MTH3041
Bayesian Statistics, philosophy and Practice

\Classical view/ • Probability is relative frequency (thing to be measured).
• data, y, generated 'at random' from a probability model parametered by a fixed unknown, θ. $P(y|\theta)$ completely specify the problem
• Data $\underline{y} = (y_1, \dots, y_n)$ (i.i.d) $P(\underline{y}|\theta)$ = likelihood
• inference by estimating true θ, maybe with an interval.

\Bayesian view/ • probability is subjective (it measures uncertainty or degree of belief)
• y is again "generated" at random from $P(y|\theta)$ ← expresses our uncertainty about y is we know θ. ($P(y|\theta)$ follows) (likelihood)
• we are also uncertain about θ. So we can specify $P(\theta)$
• we can form the joint $P(\underline{y}, \theta)$

Bayes thm $P(\theta|\underline{y}) = \frac{P(\theta)P(\underline{y}|\theta)}{P(\underline{y})}$

$P(\theta)$ ← prior distribution ($\pi(\theta)$)
$P(\theta|\underline{y})$ ← posterior distn ($\pi(\theta|\underline{y})$)
$P(\underline{y}|\theta)$ ← likelihood
$P(\underline{y})$ ← marginal likelihood (prior predictive).

\Ex/ ~1% of women [40,50] have breast cancer.
~ "90% of time a mammogram is "correct"".
r.v. is θ, where $\theta=1$ if a woman has breast cancer & $\theta=0$ otherwise
$\pi(\theta=1) = 0.01$ $\therefore \pi(\theta=0) = 0.99$.
Data y the mammogram test $y=1$

$$\pi(\theta=1 \mid y=1) \overset{\text{Bayes thm}}{=} \frac{\pi(\theta=1)\,p(y=1 \mid \theta=1)}{p(y=1)}$$

$\theta=1 \Rightarrow y=1$ "correct"

$\theta=0 \Rightarrow y=0$ "correct"

$p(y=1 \mid \theta=1) = p(y=0 \mid \theta=0) = 0.9$

$p(y=1) = p(y=1 \mid \theta=1)\,p(\theta=1) + p(y=1 \mid \theta=0)\,p(\theta=0)$ (LOTP $\theta=1,0$ form a partition.)

$$\pi(\theta=1 \mid y=1) = \frac{0.01 \times 0.9}{\left(\frac{1}{10} \times \frac{1}{100} + \frac{1}{10} \times \frac{99}{100}\right)} = \frac{9}{9+99} = \frac{1}{12}$$

$\pi(\theta \mid \underline{y}) \propto \pi(\theta)\,p(\underline{y} \mid \theta)$

$$p(\underline{y}) = \int_{-\infty}^{\infty} p(\underline{y} \mid \theta)\,\pi(\theta)\,d\theta \qquad p(\underline{y}) = \int_{-\infty}^{\infty} p(\underline{y}, \theta)\,d\theta$$

$p(\underline{y})$ is const in θ

we know that densities (pdfs) integrate to 1.

$y_i \mid \theta$ iid $\text{Ber}(\theta)$ $0 \le \theta \le 1$ let $\theta \sim \text{unif}(0,1)$

suppose $i=1,\dots,n$, find a posterior distn & write down its density.

one mark

By Bayes Theorem [1]

$\pi(\theta \mid \underline{y}) \propto \pi(\theta)\,p(\underline{y} \mid \theta)$ [1]

$$\propto \pi(\theta) \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i}$$

$$= 1 \times \theta^{\sum y_i}(1-\theta)^{n-\sum y_i}$$

$$S = \sum_{i=1}^{n} y_i \quad [= n\bar{y}]$$

$$\pi(\theta \mid \underline{y}) \propto \theta^{n\bar{y}}(1-\theta)^{n(1-\bar{y})}$$

$x \sim \text{Beta}(a,b)$ pdf $\dfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$

matching powers with a Beta Distn

$[a-1 = n\bar{y}]$ $\therefore$ $a = n\bar{y}+1$, $b = 1+n(1-\bar{y})$

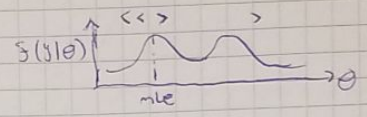
$\theta \sim \text{Beta}(n\bar{y}+1,\ 1+n(1-\bar{y}))$

density $\pi(\theta \mid \underline{y}) = \dfrac{\Gamma(n+2)}{\Gamma(n\bar{y}+1)\Gamma(1+n(1-\bar{y}))}\,\theta^{n\bar{y}}(1-\theta)^{n(1-\bar{y})}$

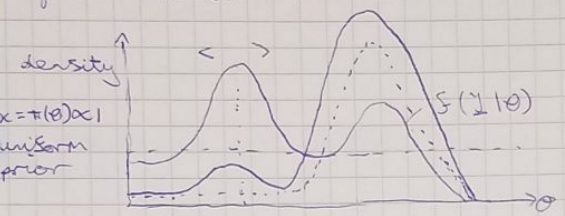
$y|\theta \sim f(y;\theta)$

$y_1, \ldots, y_n$ iid $f(y_i;\theta)$

likelihood $P(\underline{y}|\theta)$ $(f(\underline{y}|\theta))$



Bayes $\pi(\theta|\underline{y}) \propto \pi(\theta) f(\underline{y}|\theta)$



Sorry $x = \pi(\theta) \propto 1$ uniform prior

posterior inference (Bayesian inference) ≡ likelihood inference

M event Mildred is right

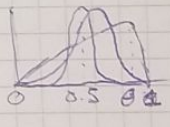
P is Paul is right

$M_i \sim Ber(\theta_M)$ $P_i \sim Ber(\theta_P)$

is $\theta_M > \theta_P$? $\theta_M = \theta_P$?

likelihood $P(\underline{M}|\theta_M) = \theta_M^9(1-\theta_M)^1$ $\hat{\theta}_M = 9/10$

$P(\underline{P}|\theta_P) = \theta_P^3(1-\theta_P)^1$ $\hat{\theta}_P = 3/10$



$P(\theta_M)$, $P(\theta_P)$

Paul is an octopus and can't see Stays and doesn't understand the game attached to any soccer ball games & ∴ is less likely to not be superstitious

$P(\theta_P > 0.5) \approx 0$ $P(\theta_M > 0.5) >> P(\theta_P > 0.5)$

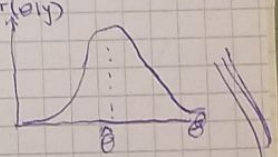
MAP (here) < MLE

Suppose $P(y|\theta)$, $\pi(\theta)$, Bayes gives $\pi(\theta|\underline{y})$

F1: compute an estimator, $\hat{\theta}$, for θ. Study bias etc. $E[\hat{\theta}] = \theta$ $(MLE = \arg\max_\theta P(\underline{y}|\theta))$

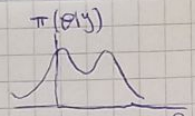
B1: $E[\theta|\underline{y}] = \int_{-\infty}^{\infty} \theta \pi(\theta|\underline{y}) d\theta$

Maximum A posteriori estimate: (MAP) $\hat{\theta} = \arg\max_\theta \pi(\theta|\underline{y})$



F2: An α% confidence interval for true θ: is an infinite sequence of repeated experiments with data $y_1, y_2, \ldots$

$P(\theta \in [a_i, b_i])$ is α%.



B2: $\pi(\theta|y)$

for any a, b $\pi(a<\theta<b|y) = \int_a^b \pi(\theta|y)d\theta$

if we choose α s.t. that $\pi(a<\theta<b|y) = \alpha$

then $[a,b]$ is a 100α% credible interval for y

F3: consider a hypothesis test involving true θ. we might test $\theta \geq T$ & $\theta < T$ for some T. Data rejects one hypothesis in favour of another

B3: $\pi(\theta|y)$ $\pi(\theta \geq T|y)$, $\pi(\theta < T|y)$

$\pi(\theta \geq T|y) = \int_T^\infty \pi(\theta|y)d\theta$

H_0 (Null) 'no effect' e.g. $X \sim Ber(\theta)$ $\theta = \frac{1}{2}$

$y|x, \theta \sim N(\theta_1 + \theta_2 x, \sigma^2)$ $\theta_2 = 0$

H_1 (Alternate) $\theta \sim Beta(a,b)$ $\theta_2 \sim N(\mu, \sigma^2)$

Bayes Factor BF $B = \frac{P(y|H_0)}{P(y|H_1)} = \frac{\int P(y|\theta, H_0)\pi(\theta, H_0)d\theta}{\int P(y|\theta, H_1)\pi(\theta, H_1)d\theta} =$

$$\frac{\int \binom{n}{y}(\frac{1}{2})^y(\frac{1}{2})^{n-y}d\theta}{\int \binom{n}{y}\theta^y(1-\theta)^{n-y}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}d\theta} \quad :(\tfrac{1}{2})^n$$

$\underline{y} = (y_1, \ldots, y_n)$ iid $P(y_i|\theta)$

Prediction: estimate $\hat{\theta} = \theta$ plugin $\hat{\theta}$ to $P(y|\theta)$

predicting $\hat{y}$ (y_{new}) involves assuming $P(\hat{y}|y) \approx P(\hat{y}|\hat{\theta})$

$P(\hat{y}, \theta|y) = P(\hat{y}|\theta, y)P(\theta|y) =$

$P(\hat{y}|\theta)P(\theta|y)$

$P(\hat{y}|y) = \int P(\hat{y}, \theta|y)d\theta$ marginalisation $= \int_{-\infty}^{\infty} P(\hat{y}|\theta)\pi(\theta|y)d\theta$

so desired. $P(\hat{y}|y)$ is called a posterior predictive distribution

$y_1, \ldots, y_n$ coin tosses $y_i|\theta \sim Ber(\theta)$

prior for θ, $\pi(\theta) = 1$ $\theta \in [0,1]$

$\sum y_i = n\bar{y} := s$

$\theta | y \sim \text{Beta}(s+1, n-s+1)$

- $\pi(\theta | y) = \frac{\Gamma(n+2)}{\Gamma(s+1)\Gamma(n-s+1)} \theta^s (1-\theta)^{n-s}$

posterior predictive

$$p(\tilde{y}|y) = \int_0^1 p(\tilde{y}|\theta)\pi(\theta|y)\,d\theta = \frac{\Gamma(n+2)}{\Gamma(s+1)\Gamma(n-s+1)} \int_0^1 \theta^{\tilde{y}}(1-\theta)^{1-\tilde{y}} \theta^s (1-\theta)^{n-s}\,d\theta =$$

$$\frac{\Gamma(n+2)}{\Gamma(s+1)\Gamma(n-s+1)} \int_0^1 \underbrace{\theta^{s+\tilde{y}}(1-\theta)^{n-s-\tilde{y}+1}}_{I} d\theta$$

is I a pdf then $\int I\,d\theta = 1$

$\therefore a+b = n+3$

$s+\tilde{y} = a-1 \therefore a = s+\tilde{y}+1$

$b-1 = n-s-\tilde{y}+1 \therefore b = n-s-\tilde{y}+2$

$$= \frac{\Gamma(n+2)}{\Gamma(s+1)\Gamma(n-s+1)} \int_0^1 \underbrace{\frac{\Gamma(n+3)}{\Gamma(s+\tilde{y}+1)\Gamma(n-s-\tilde{y}+2)} \theta^{s+\tilde{y}}(1-\theta)^{n-s-\tilde{y}+1}\,d\theta}_{1}$$

$$= \frac{\Gamma(n+2)\Gamma(s+\tilde{y}+1)\Gamma(n-s-\tilde{y}+2)}{\Gamma(s+1)\Gamma(n-s+1)\Gamma(n+3)}$$

$$p(\tilde{y}|y) = \frac{\Gamma(n+2)\Gamma(s+\tilde{y}+1)\Gamma(n-s-\tilde{y}+2)}{\Gamma(s+1)\Gamma(n-s+1)\Gamma(n+3)}$$

