta)}{\int_\Theta g(T(\boldsymbol{x}), \theta')\pi(\theta')\, d\theta'}$$

which depends on $\boldsymbol{x}$ only through $T(\boldsymbol{x})$.

A Simple Bayesian Hierarchical Model

Situation: 10 coins are sampled from a population of coins. Each coin is tossed 20 times. We desire to estimate $p_1, p_2, \ldots, p_{10}$, the probability of heads for each coin.

Let X_i be the number of heads in 20 tosses for coin i.

A Bayesian can incorporate prior knowledge about the similarity of the coins into the prior distribution.

Bayesian Model: ($i = 1, \ldots, 10$ throughout)

$$\begin{aligned} X_i &\sim \text{Binomial}(p_i\,, 20) \\ \eta_i &\sim \text{Normal}(\mu, 1/\tau) \\ &\quad \text{where } p_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}} \\ &\quad \text{and } \eta_i = \log\left(\frac{p_i}{1 - p_i}\right) \\ \mu &\sim \text{Normal}(0, 1)\,, \quad \tau \sim \text{Gamma}(10, b) \end{aligned}$$

The rv's at each level are conditionally independent given the rv's at lower levels.

$X = (X_1, X_2, \ldots, X_{10})\,, \qquad \theta = (\eta_1, \eta_2, \ldots, \eta_{10}, \mu, \tau)$

The posterior is $\pi(\theta \,|\, X) \propto$

$$\prod_{i=1}^{10} \binom{20}{X_i} p_i^{X_i} (1 - p_i)^{20 - X_i} \times \prod_{i=1}^{10} g(\eta_i \,|\, \mu, \tau)\, h(\mu)\, k(\tau)$$

where $g(\cdot\,|\mu, \tau)$ is the $N(\mu, 1/\tau)$ density, $h(\cdot)$ is the $N(0, 1)$ density, and $k(\cdot)$ is the Gamma$(10, b)$ density.

Note: Gamma(a, b) has mean a/b and variance a/b^2.

BUGS code describing the model:

```
model {

for(i in 1:N){
   x[i] ~ dbin(p[i],20)
   logit(p[i]) <- eta[i]
   eta[i] ~ dnorm(mu,tau)
}
mu ~ dnorm(0.0,1.0)
tau ~ dgamma(10.0,xxx) # Changing the scale only.

}
```

Data:

```
list(N=10,x=c(14,8,11,14,11,11,8,10,12,8))
```

Inits:

```
list(mu=0,tau=1,eta=c(0,0,0,0,0,0,0,0,0,0))
```

Story: There are 10 coins. Each is tossed 20 times.
The data
   x[1] = 14, x[2] = 8, ..., x[9] = 12, x[10] = 8
is the number of heads observed on each coin.

Goal: Estimate p[1], ..., p[10], the probability of heads for each coin.

Frequentist Answers

Assuming all p[i] are equal leads to:

phat = sum(x)/200 = 0.535

std. dev. of phat = sqrt(.535*(1-.535)/200) = 0.0353

Estimating each p[i] separately leads to:

p[1]hat = 14/20 = 0.7
p[2]hat = 8/20 = 0.4

std. dev. of p[1]hat = sqrt(.7*(1-.7)/20) = 0.1025
std. dev. of p[2]hat = sqrt(.4*(1-.4)/20) = 0.1095

Compare with the answers below.

Using: tau ~ dgamma(10.0,.001)

node	mean	sd	MC error	2.5%	median	97.5%
mu	0.1316	0.1296	0.008272	-0.1277	0.135	0.3998

node	mean	sd	MC error	2.5%	median	97.5%
tau	10001.3	3194.89	17.9655	4757.75	9670.29	17168.7

node	mean	sd	MC error	2.5%	median	97.5%
p[1]	0.5328	0.0322	0.00205	0.4680	0.5338	0.5986
p[2]	0.5326	0.0322	0.00205	0.4678	0.5336	0.5983

Using: tau ~ dgamma(10.0,10000)

node	mean	sd	MC error	2.5%	median	97.5%
mu	-0.0054	0.9929	0.004380	-1.9522	-0.0013	1.9352

node	mean	sd	MC error	2.5%	median	97.5%
tau	0.001497	3.86E-4	1.62E-6	8.377E-4	0.001464	0.002344

node	mean	sd	MC error	2.5%	median	97.5%
p[1]	0.7006	0.09994	4.151E-4	0.4871	0.7077	0.8744
p[2]	0.4008	0.10689	4.637E-4	0.2030	0.3975	0.6181

Using: tau ~ dgamma(10.0,2.0)

node	mean	sd	MC error	2.5%	median	97.5%
mu	0.1398	0.196	0.00191	-0.2462	0.1409	0.524

node	mean	sd	MC error	2.5%	median	97.5%
tau	5.512	1.603	0.01041	2.872	5.34	9.101

node	mean	sd	MC error	2.5%	median	97.5%
p[1]	0.6122	0.07851	5.701E-4	0.4547	0.614	0.7604
p[2]	0.4698	0.08052	5.129E-4	0.3135	0.4696	0.6262