$\Gamma(t+1) = t\Gamma(t) \quad \forall t, \quad \Gamma(1) = 1$

$\tilde{y}, n, s$ are all integers

$$p(\tilde{y}|y) = \frac{1}{n+2} \times \frac{(s+\tilde{y})!}{s!} \times \frac{(n-s+1-\tilde{y})!}{(n-s)!}$$

Laplace's law of succession

Yes/ potentially infinite "repeatable" trials. probab is limiting relative frequency of an event

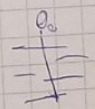
$H \sim \text{Ber}(\theta)$ toss n times. y number of heads

$\theta(y) = y/n$

$p(\hat{\theta}(y) > \frac{1}{2})$ X (question doesn't make sense)

$p(\hat{\theta}(y) - \varepsilon \le \theta_0 \le \hat{\theta}(y) + \varepsilon)$ X (question doesn't make sense since there's no randomness anywhere

- $\hat{\theta}(Y)$, $p(\hat{\theta}(Y) - \varepsilon \le \theta_0 \le \hat{\theta}(Y) + \varepsilon)$ ✓



Yes/ A random quantity is a quantity whose value is unknown to you

Week 1

1a/ By Bayes thm: ① Posterior: $\pi(\theta|y) \propto \pi(\theta) p(y|\theta)$ ④

$\propto \theta^{a-1}(1-\theta)^{b-1} \prod_{i=1}^{n} \theta(1-\theta)^{y_i-1}$ ⑤ $\propto \theta^{a+n-1}(1-\theta)^{b+n\bar{y}-n-1}$

$\left\{ \prod_{i=1}^{n}(1-\theta)^{y_i-1} = (1-\theta)^{\sum_{i=1}^{n}(y_i-1)} \right\}$ ∴

$\theta|y \sim Beta(a+n, b+n\bar{y}-n)$ ②

∵ $\pi(\theta|y)$ is proportional to a Beta density ③

$$\pi(\theta|y) = \frac{\Gamma(a+b+n\bar{y})}{\Gamma(a+n)\Gamma(b+n\bar{y}-n)} \theta^{a+n-1}(1-\theta)^{b+n\bar{y}-n-1} \qquad \theta \in [0,1]$$

1b/ $E[\theta|y] = \int_0^1 \theta \pi(\theta|y)\, d\theta = \frac{a}{a+b} = \left(\frac{a+n}{a+b+n\bar{y}}\right)$

Three prisoners / A, B, C A thinks all equally likely to go free ∴ $P(A) = P(B) = P(C) = \frac{1}{3}$

No information about A is you tell Me 2 names, Some one to be executed

Jailer says truth - B will die

Now A thinks its Me or C ∴ $P(A) = \frac{1}{2}$

b be Event jailer says B will happen General $P(b|A) = p$

x ∴ $P(A|b)$ $\left\{ x = \frac{p}{p+1} \right\}$ $x = \frac{P(b|A)P(A)}{\sum_{i=A,B,C} P(b|i)P(i)} =$

$\frac{p}{p+1}$

If A arguing $P(A|b) = P(A)$ ∴ $x = \frac{1}{3}$ ∴ $\frac{1}{3} = \frac{p}{p+1}$ ∴ $\frac{1}{3}p + \frac{1}{3} = p$ ∴

$\frac{2}{3}p = \frac{1}{3}$ ∴ $2p = 1$ ∴ $p = \frac{1}{2}$ ∴ $P = \frac{1}{2}$

If A thinks $P(A|b) \neq P(A)$ ∴ $x = \frac{1}{2}$ ∴ $\frac{p}{p+1} = \frac{1}{2}$ ∴ $\frac{1}{2}p + \frac{1}{2} = p$ ∴

$\frac{1}{2}p = \frac{1}{2}$ ∴ $P = 1$ but this goes against common sense

SHELF ④ Elicit individual judgements

2) group discussion 3) group consensus

Jeffreys prior Beta $(\frac{1}{2}, \frac{1}{2})$ Beta: $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}$

Beta(1,1) = Unif(0,1) Beta(a,b)

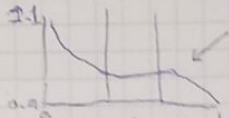
Median for quantity of interest (QOI) ×

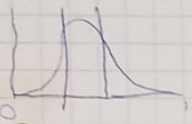
Median MSt quantity is equally likely to be lower or higher

0.5 → θ is prob of tails

Quartiles for a quantity of interest (QOI) so its the Median of 2 Median and 2 bounds

0.25 & 0.75





Beta(10,10) optics.eee.nottingham.ac.uk

is Median is 0.505 Sobris on Coin ∴ [S(7), S(3)] = [α, β]

Beta(α+y, b+n-y) n=tosses y=heads

Code on the week 2 MLE is 5

with 2 tails out of 5 you get 2 pseudo prior

its becoming dissicult for Beta to have any influence at all because of your prior knowledge

$P(y|\theta)$ likelihood choose $\pi(\theta)$ be anything $\pi(\theta|y) \propto \pi(\theta)P(y|\theta)$ (unnormalised posterior density)

(for density $\int_{-\infty}^{\infty} P(y|\theta)\pi(\theta)d\theta$)

\Def 1.3/ is F is a class of sampling distris with params θ, $P(y|\theta)$ & P is a class of prior distris for θ then is $\forall \pi(\theta) \in P$, we have $\pi(\theta|y) \in P$ then P class P is conjugate for F

\Ex/ y is 1 is a coin toss is heads Suppose n coin tosses

$y|\theta \sim Bin(n,\theta)$ $\theta \sim Beta(a,b)$ ∴

By Bayes thm $\pi(\theta|y) \propto \pi(\theta)P(y|\theta)$

$\propto \theta^{a-1}(1-\theta)^{b-1}\theta^{y}(1-\theta)^{n-y} \propto \theta^{a+y-1}(1-\theta)^{b+n-y-1}$

$\pi(\theta|y)$ is proportional to a Beta density ①

$\theta|y \sim \text{Beta}(a+y, b+n-y)$ ①

∴ Beta dist'n is conjugate for Binomial models

$E[\theta|y] = \frac{a+y}{a+y+b+n-y} = \frac{a+y}{a+b+n}$ as $n \to \infty$ $E[\theta|y] \to \frac{y}{n} = \bar{y}$

$\hat{\theta} = \text{argmax}\ \pi(\theta|y) = \frac{a+y-1}{a+b+n-2}$ tends to $\frac{y}{n} = \bar{y}$ as n tends to ∞

measuring an object of length θ, instrument with known error var σ^2 ∴ $y_i|\theta, \sigma^2 \sim N(\theta, \sigma^2)$

suppose have prior for $\theta \sim N(\theta_0, \sigma_0^2)$

given n measurements $y_1, \dots, y_n = y$, find $\pi(\theta|y)$

By Bayes Theorem $\pi(\theta|y) \propto \pi(\theta)\, p(y|\theta)$

$$\propto \frac{1}{\sqrt{2\pi}\,\sigma_0} \exp\left\{-\frac{1}{2\sigma_0^2}(\theta-\theta_0)^2\right\} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y_i-\theta)^2\right\}$$

$$\propto \exp\left\{-\frac{1}{2\sigma_0^2}(\theta-\theta_0)^2\right\} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma_0^2}(\theta-\theta_0)^2 + \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right)\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma_0^2}(\theta^2-2\theta_0\theta+\theta_0^2) + \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i^2-2\theta y_i+\theta^2)\right)\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma_0^2}(\theta^2-2\theta_0\theta) + \frac{1}{\sigma^2}\sum_{i=1}^{n}(\theta^2-2y_i\theta)\right)\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma_0^2}(\theta^2-2\theta_0\theta) + \frac{1}{\sigma^2}(n\theta^2-2n\bar{y}\theta)\right)\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\left(\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)\theta^2 - 2\left(\frac{\theta_0}{\sigma_0^2}+\frac{n\bar{y}}{\sigma^2}\right)\theta\right)\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)\left(\theta^2 - 2\left(\frac{\left(\frac{\theta_0}{\sigma_0^2}+\frac{n\bar{y}}{\sigma^2}\right)}{\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)}\right)\theta\right)\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)\left(\theta - \left(\frac{\left(\frac{\theta_0}{\sigma_0^2}+\frac{n\bar{y}}{\sigma^2}\right)}{\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)}\right)\right)^2 + \left(\frac{\left(\frac{\theta_0}{\sigma_0^2}+\frac{n\bar{y}}{\sigma^2}\right)}{\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)}\right)^2\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)\left(\theta - \left(\frac{\left(\frac{\theta_0}{\sigma_0^2}+\frac{n\bar{y}}{\sigma^2}\right)}{\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)}\right)\right)^2\right\} \quad \therefore$$

$\theta|y \sim N(\theta_n, \sigma_n^2)$ ∴ $X \sim N(M, V^2)$ is $S(x) \propto \exp\left(-\frac{1}{2V^2}(x-M)^2\right)$ ∴

$$\theta_n = \frac{\left(\frac{\theta_0}{\sigma_0^2} + \frac{n\bar{y}}{\sigma^2}\right)}{\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)} \quad \therefore \quad \sigma_n^2 = \frac{1}{\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)}$$

$\therefore \pi(\theta|y) \propto$ to a Normal density

$\therefore$ as $n \to \infty$: $\theta_n \to \bar{y}$

two data sets y_1, y_2

consider a prior, $\pi(\theta)$, that's uninformative

$\pi(\theta|y_1, y_2) = \pi(\theta|y_1)p(y_2|\theta, y_1) = \pi(\theta|y_1)p(y_2|\theta)$

$= \pi(\theta)p(y_1|\theta)p(y_2|\theta)$

'Objective' Bayes (O-Bayes)

$p(y|\theta) \xrightarrow[\text{choice}]{\text{automatic}} \pi(\theta)$

- Not to inform a posterior (uninformative)
- Have 'good' frequentist properties

$p(\theta \in [a,b]) = 0.95$

- Be transformation invariant

one approach is a 'uninformative prior' $\pi(\theta) \propto 1$

$\pi(\theta|y) \propto p(y|\theta)$ (same inference as for classical, different interpretations)

Def: / A prior (or posterior) is proper if it is a valid prob distn (eg its pdf integrates to 1) otherwise its improper

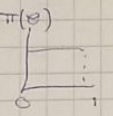
Ex/ $y|\theta \sim N(\theta, \sigma^2)$ $\pi(\theta) \propto 1$ $\theta \in (-\infty, \infty)$ $\therefore$

$\int_{-\infty}^{\infty} 1 \, d\theta = \infty$ $\therefore \pi$ is improper

but $\pi(\theta|y) \propto \exp\left\{-\frac{1}{2\sigma^2}(y-\theta)^2\right\}$ $N(y, \sigma^2)$

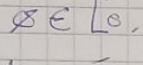
$\therefore$ posterior is proper

$y|\theta \sim Ber(\theta)$



uninformative prior $\pi(\theta) = 1 \quad \theta \in [0,1]$

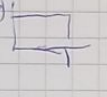
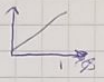
consider a reparameterisation $\phi = \sqrt{\theta}$ ∴

$\pi(\phi) = 1 \quad \phi \in [0,1]$ ∴  this approach is incoherent since $\pi(\theta) = 1 \Rightarrow \pi(\phi)$ }

i.e. $\phi = g(\theta) \quad \pi_\phi(\theta) = \left| \det\left(\frac{\partial g^{-1}(\phi)}{\partial \phi} \right) \right| \pi_\theta(g^{-1}(\phi))$

$g(\theta) = \sqrt{\theta}$ ∴ $g^{-1}(\phi) = \phi^2$ { $g(\theta) = \sqrt{\theta}$ ∴ i.e. $\phi = \sqrt{\theta}$ ∴ $\theta = \phi^2$ ∴ $g^{-1}(\phi) = \phi^2$ }

∴ $g^{-1}(\theta) = \phi^2 = \theta$ ∴ $\frac{\partial g^{-1}(\phi)}{\partial \phi} = \frac{\partial}{\partial \phi}(\phi^2) = 2\phi$ ∴

$\pi(\phi) = 2\phi$  ⟹ 

\Def/ for prob model $p(y|\theta)$ with scalar θ & expected info $J(\theta) = E\left[-\frac{\partial^2 l}{\partial \theta^2} \right]$, Z Jeffreys prior is Z distri with density proportional to $J(\theta)^{1/2}$ $\pi(\theta) \propto J(\theta)^{1/2}$

\Ex/ $y|\theta \sim Bin(n, \theta)$ ∴ $p(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$ ∴

$l(\theta) = \log\binom{n}{y} + y \log\theta + (n-y)\log(1-\theta)$ ∴

$l'(\theta) = \frac{y}{\theta} - \frac{n-y}{1-\theta}$ ∴ $l''(\theta) = -\frac{y}{\theta^2} - \frac{n-y}{(1-\theta)^2}$

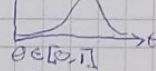
∴ $J(\theta) = E\left[\frac{y}{\theta^2} + \frac{n-y}{(1-\theta)^2} \right] = \frac{E[y]}{\theta^2} + \frac{n - E[y]}{(1-\theta)^2} =$

$\frac{n\theta}{\theta^2} + \frac{n(1-\theta)}{(1-\theta)^2} = n\left(\frac{1}{\theta} + \frac{1}{(1-\theta)} \right) = n\left(\frac{1-\theta+\theta}{\theta(1-\theta)} \right) = \frac{n}{\theta(1-\theta)}$ ∴

$\pi(\theta) \propto J(\theta)^{1/2} \propto \theta^{-1/2}(1-\theta)^{-1/2}$ ∴ $\theta \sim Beta\left(\frac{1}{2}, \frac{1}{2}\right)$

$p(y|\theta) \quad \pi(\theta) \rightarrow \pi(\theta|y) \propto \pi(\theta) p(y|\theta)$

$\pi(\theta)$ represent my or a scientists beliefs. prior elicitation

$\pi(\theta)$ 

- known psychological pitfalls
- stick to observables

θ is abstract so try to elicit y.

- Be transparent
- Do sensitivity analysis

Match uncertainty elicitation tool

die from lung cancer is more likely than [illegible] accident

book - 1.1 m person A most likely to be a librarian

\Heuristic/ 1) Availability

2) Anchoring - is 40ok in interval because he kept saying it?

3) Most people put librarian but there are more teachers & farmers in the world ∴ more likely to be one of those

∴ Representativeness: conditional probs are judged by symmetry (how similar A is to B) but $P(A|B) \neq P(B|A)$

eg prob of having this description given they're a librarian is quite high but prob they're a librarian given this description is quite low

eg $P(A|\text{Teacher})$ low but $P(\text{Teacher}|A)$ high

$\pi(\theta)$ $y|\theta \sim f(y;\theta)$ or $y|\underline{\theta} \sim f(y;\underline{\theta})$ $\pi(\underline{\theta})$

∴ try to infer $\pi(\theta) \to \pi(\underline{\theta})$

hard to think about $\pi(\theta)$ $P(y|\theta)$

$\pi(\theta)$ induces beliefs about y through prior predictive distrib $P(y) = \int_{-\infty}^{\infty} P(y,\theta)\,d\theta =$

$\int_{-\infty}^{\infty} P(y|\theta)\pi(\theta)\,d\theta$

give a graph you think  ∴ $P(y|\theta)$ ∴ $\pi(\theta)$ ∴ leads to:

Ω, $E \subseteq \Omega \leftarrow P$

R.V.

\Def/ a random quantity (r.q.) X is a number whose true val is unknown (to you)

\Def/ an event admits only two possible vals, True or False

we use 2 convention, if (A is a random event), that A=1 is A is True, A=0 is A is False

$\wedge$ 'and' , $\vee$ 'or' , $\sim$ ('Not' eg $\sim A$, $\tilde{A}$)

$A \vee B$ (TRUE is A or B happen)

$A \wedge B$ (true is A and B happen)

$\sim(A \vee B) = \tilde{A} \wedge \tilde{B} = \tilde{A}\tilde{B}$

\Def/ let x, y be real numbers, then $x \wedge y = \min(x,y)$ $x \vee y = \max(x,y)$

$\hat{x} = 1 - x$

this establishes a duality between r.q's and events

$A \vee B = \sim(\tilde{A} \wedge \tilde{B}) = \sim(\tilde{A}\tilde{B}) = 1 - ((1-A)(1-B)) = 1 - (1 - A - B + AB) =$

$A + B - AB$

$A \vee (B \vee C) = A \vee \sim(\tilde{B} \wedge \tilde{C}) = \sim(\tilde{A} \wedge (\tilde{B} \wedge \tilde{C})) = 1 - ((1-A)(1-B)(1-C)) =$

$1 - ((1-A)(1-B-C+BC)) = 1 - (1 - B - C + BC - A + AB + AC - ABC) =$

$A + B + C - AB - BC - AC + ABC$

these questions will be in exam with probably from 5 than ... change 2 'or's into 'and's & nots

\Def 2.4/ given a r.q., X your Expectation for X, $E[X]$ is that val, $\bar{x}$, you would choose on 2 understanding that, you are committed (having chosen $\bar{x}$) to accepting any bet with gain $C(X - \bar{x})$, where C is both arbit and chosen by an "opponent"

Assumption: Coherence: we assume that you dont wish to lay down bets that with certainty result in a loss for you

\Lemma 2.1/ Expectation is a linear func: $E\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i E[X_i]$

for a r.q. X_i, Constas a_i

\proof/ $E[aX] = aE[X]$ $E[X+Y] = E[X] + E[Y]$

(1) let $\bar{x}$ be $E[X]$, $\bar{a} = E[aX]$ we must accept all bets with gain $G = C_1(aX - \bar{a}) + C_2(X - \bar{x})$ ...

• Be transparent • 1/0 sensitivity ...

set $C_2 = -aC_1$ $C_T = C_1 aX - C_1 \bar{a} - aC_1 X + aC_1 \bar{x} =$
$C_1(a\bar{x} - \bar{a})$

by coherence (to avoid sure loss) $a\bar{x} = \bar{a}$

(2) $E[X] = \bar{x}$ $E[Y] = \bar{y}$ $Z = X + Y$ $\bar{z} = E[Z]$ $\bar{z} = \bar{x} + \bar{y}$

$C_T = C_1(X+Y-\bar{z}) + C_2(X-\bar{x}) + C_3(Y-\bar{y}) =$
$(C_1 + C_2)X + (C_1 + C_3)Y - C_1\bar{z} - C_2\bar{x} - C_3\bar{y}$

$C_1 = -C_2 = -C_3$

$C_T = -C_1\bar{z} + C_1\bar{x} + C_1\bar{y} = C_1(\bar{x} + \bar{y} - \bar{z})$

to be coherent, $\bar{x} + \bar{y} = \bar{z}$ proves (2) $\therefore$
2 general proofs follows by induction □

\Def 2.5/ given r.q. X, your expectation for X, $E[X]$, is that val $\bar{x}$, you would choose if you must suffer penalty $L = \left(\frac{X - \bar{x}}{k}\right)^2$
where X is revealed. k is an arbit const

assumption: Coherence, you do not have a preference for given penalty if you have 2 option fs another that is certainly smaller

\Thm 2.1 Def 2.4 & 2.5 are equivalent/ \proof/ let $\bar{x}$ be $E[X]$
under 2.4 let $\bar{\bar{x}}$ be $E[X]$ under 2.5
under 2.5 gain $-(X - \bar{\bar{x}})^2$ is preferred to $-(X-x)^2$ for $x \neq \bar{\bar{x}}$
$\therefore$ Gain $G_T = (X-x)^2 - (X - \bar{\bar{x}})^2$ is preferred to 0 $\forall x \neq \bar{\bar{x}}$

$a = \bar{\bar{x}}$, $b = x$, $c = \frac{1}{2}(a+b)$

$G_T = (X-b)^2 - (X-a)^2 = (X^2 - 2bX + b^2) - (X^2 - 2aX + a^2) = 2(a-b)X - (a^2 - b^2) =$
$2(a-b)(X-c)$ $\therefore E[G_T] > 0$
$\therefore$ under 2.4 $E[G_T] = 2(a-b)(\bar{x} - c) > 0$
$E[G_T]$ is positive iff $a > b$ & $\bar{x} > c$ or $b > a$ & $c > \bar{x}$

$\bar{x}$ must be closer to $\bar{\bar{x}}$ (a) than arbit x (b) $(\forall x)$
letting $x \to \bar{\bar{x}}$ $\bar{x} \to a$ $\bar{x} = \bar{\bar{x}}$ is 2 optimal choice □

\thm 2.2 (Boundedness of expectation)\
$\inf X \le E[X] \le \sup X$

\proof\ first suppose $E[X] < \inf X$ ∴

$-\inf X < -E[X] \iff X - \inf X < X - E[X] \iff$

$(X-\inf X)^2 < (X-E(X))^2$ this implies we could reduce our penalty (L) by changing $E[X]$ to $\inf X$ (indep of X) ∴ by coherence $E[X] \ge \inf X$

suppose instead $E[X] > \sup X$ then $X - \sup X > X - E[X] \iff$

$(X-\sup X)^2 < (X-E(X))^2$ hence you can reduce your penalty independently of X, by choosing $E[X] = \sup X$ instead (∴ $\perp X$) ∴

so by coherence $E[X] \le \sup X$ □

$E[X] = \int x S(x)\,dx \quad \sum_x x\, p(X=x)$

let A be an event

\def 2.6 (probab)\ your probab for A ($P(A)$) is $E[A]$

\thm 2.3\ $0 \le P(A) \le 1$

\proof\ this follows immediately from boundedness of expectation □

\thm 2.4\ let $A_1, \dots, A_n$ be incompatible (mutually exclusive) events

let A be a event that one of A happens $P(A) = P(A_1) + \dots + P(A_n)$

\proof\ $A = A_1 \vee A_2 \vee \dots \vee A_n = A_1 + A_1 + \dots + A_n$ (incompatible)

$P(A) = P(A_1 + \dots + A_n) = P(A_1) + P(A_2) + \dots + P(A_n)$ by linearity of Expectation

\thm 2.5\ for finite partition $A_1, \dots, A_n$ then event $A = A_1 \vee \dots \vee A_n$ has probab 1

\proof\ $P(A) = P(A_1) + \dots + P(A_n)$ {by thm 2.4} $= 1$ ∴

$A = A_1 + \dots + A_n = 1$ So $P(A) = 1$ in order to be coherent

{give this proof in exam}

2 partition with $n=2$: $A, \tilde{A}$ ∴ (thm 2.5) implies $P(\tilde{A}) = 1 - P(A)$

$P(A \vee B) = P(\sim(\tilde{A} \wedge \tilde{B})) = P(1 - ((1-A)(1-B)) = P(A+B-AB) =$

$P(A) + P(B) - P(AB)$ by linearity of expectation ① ← 1 mark in exam

or $\vee$ — 2 partition

$E[X]$, $E[X|B]$ $\quad P(A|B) = \frac{P(A \wedge B)}{P(B)}$

\Def 2.7 (Conditional Expectation) / given r.v. X & possible event H

Suppose you are subject to a penalty $L = H\left(\frac{X - \bar{x}}{k}\right)^2$

where X is revealed. your conditional expectation, $E[X|H]$ is $\exists$ val $\bar{x}$ that you would choose for this

\Def 2.5/ Conditional probability

is A is an event with H, $\exists$ as above: $P(A|H) = E[A|H]$

\thm 2.6 (compound probabs)/ for events H, A, $\exists$ probabs $P(HA)$ $P(H)$ & $P(A|H)$ where is $P(HA) = P(A|H)P(H)$

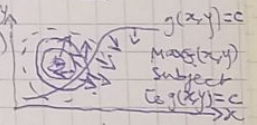
\proof/ take $k=1$ $\exists$ L we since having chosen $P(A|H) = x$, $P(H) = y$ & $P(HA) = z$ is $L = H(A-x)^2 + (H-y)^2 - (HA-z)^2$

$HA = 1$ $(H = A = 1)$, $H = 1$, $A = 0$ $(HA = 0)$, $H = 0$, $HA = 0$

3 cases: $HA: L = u(x,y,z) = (1-x)^2 + (1-y)^2 + (1-z)^2$

$H\tilde{A}: L = v(x,y,z) = x^2 + (1-y)^2 + z^2$

$\tilde{H}: L = w(x,y,z) = y^2 + z^2$ have to show $z = xy$ or there is a choice of x, y, z that simultaneously reduces u, v, w



this only happens gradients ∇x, ∇y, ∇z are coplanar

$\nabla x, \nabla y, \nabla w$ are coplanar $\iff$ $J(x,y,z) = \begin{vmatrix} \partial x/\partial x & \partial x/\partial y & \partial x/\partial z \\ \partial y/\partial x & \partial y/\partial y & \partial y/\partial z \\ \partial z/\partial x & \partial z/\partial y & \partial z/\partial z \end{vmatrix} = 0$

$$= \begin{vmatrix} -2(1-x) & -2(1-y) & -2(1-z) \\ 2x & -2(1-y) & 2z \\ 0 & 2y & 2z \end{vmatrix} = 8z(1-x) + 8xz(1-y) - 8xy(1-z) =$$

$8(z - xy) = 0 \iff z = xy$ $\square$

$\exists$ thm extends to expectation $E[XH] = E[X|H]P(H)$ $\{X$ an r.v.$\}$

\Bayes thm/ $P(A|H) = \frac{P(A)P(H|A)}{P(H)}$ follows immediately

$Y_1, \ldots, Y_n$ iid $\quad Y_i|\theta \sim Ber(\theta) \quad \theta \sim Beta(a,b)$

\Def 2.9/ a sequence $Y_1, Y_2, \ldots$ of r.v.'s is said to be exchangeable if $\exists$ joint probab distri of each sub-collection of n quantities $(Y_{i_1}, \ldots Y_{i_n})$ is $\exists$ same

Thm 2.7 (Representation thm) / let $Y_1, Y_2, \ldots$ an infinite exchangeable sequence of r.v.'s with vals in $\mathcal{Y}$. Then $\exists$ a measure F (prior) on a set of prob measures $Q(\mathcal{Y})$ (on $\mathcal{Y}$) st for every n & subsets $E_1, \ldots, E_n$ of $\mathcal{Y}$.

$P(Y_1 \in E_1, Y_2 \in E_2, \ldots, Y_n \in E_n) = \int Q(E_1) \cdots Q(E_n)\, dF$

$\int dF(Q) = \int \pi(Q)\, dQ$

$P(y) = \int p(y|\theta)\, \pi(\theta)\, d\theta$ ← $\pi(\theta)$ exists

$y_1, \ldots, y_n$ indep observations from some $S(y)$, true but unknown data generating process. Suppose we model y with $p(y|\theta)$ & $\pi(\theta)$. Consider a Kullback-Leibler divergence (KL), $KL(\theta)$ measures "distance" betw 2 prob dists is

$KL(\theta) = \| S(y) - p(y|\theta) \|_{KL} = E\left[\log\left(\frac{S(y)}{p(y|\theta)}\right)\right] = \int \log\left(\frac{S(y)}{p(y|\theta)}\right) S(y)\, dy$

if likelihood is continuous in θ, θ_0 not on a boundary, then as $n \to \infty$: $\theta|y \to N(\theta_0, (nJ(\theta_0))^{-1})$ where $\theta_0 = \arg\min_{\theta} KL(\theta)$

$\theta_0 = \arg\min KL(\theta)$ $\quad \| S(y) - p(y|\theta) \|_{KL}$ $\quad \hat{\theta}$ is a MAP (posterior mode)

$\hat{\theta} \to \theta_0$ as $n \to \infty$ under assumed conditions is well known result (consistency of MAP). $\theta|y$ as $n \to \infty$: $\theta|y \to N(\theta_0, (nJ(\theta_0))^{-1})$

let $L(\theta) = \log \pi(\theta|y)$, Taylor expansion around $\hat{\theta}$;

$L(\theta) = L(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^2 \frac{d^2}{d\theta^2} L(\theta)\Big|_{\theta=\hat{\theta}} + \frac{1}{6}(\theta - \hat{\theta})^3 \frac{d^3}{d\theta^3} L(\theta)\Big|_{\hat{\theta}} + \ldots$

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

2 coeff is ~ const + sum of n iid r.v.'s with negative mean if $S(y) = p(y|\theta_0)$ each of 2 n terms has mean $-J(\theta_0)$

else 2 mean is 2 negative of 2 2nd derivative of 2 1st at θ_0,

by law of large nums 2 sum $\to -\infty$ as $n \to \infty$ ∴ as $n \to \infty$ $|\hat{\theta} - \theta_0| \to 0$

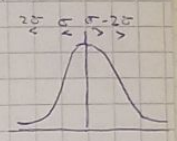
away from $\hat{\theta}$, $L(\theta) \to -\infty$ $\pi(\theta|y) \to 0$ ∴ ~~$\pi(\theta) \propto \exp$~~

$\pi(\theta|y) \propto \exp\left\{-\frac{1}{2} n J(\theta_0)(\theta - \theta_0)^2\right\}$ ∴ $\theta|y \sim N(\theta_0, (nJ(\theta_0))^{-1})$

Babys weight $\tilde{y}$ $y|\mu,\sigma^2 \sim N(\mu,\sigma^2)$

$\pi(\mu,\sigma^2)$ $\mu|\sigma^2 \sim N(\mu_0, \sigma^2/\tau)$ $\sigma^2 \sim IG(\alpha,\beta)$

$\mu \pm 2\sigma$



$$p(\sigma^2) = \frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{1}{\sigma^2}\right)^{\alpha+1}\exp\left\{\frac{-\beta}{\sigma^2}\right\} \qquad E[\sigma^2] = \frac{\beta}{\alpha-1} \qquad sd(\sigma^2) = \frac{\beta}{(\alpha-1)\sqrt{\alpha-2}}$$

$$p(\mu,\sigma^2) = p(\sigma^2)p(\mu|\sigma^2) = \frac{\beta^\alpha\sqrt{\tau}}{\Gamma(\alpha)\sqrt{2\pi}}(\sigma^2)^{-1/2}(\sigma^2)^{-(\alpha+1)}\exp\left\{\frac{-\beta}{\sigma^2}\right\}\exp\left\{-\frac{\tau}{2\sigma^2}(\mu-\mu_0)^2\right\}$$

$$= \frac{\beta^\alpha\sqrt{\tau}}{\Gamma(\alpha)\sqrt{2\pi}}(\sigma^2)^{-(\alpha+1+\frac{1}{2})}\exp\left\{-\frac{1}{2\sigma^2}(2\beta+\tau(\mu-\mu_0)^2\right\}$$

$y_1,\dots,y_n$ data: Bayes thm $\Rightarrow$

$$\pi(\mu,\sigma^2|\underline{y}) \propto \pi(\mu,\sigma^2)\prod_{i=1}^n \frac{1}{\sigma}\exp\left\{-\frac{1}{2\sigma^2}(y_i-\mu)^2\right\}$$

$$\propto (\sigma^2)^{-(\alpha+1+\frac{1}{2}+\frac{n}{2})}\exp\left\{-\frac{1}{2\sigma^2}\left(2\beta+\tau(\mu-\mu_0)^2+\sum_{i=1}^n(y_i-\mu)^2\right)\right\}$$

$$\propto (\sigma^2)^{-(\alpha+1+\frac{n+1}{2})}\exp\left\{-\frac{1}{2\sigma^2}\left(2\beta+\tau\mu^2-2\mu_0\tau\mu+\mu_0^2\tau+\sum_{i=1}^n y_i^2-2n\bar{y}\mu+n\mu^2\right)\right\}$$

$$\propto (\sigma^2)^{-(\alpha+1+\frac{n+1}{2})}\exp\left\{-\frac{1}{2\sigma^2}\left(2\beta+(\tau+n)\mu^2-2(\mu_0\tau+n\bar{y})\mu+\mu_0^2\tau+\sum_{i=1}^n y_i^2\right)\right\}$$

$$\propto (\sigma^2)^{-(\alpha+1+\frac{n+1}{2})}\exp\left\{-\frac{1}{2\sigma^2}\left((\tau+n)\left(\mu-\frac{\mu_0\tau+n\bar{y}}{\tau+n}\right)^2-(\tau+n)\left(\frac{\mu_0\tau+n\bar{y}}{\tau+n}\right)^2+2\beta+\mu_0^2\tau+\sum_{i=1}^n y_i^2\right)\right\}$$

$\therefore \alpha_n \to \alpha+\frac{n}{2}$ $\mu_n \to \frac{\mu_0\tau+n\bar{y}}{\tau+n}$ $\tau_n \to \tau+n$ $\beta_n = \beta+\frac{1}{2}\sum_{i=1}^n(y_i-\bar{y})^2+\frac{n\tau}{2(n+\tau)}(\bar{y}-\mu_0)^2$

```
>> library(invgamma)
>> BirthWeightPrior <- list(mu=3, alpha=6, beta=5, tau=16)
>> PosteriorUpdate <- function(Data, Prior){
>> n <- length(Data)          >> Posterior <- list()
>> Posterior$alpha <- Prior$alpha + n/2
>> Posterior$tau <- Prior$tau + n
>> Posterior$mu <- (Prior$mu*Prior$tau + n*mean(Data))/(Prior$tau + n)
>> Posterior$beta + 0.5*(n*var(Data) + (n*Prior$tau/(n+Prior$tau))*(mean(Data)-Prior$mu)^2)
>> return(Posterior)  >>}
>> BirthWeightPrior <- list(mu=3, alpha=6, beta=5, tau=16)
>> BirthWeightData <- rnorm(40, mean=3.5, sd=0.5)
>> BirthWeightPosterior <- PosteriorUpdate(Data=BirthWeightData, Prior=
BirthWeightPrior)
```

$P(\tilde{y} > 4.54 \mid \underline{y}) = \int_{4.54}^{\infty} \iint p(y \mid \mu, \sigma^2)\, \pi(\mu, \sigma^2 \mid \underline{y})\, d\mu\, d\sigma^2\, d\tilde{y}$

$= \int \int \mathbb{1}(\tilde{y} > 4.54)\, p(y \mid \mu, \sigma^2)\, \pi(\mu, \sigma^2 \mid \underline{y})\, du\, d\sigma^2$

Draw μ, σ^2 from $\pi(\mu, \sigma^2 \mid \underline{y})$

Draw $y \mid \mu, \sigma^2$ repeat Count num > 4.54

$\pi(\sigma^2 \mid \underline{y}) \sim$ Inv $\Gamma(\alpha_n, \beta_n)$

$\pi(\mu \mid \sigma^2, \underline{y}) \sim N(\mu_n, \frac{\sigma^2}{\tau_n})$ #

```
>> N=1000
>>SigmaSq <- rinvgamma(N, BirchweightPrior$alpha, BirchweightPrior$Beta)
```

$\mu \mid \sigma^2 \sim N(\mu_0, \frac{\sigma^2}{\tau})$

P(Y > Perry)

```
>>phat <- sum(YS>4.54)/N
```

$E[g(\theta) \mid \underline{y}] = \int_{-\infty}^{\infty} g(\theta)\, \pi(\theta \mid \underline{y})\, d\theta$

$\pi(\theta \mid \underline{y}) \propto \pi(\theta) p(\underline{y} \mid \theta)$ $\int_{-\infty}^{\infty} p(\underline{y} \mid \theta)\, \pi(\theta)\, d\theta$

let $\theta \sim Beta(a, b)$, $y \mid \theta, \underline{y} \sim Gamma(\theta + 1, \beta)$

$y_1, \ldots, y_n \Rightarrow \pi(\theta \mid \underline{y}) \overset{Bayes}{\propto} \pi(\theta) p(\underline{y} \mid \theta) \propto \theta^{a-1}(1-\theta)^{b-1} \prod_{i=1}^{n} y_i^{\theta} e^{-\beta y_i}$

$\propto \theta^{a-1}(1-\theta)^{b-1} \exp\{\theta \sum_{i=1}^{n} \log y_i\}$ $\theta \in [0, 1]$

Def 3.1 / Suppose your probab distri for r.v. θ is $P(\theta)$ let $\theta_1, \ldots, \theta_N$ be N indep random draws from $P(\theta)$. A Monte Carlo estimate for a expectation of a real valued func of θ, $g(\theta)$ where $E[g(\theta)] = \int_{\theta} g(\theta)\, dP(\theta)$ is $\bar{g}(\theta) = \frac{1}{N} \sum_{i=1}^{N} g(\theta_i)$

let $\theta \sim Beta(2, 2)$, $E[\theta] = \frac{1}{2}$ can estimate $E[\theta]$ by MC by sampling $\theta_1, \ldots, \theta_N$ from Beta(2, 2) & averaging $g(\theta_i) = \theta_i$ ~~E[θ]~~

$E[\theta] = \int_0^1 \theta \frac{\Gamma(4)}{\Gamma(2)^2} \theta(1-\theta)\, d\theta \approx \frac{1}{N} \sum_{i=1}^{N} \theta_i$

```
>> theta=0.5   >>N=10000   >>Samples <- rbeta(N, 2, 2)
>>MCestimate <- Mean(Samples)
```

$\theta \in [-1.96, 1.96]$ and $\int_\theta e^{-\frac{1}{2}\theta^2} d\theta$

$\therefore \int_{-1.96}^{1.96} e^{-\frac{1}{2}\theta^2} d\theta = \sqrt{2\pi} \int_{-1.96}^{1.96} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\theta^2} d\theta = \sqrt{2\pi}\left[\Phi(1.96) - \Phi(-1.96)\right] =$

$0.95\sqrt{2\pi}$

$\theta \sim unif(-1.96, 1.96)$, $P(\theta) = \frac{1}{2 \times 1.96} = \frac{1}{3.92}$ $\therefore$

$\int_{-1.96}^{1.96} 3.92 e^{-\frac{1}{2}\theta^2} P(\theta) d\theta$ $\quad g(\theta) = 3.92 e^{-\frac{1}{2}\theta^2}$ sample N θ

average $g(\theta)$ as usual

```
>> integral <- 0.95*sqrt(2*pi)   >> N<-10000
>> samples<-runis(N, -1.96 , 1.96)
>> g <- function (theta) {
>> 3.92 *exp(-0.5*theta^2)  >>}
>> MCestimate <- mean(g(Samples))   >>integral  >>MCestimate
>> hist (Samples)   >> hist(g(Samples))   >>
>> abs(integral - MCestimate)
```

$E[g(\theta)]$ $\quad \bar{g}(\theta) = \frac{1}{N}\sum_{i=1}^{N} g(\theta_i)$ $\quad \sigma_g^2 = var[g(\theta)]$

$var[\bar{g}(\theta)] = var\left[\frac{1}{N}\sum_{i=1}^{N} g(\theta_i)\right] = \frac{1}{N^2}\sum_{i=1}^{N} var[g(\theta_i)] = \frac{N\sigma_g^2}{N^2}$

Sample $var[g(\theta)] = \frac{1}{n-1}\sum_{i=1}^{n}(g(\theta_i) - \bar{g})^2 = S^2$ $\therefore$ S^2 is an unbiased estimator for σ_g^2 $\therefore$ Monte Carlo standard Error is $\frac{S}{\sqrt{N}}$ >>MCerror = S/sqrt(N)

```
>> abline(v=MCestimate , col=2)
>> abline(v=MCestimate -1.96*MCerror, col=4, lty=2)
>> abline(v=MCestimate +.96*MCerror, col=4, lty=2)
>> abline(v=Etheta, col=5 , lwd=(-4)
```

```
>> MCerror <- sd(g(Samples))/sqrt(N)
```

$\bar{g} = \frac{1}{N}\sum_{i=1}^{N}[g(\theta_i)]$ $\quad \theta_i \sim P(\theta)$ is $iid(g(\theta)) = 1$, for large n, by CLT

$\bar{g}(\theta) \sim N\left(E[g(\theta)], \frac{var[g(\theta)]}{n}\right)$ $\quad \frac{\bar{g}(\theta) - E[g(\theta)]}{\sqrt{var[g(\theta)]/n}} \xrightarrow[n\to\infty]{} N(0,1)$

$\bar{g}(\theta) \pm z_{\alpha/2}\frac{S}{\sqrt{n}}$ is an α level C.I. for $E[g(\theta)]$ this idea

generalises for $g: \mathbb{R}^p \to \mathbb{R}^q$ $q<p$

$\pi(\theta \in S|y) = \int_{\theta \in S} \pi(\theta|y)d\theta$ $\quad g(\theta) = \mathbb{1}(\theta \in S) = \begin{cases} 1 & \theta \in S \\ 0 & \text{otherwise} \end{cases}$

$\pi(\theta \in S|y) = \int_\theta g(\theta)\pi(\theta|y)d\theta$ is why we can sample $\theta_1, \dots, \theta_N$ from $\pi(\theta|y)$ then our MC estimate for $\pi(\theta \in S|y)$ is $\hat{p} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(\theta_i \in S)$

is $X \sim Ber(p)$ $var(X) = p(1-p)$ Monte Carlo Standard error is $\sqrt{\hat{p}(1-\hat{p})/N}$ $\quad s^2 = \hat{p}(1-\hat{p})$

Temps are $Y_i|\theta \sim N(\theta, \sigma^2)$ $\quad \theta \sim N(\theta_0, \sigma_0^2)$

$\theta_0 = 20$, $\sigma_0^2 = 0.75^2$ $\quad \sigma^2 = 15.6$ $\quad Y_1, \dots, Y_n$

$P(\tilde{y} < 10.3|y) = \int_{-\infty}^{10.3} p(\tilde{y}|y)d\tilde{y} = \int_{-\infty}^{10.3}\int_{-\infty}^{\infty} p(\tilde{y}|\theta)\pi(\theta|y)d\theta\, d\tilde{y} =$

$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbb{1}(\tilde{y}<10.3)\, p(\tilde{y}|\theta)\, \pi(\theta|y)d\theta\, d\tilde{y}$ sample θ_i from $\pi(\theta|y)$

$p(\tilde{y}, \theta|y)$

Sample $\tilde{y}_i$ from $p(\tilde{y}|\theta = \theta_i) = N(\theta_i, \sigma^2)$ {>>rnorm}

$\theta|y \sim N(\theta_n, \sigma_n^2)$ $\quad \frac{1}{\sigma_n^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}$ $\quad , \theta_n = \frac{\theta_0/\sigma_0^2 + n\bar{y}/\sigma^2}{1/\sigma_0^2 + n/\sigma^2}$

>>Temps<-c(>>Temperature<-tibble(Temp=Temps)

>>ggplot(Temperature)+geom_hist(aes(x=Temp))+xlab("Temp")

>>n<-11 >>sigma0.sq=0.75^2 >>sigma.sq=var(Temps) >>theta=20

>>sigma.n.sq<-1/(1/sigma0.sq+n/sigma.sq)

>>theta.n=(theta0/sigma0.sq+n*mean(Temps)/sigma.sq)/(1/sigma0.sq+n/sigma.sq)

>>N<-10000 >>thetaSamples<-rnorm(N,mean=theta.n,sd=sqrt(sigma.n.sq))

>>hist(thetaSamples) >>yhatSamples<-rnorm(N,mean=thetaSamples,sd=sqrt(sigma.sq))

>>hist(yhatSamples) >>phat<-sum(yhatSamples<=10.3)/N

>>MCerror<-sqrt(phat*(1-phat)/N) >>phat >>MCerror ~~>>print~~

>>paste("[",max(0,phat-1.96*MCerror),",",phat+1.96*MCerror,"]",sep="")

>>hist(thetaSamples) >>quantile(thetaSamples,probs=c(0.02,0.97))

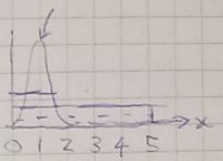
>>abline(v=18.84,col=2,lty=2) >>abline(v=21.03,col=2,lty=2)

>>quantile(yhatSamples,probs=c(0.005,0.995))

Suppose pdf $g(x)$ hard (impossible) to sample from. $E_g[X]$

$X \in [0,1]$ $\therefore \int_0^1 x g(x)\,dx$ $\quad f(x)=1$

$\therefore = \int_0^1 x g(x) f(x)\,dx$ set $g^*(x) = x g(x)$



Samples N $X_i \sim Unif(0,1)$ $\frac{1}{N}\sum g^*(x_i)$

$\int_0^1 x g(x)\,dx = \int_0^5 x g(x)\,dx = \int 5x g(x) f(x)\,dx$

$f(x)$ $Unif(0,5)$ $g^{**}(x) = 5x g(x)$

```
>>gs$<-function(x) {      >>x*dbeta(x,2,2)  >>}  >>library(tidyverse)
>>xs<-seq(from=0,to=5,length=200)
gdata2<-tibble(x=xs, g=dbeta(xs,2,2), f=rep(1,length(xs)))
>>plot2<-ggplot(gdata2)+geom_line(aes(x=x,y=g),col="blue")+
geom_line(aes(x=x,y=f),col="red")   >>plot2
>>N<-100   >>MCsamples<-gs$(runif(N,0,5))
>>MCestimate<-mean(MCsamples)   >>MCerror<-sd(MCsamples)/sqrt(N)
>>plot2+geom_vline(aes(xintercept=MCestimate),col=3)+geom_vline(aes(
xintercept=MCestimate-1.96*MCerror),col=3,lty=3)+geom_vline(aes(
xintercept=MCestimate+1.96*MCerror),col=3,lty=2)
```

Importance Sampling Theory: Let X be on domain D & let $h(x)$ be a density for X with $h(x) \neq 0$ for $x \in D$ $\quad X \sim f(x)$, $E[g(x)]$

$E[g(x)] = \int_D g(x) f(x)\,dx = \int_D \frac{g(x) f(x)}{h(x)} h(x)\,dx = E_h\left[\frac{g(x) f(x)}{h(x)}\right]$ 2 var of f E_h is minimised when $h(x) \propto |g(x) f(x)|$ eg if $g(x) \geq 0$ $\therefore$

$E_h\left[\frac{g(x) f(x)}{h(x)}\right] = \text{const}\ \forall x$ $\therefore$ $\text{var}[\hat{g}(x)] = 0$

need 1) $h(x) \geq 0$ when $g(x) \neq 0$

2) $h(x)$ 'nearly' proportional to $|g(x) f(x)|$

3) $h(x)$ easy to sample from

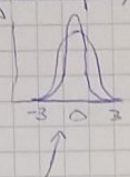
4) $h(x)$ easy to evaluate

Importance Sampling Example

$\int_{-50}^{50} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$ $[-50, 50]$ standard normal $\therefore$

$\int_{-50}^{50} g(x) S(x) dx$ $S(x) \sim \text{Unif}(-50, 50)$

$\therefore g(x) = \frac{100}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$



$h(x) \sim t_1$
$h(x) \sim t_{30}$

$\int_{-\infty}^{\infty} \frac{\mathbb{1}_{(-50 < x < 50)} g(x)}{h(x)} h(x)$

t, sat tailed, t_{30} similar to normal distri

Week 5 / Introduction to Random Number Generation

RNG's $X_{i+1} = (aX_i + c) \bmod M$ $x \in \mathbb{N}$, M large

$\frac{X_1}{M}, \ldots, \frac{X_N}{M}$ gets 'unif(0,1)'

>> ?runif >> random.seed "Mersenne-Twister"

r.v. X cdf $F(x) = P(X \le x)$ $F(X) = W$ (r.v.) what $P(W \le w)$

$W \in [0,1]$ $P(W < 0) = P(W > 1) = 0$

For $w \in [0,1]$ $P(W \le w) = P(F^{-1}(W) \le F^{-1}(w))$ if F^{-1} exists

$= P(F^{-1}(F(X)) \le F^{-1}(w)) = P(X \le F^{-1}(w)) = F(F^{-1}(w)) = w$ $\therefore$

$S(w) = F'(w) = 1$ i.e. $W \sim \text{Unif}(0,1)$

if F^{-1} exists to obtain a sample with same distri as X,

Sample U from Unif(0,1), & compute $F^{-1}(U)$

Ex/ $X \sim \text{Exp}(\lambda)$ $S(x) = \lambda e^{-\lambda x}$ $F(x) = \int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x}$ $x \ge 0$

$U = F(x) = 1 - e^{-\lambda x}$ $\therefore x = -\frac{1}{\lambda} \log(1 - U)$

>> expsamples <- rexp(100, 1) >> u <- runif(100)

>> expsamples2 <- -1 * log(1 - u) >> par(mfrow = c(1,1))

>> plot(sort(expsamples), sort(expsamples2))

Consider discrete r.v. X with values $x_1, \ldots, x_m$, probabs $p_1, \ldots, p_m$

(w.l.o.g.) $x_1 < x_2 < \ldots < x_m$ $\therefore F(x_j) = p_1 + p_2 + \ldots + p_j : F_j$

$F(x) = F_j$ $x_j \le x \le x_{j+1}$ for $p \in (F_{j-1}, F_{j+1})$ F^{-1} doesn't exist

non-unique when $p = F_j$ inverse CDF for Discrete

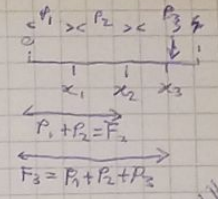
Define $F_*^{-1}(p) = x_j$ $p \in (F_{j-1}, F_j]$ then if $U \sim \text{Unif}(0,1)$,

$X_* = F_*^{-1}(U)$ $\therefore$

(i) X_* can only take vals $x_1, \dots, x_m$

(ii) $P(X_* = x_j) = P(U \in (F_{j-1}, F_j]) = F_j - F_{j-1} = P_j$

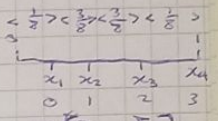
- X_* has 2 same distri as X.



Toss coin (Prob $\frac{1}{2}$) 3 times X = Num of heads $X \sim Bin(3, \frac{1}{2})$

$F_1 = P(X=x_1) = P(X=0) = \frac{1}{8}$ $\quad F_2 = P(X=1) + F_1 = \frac{3}{8} + \frac{1}{8} = \frac{4}{8}$

$F_3 = P(X=2) + F_2 = \frac{3}{8} + \frac{4}{8} + \frac{7}{8}$ $\quad F_4 = 1$ $\quad x_1=0, x_2=1, \dots, x_4=3$



Draw from $U \sim Unif(0,1)$ assign $X = F^{-1}(U)$ $\quad F^{-1}(U) = x_j$ $U \in [F_{j-1}, F_j]$

```
>>ThreeSamples <- function(N) {      >>Us <- runif(N)
>>ifelse(Us<=1/8,0, ifelse(Us<=4/8, 1, ifelse(Us<=7/8, 2, 3)))   >>}
>>hist(ThreeSamples(1000), seq, =FALSE)  >>hist(rbinom(1000, 3, prob=0.5))
```

Box-Muller X_1, X_2 iid $N(0,1)$ $\quad p(x_1,x_2) = \frac{1}{2\pi} \exp\left\{-\frac{1}{2}(x_1^2+x_2^2)\right\}$ ...

$X_1 = R\cos\theta$, $X_2 = R\sin\theta$, $|J| = R$ ∴

$f(r,\theta) = \frac{1}{2\pi} r e^{-\frac{r^2}{2}}$ $\Rightarrow$ r, θ also indep $p(\theta) = \frac{1}{2\pi}$ $\quad p(r) = re^{-\frac{r^2}{2}}$

∴ $p(\theta) \sim Unif[0, 2\pi]$

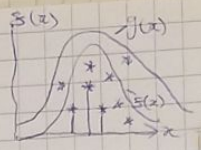
let $W = R^2$, $\frac{\partial R}{\partial W} = \frac{1}{2\sqrt{W}}$ ∴ $p(W) = \frac{1}{2\sqrt{W}} \sqrt{W} e^{-\frac{W}{2}} = \frac{1}{2} e^{-\frac{W}{2}}$ ∴ $W \sim Exp(\frac{1}{2})$

To sample $X_1, X_2 \sim N(0,1)$ generate θ from $Unif(0, 2\pi)$

generate $U \sim Unif(0,1)$ $\quad W = -2\log(1-U)$ $\quad X_1 = \sqrt{W}\cos\theta$, $X_2 = \sqrt{W}\sin\theta$

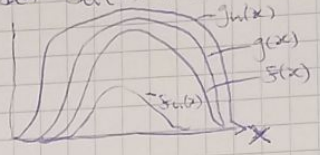
```
>>BoxMuller <- function(N) {
>>M<-ceiling(N/2)   >>thetas <- 2*pi*runif(M)
>>Us <- runif(M)    >>Ws <- -2*log(1-Us)
>>X1 <- sqrt(Ws)*cos(thetas)   X2 <- sqrt(Ws)*sin(thetas)
>>c(X1, X2)[1:N]    >>}
>>BoxMuller(100)    >>hist(BoxMuller(1000))
>>plot(sort(rnorm(1000)), sort(BoxMuller(1000)))
```

X density $f(x)$ can sample from $g(x)$
but reject points not under $f(x)$

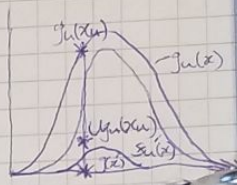


X pdf $f(x)$ let $g(x) = cf(x)$ const c
we can sample from
first set $g_u(x) = k f_u(x) \geq g(x)$

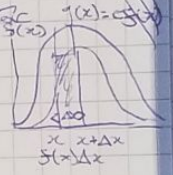
let $f_u(x)$ be a distri



1) Draw X_u from $f_u(x)$, $U \sim \text{Unif}(0,1)$
2) accept X_u (as a sample from $f(x)$) if $U g_u(X_u) \leq g(X_u)$



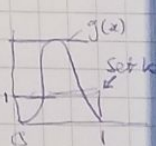
Claim
x coord of a random pt from Z area under $f(x)$ ($X \sim f(x)$),
X_*, has Z same distri as X. $P(X_* \in [x, x+\Delta x]) \approx \Delta x f(x)$ for Δx small $\therefore f_{X_*}(x) = F'_{X_*}(x) =$



$$\lim_{\Delta x \to 0} \frac{F_{X_*}(x+\Delta x) - F_{X_*}(x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{P(X_* \in [x, x+\Delta x])}{\Delta x} = f(x)$$

let $g(x) = cf(x)$ & X_* is a random pt from Z area under $g(x)$
because $P(X_* \in [x, x+\Delta x]) \approx \frac{\Delta x\, g(x)}{c} = \frac{\Delta x\, c f(x)}{c} = \Delta x f(x)$
rejection sampling samples from a random pt from area under $g_u(x) = k f_u(x)$
accept is that pt is in Z area under $g(x)$
acceptance Rate $= \frac{\text{area under } g}{\text{area under } g_u} = \frac{c}{k}$

$X \sim \text{Beta}(a,b)$ sample X letting $A = a-1$, $B = b-1$
$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^A (1-x)^B$ $g(x) = c f(x)$ $g(x) = x^A (1-x)^B$



$f_u(x)$ a distri can sample from $g_u(x) = k f_u(x) \leq g(x)$ $\forall x$
$f_u(x)$ Uniform on domain of X. $f_u(x) = 1$
choose k s.t. $k f_u(x) = k \geq g(x)$ $\forall x$
maximise $g(x)$ $h(x) = \log g(x) = A \log x + B \log(1-x)$ $\therefore$
$h'(x) = \frac{A}{x} - \frac{B}{1-x} = 0 \iff x = \frac{A}{A+B}$ $\therefore k = \max g(x) = g\left(\frac{A}{A+B}\right)$

```
>>a=5   >>b=1.2   >>A<-a-1   >>B<-b-1
>>gfun<-function(x){   >>(x^A)*(1-x)^B   >>}
   >>tBound<-gfun(A/(A+B))   #k#
>>xs<-seq(from=0, by=0.005, to=1)   >>gx<-gfun(xs)
>>plot(xs,gx,type='l',xlab="x",ylab="g(x)",ylim=c(0,tBound+0.5))
>>points(xs,dbeta(xs,a,b),type='l',lty=2)
>>abline(h=tBound,lty=2,col=4)   >>abline(v=0,lty=1,col=1)
>>abline(v=1,lty=1,col=1)   >>abline(h=1)
>>addpoint<-function(tgfun=gfun,bound=tBound){
>>Us<-runif(2)   >>if(bound*Us[2]<=tgfun(Us[1])){
   >>points(Us[1],bound*Us[2],col=3,pch=16)   >>return(Us[1])   >>}
>>else{points(Us[1],bound*Us[2],col=2,pch=4)   >>return(NA)   >>}}
>>addpoint()   >>Sim100<-sapply(1:100,function(k) addpoint())
>>sum(!is.na(Sim100))/100
```

A density $g(x)$ is log concave if $\frac{d^2 \log g(x)}{dx^2} < 0 \ \forall x$

Proposition Let $h(x) = \log g(x)$, suppose $0 > H = \sup h''(x)$, & $\hat{x} = \arg\max_x h(x)$, then with $g_u(x)$ 2 pdfs of $N(\hat{x}, -\frac{1}{H})$

$g_u(x) = k S_u(x) = e^{h(\hat{x})} e^{\frac{H}{2}(x-\hat{x})^2} \geq g(x) \ \forall x$

i.e. a rejection sampler is: i) Draw X_u from $N(\hat{x}, -\frac{1}{H})$

(set $X_u = \hat{x} + \sqrt{-\frac{1}{H}}\, Z \quad Z \sim N(0,1)$)

ii) accept X_u if $U g_u(X_u) \leq g(X_u) \quad U \sim Unif(0,1)$

Proof $h(x) < h(\hat{x}) + \int_{\hat{x}}^{x} h'(t)\,dt = h(\hat{x}) + \int_{\hat{x}}^{x} \left[h'(\hat{x}) + \int_{\hat{x}}^{t} h''(s)\,ds \right] dt$

$= h(\hat{x}) = \int_{\hat{x}}^{x} \int_{\hat{x}}^{t} h''(s)\,ds\,dt \leq h(\hat{x}) + \int_{\hat{x}}^{x} \int_{\hat{x}}^{t} H\,ds\,dt = h(\hat{x}) + \frac{H}{2}(x-\hat{x})^2$

$h_u(x) = h(\hat{x}) + \frac{H}{2}(x-\hat{x})^2 \quad \therefore h_u(x) \geq h(x) \ \forall x$

$\Rightarrow g_u(x) = \exp\{h_u(x)\} \geq g(x) \ \forall x$

$g_u(x) = e^{h(\hat{x})} e^{\frac{H}{2}(x-\hat{x})^2} = \underbrace{e^{h(\hat{x})} \sqrt{\frac{2\pi}{-H}}}_{k} \underbrace{\sqrt{\frac{-H}{2\pi}}\, e^{\frac{H}{2}(x-\hat{x})^2}}_{\text{Normal}(\hat{x}, -\frac{1}{H})} = k S_u(x)$

$X \sim Beta(a,b)$ $A=a-1$, $B=b-1$ $h'(x) = \frac{A}{x} - \frac{B}{1-x}$, $\hat{x} = \frac{A}{A+B}$

$h''(x) = -\frac{A}{x^2} - \frac{B}{(1-x)^2}$ if $a,b>1$ then $h''(x)<0$

g is log concave $h'''(x) = \frac{2A}{x^3} - \frac{2B}{(1-x)^3} = 0 \iff x = \frac{A^{1/3}}{A^{1/3}+B^{1/3}}$

$H = h''\left(\frac{A^{1/3}}{A^{1/3}+B^{1/3}}\right) = -\left(A^{1/3}+B^{1/3}\right)^3$

$g_u(x) = e^{h(\hat{x})} e^{\frac{H}{2}(x-\hat{x})^2} = \hat{x}^A (1-\hat{x})^B e^{\frac{H}{2}(x-\hat{x})^2}$

Sample from $N\left(\frac{A}{A+B}, \frac{1}{(A^{1/3}+B^{1/3})^3}\right)$

```
>> # Rejection with log concave beta example #
>> betaSams <- function(n,a,b){   >> A=a-1  >> B=b-1  >> xhat <- A/(A+B)
>> H <- (-1)*(A^(1/3) + B^(1/3))^3
>> z <- rnorm(n)  # normalSams(n) via Box-Muller #
>> Xus <- xhat + z*(1/sqrt(-H))   >> Us <- runif(n)
>> hxhat <- A*log(xhat) + B*log(1-xhat)
>> c <- gamma(a)*gamma(b)/gamma(a+b)  >> k <- exp(hxhat)*sqrt(2*pi/(-H))
>> xs <- seq(from=0, by=0.005, to=1)  >> gx <- dbeta(xs,a,b)
>> plot(xs, c*gx, type='l', xlab="x", ylab='g(x)', ylim=c(0, k*dnorm(xhat, xhat,
sqrt(-(1/H)))))
>> points(xs, k*dnorm(xs, xhat, sqrt(-(1/H))), type='l', col=2)
>> print(paste("acceptance rate = ", c/k, sep=" "))
>> Xus[Us*exp(hxhat)*exp((1/H/2)*(Xus-xhat)^2) <= (Xus^A)*((1-Xus)^B)]}
>> # undebug(betaSams) #
>> betaSams(10000, 5, 1-2)
>> tbetaSams(10000, 5, 1-2)  >> hist(tbetas, prob=F, breaks=100, xlim=c(0,1))
>> xs <- seq(from=0, by=0.005, to=1)  >> gx <- dbeta(xs, 5, 1-2)
>> plot(xs, gx, type='l', xlab="x", ylab='g(x)', ylim=c(0, 3.5))
```

Monti / 1/

$\int_0^i e^{\sin x} x^2 (1-x)^3$ can be solved by Monte Carlo

Beta(3,4)

$\{\int g(x) f(x) dx\}$ $f(x)$ "is density we sample from $E[g(x)] = \int g(x) f(x) dx$

$X_1, \ldots, X_n \sim f(x)$ $E[g(x)] \approx \frac{1}{N} \sum_{i=1}^{N} g(X_i)$

) $X \sim Beta$

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

or $f(x) = 1$ $X \sim Unif(0,1)$

2 Monte Carlo error shrinks at a rate proportional to $\frac{1}{\sqrt{N}}$

$E[g(x)] \triangleq \int_{-\infty}^{\infty} g(x) f(x) dx$ $f(x) = 0$ if $x \notin [0,1]$

always use "exchangeable" instead of iid but is much weaker

$\pi(\theta|y) \propto \pi(\theta) p(y|\theta)$

mentí

\4/ $\pi(\theta|y_i) \propto \pi(\theta) p(y_i|\theta)$?

$y = y_1, \ldots, y_n$ exchangeable with $p(y_i|\theta)$ & $\pi(\theta)$ ∴ to eval $p(y_{n+1} > 0.5|y)$ must sample $y'_{n+1} \sim p(y_{n+1}|\theta = \theta_i)$

\5/ $y = y_1, \ldots, y_n$ exchangeable with $p(y_i|\theta)$ & $\pi(\theta)$ ∴ to eval $p(y_{n+1} > 0.5|y)$ must eval $\mathbb{1}(y'_{n+1} > 0.5)$

$$p(y_{n+1} > 0.5|y) = \int_{0.5}^{\infty}\int_{-\infty}^{\infty} p(y_{n+1}, \theta|y) dy d\theta = \int_{0.5}^{\infty}\int_{-\infty}^{\infty} p(y_{n+1}|\theta)\pi(\theta|y) dy_{n+1} d\theta$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbb{1}(y_{n+1} > 0.5) p(y_{n+1}|\theta)\pi(\theta|y) dy d\theta$$

~~1 of 2~~ $\int g(x) f(x) dx$ $\mathbb{1}(y_{n+1} > 0.5)$ is $g(x)$ $p(y_{n+1}, \theta|y)$ is $f(x)$

1) $\theta_1, \ldots, \theta_n \sim \pi(\theta|y)$

2) $y'_{n+1}, \ldots, y'_{n+1} \sim p(y_j|\theta = \theta_j)$ $j = n+1, \ldots, n+n$

$p(x,y) = p(x|y)p(y) = p(y|x)p(x)$

\6/ have $y = y_1, \ldots, y_n$ exchangeable with $p(y_i|\theta)$ & $\pi(\theta)$ to eval $p(y_{n+1} > 0.5|y)$ finally repeat the last 3 steps as many times and and average the $I(y'_{n+1} > 0.5)$

\4/ $y|\theta \sim S(y|\theta)$ $\pi(\theta) \xrightarrow{\text{Bayes}} \pi(\theta|y)$

$E[g(\theta)|y] = \int g(\theta)\pi(\theta|y)d\theta$

$E[g(y)|y] = \int g(y)p(y|\theta)\pi(\theta|y)d\theta$

Sample $\pi(\theta|y) \rightarrow p(y|\theta)$

$\pi(\theta_1, \theta_2, \cdots, \theta_d|y) = \pi(\theta_d|\theta_{d-1}, \cdots, \theta_1, y)$

$\pi(\theta_d|\theta_{d-1}, \cdots, \theta_1, y)\,\pi(\theta_{d-1}|\theta_{d-2}, \cdots, \theta_1, y) \cdot \pi(\theta_{d-2}|\theta_{d-3}, \cdots)\cdots$

$\pi(\theta_2|\theta_1, y)\,\pi(\theta_1|y)$

inverse CDF / Discrete / continuous

Unif → N (Box Muller)

Unif → Anything (Rejection Sampling $g(x) \propto S(x)$)

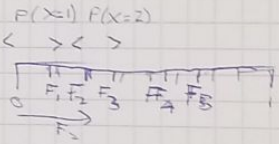
\Wenti/ \1/ Sie die inverse CDF method & $U \sim Unif(0,1)$

what val of X does true $U = 0.345$ give? : 3

$F(x)$ $F_1 = P(X=1) = \frac{1}{6}$

$F_2 = P(X \le 2) = P(X=2) + P(X=1) = \frac{2}{6}$

$F_3 = \frac{3}{6}$ $F_4 = \frac{4}{6}$ $F_5 = \frac{5}{6}$ $F_6 = 1$

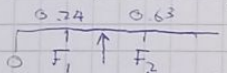


\2/ inverse CDF für die

\3/ $X \sim Bin(5, 0.25)$ convert $U = 0.55 \sim Unif(0,1)$ into a sample of 2 same distri as X: 1

>> placeholder(0.5, 5, 0.25)

⇒ 0.237, 0.632, ...



$F_1 = P(X = x_1) = P(X=0) = \binom{5}{0} 0.25^0 0.75^5 = 0.237$

$F_2 = P(X = x_2) + F_1 = 0.237 + P(X=1) = \binom{5}{1} 0.25^1 0.75^4 = 0.632$

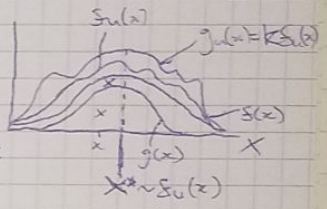
\4/ $X \sim Bin(5, 0.25)$ convert $U = 0.922 \sim Unif(0,1)$ into a Sample from 2 same distri of X

∴ 0.237, 0.632, 0.896, 0.984, 0.999, 1 ∴ 3

\5/ $U = 0.102 \sim Unif(0,1)$ ∴ 0

\6/ 2 rejection sampling algorithms involves 4 funcs
f, f_u, g, g_u which of these are densities? f and f_u

\7/ let $X \sim f(x)$ & $g(x) = cf(x)$ which of 2 following is most precise?
2 x-coords of a random sample from 2 area under $f(x)$ has 2 same distri as X



$X \sim f(x)$, $g(x)$ $f_u(x)$ $g_u(x)$
$g(x) = cf(x)$ points in $g(x)$ which is under $f(x)$ has distrix
$f_u(x)$: Easy to evaluate
$g_u(x) = kF_u(x)$

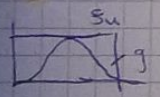
\8/ let $X \sim f(x)$ & $g(x) = cf(x)$. 2 x-coord of a sample from under $g(x)$ has pdf $f(x)$

\9/ let f, f_u, g, g_u be 2 functions used in rejection sampling
which of these is 2 target density of our sample?: f
$g_u(x) \geq g(x) \quad \forall x$
sample at random from area under $g_u(x)$ (X^*, y^*) is $y^* \leq g(x^*)$
(X^*, y^*) is a random pt from area under $g(x)$
$\therefore$ keep all samples under g . x-coords are samples from f
$U \sim Uni(0,1)$ $\therefore$ $Ug_u(X^*)$ $Uni(0, g_u(X^*))$ $\therefore$ $(X^*, Ug_u(X^*))$ is random under $g_u(X)$ $\therefore$ $Ug_u(X^*) \leq g(X^*)$

\10/ let f, f_u, g, g_u be 2 funcs used in rejection sampling. which func do we sample pt randomly from 2 area underneath: g_u or g

\11/ f, f_u, g, g_u which use to bound our target? g_u

\12/ for optimal bounding box algor which is univar density: f_u



\13/ $g(x) = cf(x)$ & $g_u(x) = kf_u(x)$
want c/k large

Sheet Week 5

3a) $\delta(x) \propto \begin{cases} x^3 & x \in [0,3] \\ 0 & \text{otherwise} \end{cases}$

$F(x) = \int_{-\infty}^{x} \delta(t)\,dt \qquad \delta(x) = kx^3 \qquad \int_{-\infty}^{\infty} kx^3\,dx = 1 \Rightarrow$

$\int_0^3 kx^3\,dx = \left[\frac{kx^4}{4}\right]_0^3 = \frac{81}{4}k = 1 \quad \therefore k = \frac{4}{81}$

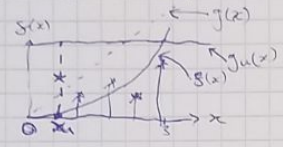
$F(x) = \int_0^x \frac{4}{81}t^3\,dt = \left[\frac{1}{81}t^4\right]_0^x = \frac{x^4}{81}$

$U = \frac{x^4}{81} \Rightarrow x = \sqrt[4]{81U} = 3U^{1/4} = F_0^{-1}(U)$

>> Us ← runis(1000) >> Xs ← 3*Us^(1/4)

>> hist(Xs)

3b)



$g(x) = x^3 \quad g(x) = cs(x) \quad c = \frac{81}{4}$

$g_u(x) = k\delta_u(x) \geq g(x)\ \forall x \quad \therefore \delta_u(x) = \frac{1}{3} \quad g_u(x) = 3^3 = 27 \quad \therefore k = 81$

$\therefore \frac{c}{k} = \frac{81}{4 \cdot 81} = \frac{1}{4}$ { sanity check is it an acceptance rate? is it between 0 & 1?

generate X_u from $\delta_u(x)$

$(X_u, U_{g_u}(X_u)) \quad U \sim runis(0,1) \quad \therefore (X_u, U_{g_u}(X_u)) = (X^*, Y^*)$

$\therefore$ accept X^* is $Y^* \leq g(X^*)$

>> reject ← function(N) { >> Xus ← runis(N,0,3) # 3*runis(N) →

>> Us ← runis(N) >> Xus[Us*27 <= Xus^3] >> }

>> Nsam ← length(Sams)

>> while(Nsam<1000) {

>> sams ← c(Sams, reject(1000))

>> Nsams ← length(Sams) >> }

>> Sams ← Sams(1:1000)

>> plot(sort(Xs), sort(Sams))

→ produces line y=x to show they are sampled from the same model

Sheet Week 5/ 5a/ $g(x) \propto \begin{cases} (2+x)(2-x) & x \in [-2,2] \\ 0 & \text{otherwise} \end{cases}$

Let $g(x) = (2+x)(2-x)$ find c s.t. $g(x) = c\,\delta(x)$

$\therefore g(x) = 4 - x^2 \quad \therefore \int_{-2}^{2} g(x)\,dx = c = \left[4x - \frac{x^3}{3}\right]_{-2}^{2} = 8 - \frac{8}{3} + 8 - \frac{8}{3} = \frac{32}{3} = c$

5d/ Show $g(x)$ is log concave

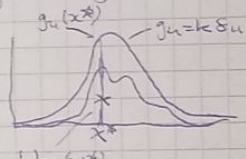
$\therefore$ Show $h''(x) \leq 0 \; \forall x \quad h(x) = \log(x) \; \therefore$

$h'(x) = \frac{-2x}{4-x^2} \quad \therefore h''(x) = \frac{-1(2x)}{(4-x^2)^2} - \frac{4x^2}{(4-x^2)^2}$ ← B $\quad h''(x) = \frac{-2}{(4-x^2)} - \frac{4x^2}{4-x^2)^2} < 0 \; \therefore$

log concave

5e/ $\hat{x} = \arg\max g(x) \quad , H = \sup h''(x) < 0$

$\therefore \delta_u(x)$ pdf $N(\hat{x}, -\frac{1}{H}) \quad g_u(x) = k\,\delta_u(x) \quad g_u(x) \geq g(x) \quad \forall x$



$K = e^{h(\hat{x})}\sqrt{\frac{2\pi}{-H}}$

$x^* = \sqrt{\frac{1}{H}}\,Z + \hat{x} \quad U \sim Unif(0,1)$

accept x^* if $Ug_u(x^*) \leq (2+x)(2-x)$

$\delta_u(x) = \frac{\sqrt{-H}}{\sqrt{2\pi}} e^{H(x-\hat{x})^2} Ug_u(x^*)$

$\theta = (\theta_1, \dots, \theta_d)$ $\pi(\theta|y)$ $\pi(\theta_1|\theta_{-1}, y)$ $\pi(\theta_2|\theta_{-2}, y)$

$\pi(\theta_d|\theta_{-d}|y)$

$y^{(1)} = (\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_d^{(0)}) = (\theta_1^{(1)}, \theta_2^{(1)}, \theta_3^{(0)}$

>>theta_j_post <- function(j, mu, sigma2, tau2, ybars, nj) {

>> rnorm(1, mean(mu/tau2 + nj*[j]*ybars[j]/sigma2)/(1/tau2 + nj/

>> library(coda)

$P(\theta_4 \le \theta_1) = E[\mathbb{1}(\theta_4 > \theta_1)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbb{1}(\theta_4 > \theta_1)\,\pi(\theta_1, \theta_4|y)\,d\theta$

∴ Monte Carlo estimate using 2 samples is ∴

>> phat <- sum(ThinnedSamples$Theta.4 > ThinnedSamples$Theta.1)/Nthinned

θ^{t-1} $\theta^* \sim N(\theta^{t-1}, \Sigma)$ Propose $\theta^* \sim q(\theta^*|\theta^{t-1})$

accept with prob $p(\theta^*, \theta^{t-1})$

metropolis ratio $\theta^* \sim N(\theta^{t-1}, \Sigma)$ $r(\theta^*, \theta^{t-1}) = \frac{\pi(\theta^*|y)/q(\theta^*|\theta^{t-1})}{\pi(\theta^{t-1}|y)/q(\theta^{t-1}|\theta^*)}$

$\log(\pi(\theta^*|y)) - \log(\pi(\theta^{t-1}|y))$

$\Sigma = \begin{pmatrix} S_L & 0 \\ 0 & S_x \end{pmatrix}$ $\frac{1}{\sqrt{N_{est}}} S_c(\theta)$ {Nest is effective sample size}

Week 8 tutorial lecture / Data $y|\theta \sim f(y;\theta)$

prior $\pi(\theta) \xrightarrow{\text{Bayes}} \pi(\theta|y) \propto \pi(\theta) \prod_{i=1}^{N} f(y_i|\theta)$

$P(a < \theta < b|y) = \int_a^b \pi(\theta|y)\,d\theta$ $P(y_{new}|y) = \int p(y|\theta)\,\pi(\theta|y)\,d\theta$

linear model: $y = \beta_0 + \beta_1 x$ $N(\beta_0 + \beta_1 x, \sigma^2)$ can do 95% error bars: $\mathbb{I}$

list: $\pi(\theta_1, \dots, \theta_n)$ is a prior ∴ when condition on a param: $\pi(\theta_1, \dots, \theta_n|y)$

$\pi(\theta_1|\theta_n, \theta_{n-1}, \dots, y)$

$\pi(\theta_1, \theta_2|y)$ ∴ $\pi(\theta_1|\theta_2, y) \overset{\text{bayes}}{\propto} \pi(\theta_1|y)\,p(\theta_2|y, \theta_1)$

$\pi(\theta_1|y)$ $p(\theta_2|\theta_1, y)$

\Monti/ \1/ what call property: $p(\theta^t|\theta^{t-1}, \theta^{t-2}, \dots, \theta^1) = p(\theta^t|\theta^{t-1})$ is Markov property (Markov chains)

\2/ let X^1, X^2 be a Markov chain. what 2 rules of $q(y|x) = p_{X^{(t+1)}|X^{(t)}}(y|x)$? 2 transition density

\3/ Markov chain $X^1, X^2, \dots$ $p(x)$ s.t. $\sigma(y) = \int q(y|x)\,p(x)\,dx$

2 equilibrium distn

\4/ whats 2 role of equilibrium distri for Bayesian sampling
so 2 posterior is 2 equalibrium

\5/ let X be a vec of params & want to sample from $P(X|y)$ by Gibbs sampling. Step 1? start at X^0 with $P(X^0|y)>0$

\6/ let X be a vec of params & want to sample from $P(X|y)$ by Gibbs sampling. Step 2? Sample from $P(X_1^1 | X_{-1}^0, y)$
since whole start vec $X_{-1}^0 = X_2^0, \ldots X_n^0$

\7/ let X be a vec of params & want to sample from $P(X|y)$ by Gibbs sampling. Step j+1? sample from $P(X_j^1 | X_{-j}^0, y)$
since desired: $X_{-j}^0 = X_1^1, \ldots, X_{j-1}^1, X_{j+1}^0, X_n^0$

\8/ 2 Metropolis hastings algorithm, what is 2 role of 2 posterior distri? 2 equalibrium distri of 2 Markov chain

\9/ in 2 metropolis hastings alg what $q(\cdot|\cdot\cdot)$
2 proposal distri

\10/ $\pi(\theta^*) q(\theta^*|\theta^{t-1}) / \pi(\theta^{t-1}) q(\theta^{t-1}|\theta^*)$
is 2 Metropolis ratio

\11/ what makes 2 Metropolis alg & a special case of Metropolis hastings? $q(y|x) = q(x|y)$ ∴

$$\frac{\pi(\theta^*|y) q(\theta^*|\theta^{t-1})}{\pi(\theta^{t-1}|y) q(\theta^{t-1}|\theta^*)} = \frac{\pi(\theta^*|y)}{\pi(\theta^{t-1}|y)}$$

start at θ^{t-1} ∴ propose $\theta^* \sim q(\theta^*|\theta^{*-1})$
calc $r(\theta^*, \theta^{t-1})$
accept: set $\theta^t = \theta^*$ prob $\min(1, r)$ $\theta^t = \theta^{t-1}$ otherwise

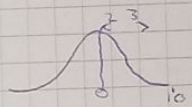
\12/ has this Markov chain converged? yes
\13/ is this Markov chain well mixed? yes
\14/ is this Markov chain well mixed? yes
\15/ has this Markov chain converged? No
\16/ has this Markov chain converged? No have to run for longer

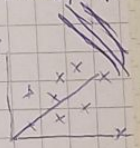
\17/ is this Markov Chain well mixed? No

\18/ has this Markov chain converged? No

```
>>library(brms)    >>library(tidyverse)
>>options(mc.cores=parallel::detectCores())
>>library(bayesplot)    >>library(reshape2)
>>load("~Dropbox/Teaching/Bayes Courses/Lecs/drinking.RData")
>>drinking<-as_tibble(drinking)
>>drinking<-drinking %>% rename(births=Births)
>>?melt
>>drinking %>% melt(id.vars='cirrhosis') %>% ggplot() +
geom_pt(aes(x=value, y=cirrhosis, colour=variable)) +
facet_wrap(~variable, scales="free_x") +
geom_smooth(aes(value, y=cirrhosis, colour=variable), method='lm')
# y~N(a+bx, σ²) #  >>lm(cirrhosis~urban, data=drinking)
>>summary(lm(cirrhosis~...))  # isn't for bayes :. need to do bayes:#
>>mod1<-brm(cirrhosis~urban, data=drinking)
>>summary(mod1)    >>plot(mod1)    >>mod1$prior  #{P(ynew|y)}#
>>nrow(drinking$urban)  >>urban_new<-seq(from=27, to=37, by=1)
>>pred1<-predict(mod1, newdata=X
>>urban<-tibble(urban=seq(from=27,to=37,by1))
>>pred1<-predict(mod1, newdata=urban)  >>pred1
# E[y] ∫∫P(y|y,θ)π(θ|y)dθdy   P(Q2.5 < ynew < Q97.5) = 0.95
>>plotData<-cbind(urban, pred1)  >>plotData
>>ggplot(plotData)+geom_line(aes(x=urban, y=Estimate), col=2)+
geom_line(aes(x=urban, y=Q2.5), col=2, lty=2)+
geom_line(aes(x=urban, y=Q97.5, col=2, lty=2)+ylab("cirrhosis")+)
geom_pt(data=drinking, aes(x=urban, y=cirrhosis), col=4, pch=16)
#only 3 pts outside 2 quartiles but 2 outliers are fine (4 standard)#
```

```
>>Model2 <- brm(cirrhosis ~ Urban + birthus + wine + spirits, data = drinking)
>>Summary(Model2)  # check its converged sirst :. rhat = 1
   traceplots are sine   >>plot(model2, N=3)  #  >>plot(model2)#
>>predsull <- predict(Model2, newdata = drinking)   >>predsull
>> Sulldata <- cbind (drinking, predsull)   >> Sull data <- Sullstate[, -7]
>> Sull data %>% Melt(id. var("cirrhosis", "Estimate", "Q2.5", "Q97.5"))
% >%  ggplot() + geom_errobar(...) + geompt(y=Estimate) + geom-pt(y=Cirrhosis) +
Facet-wrap(~variables, scales="free_x")  # priors are cutteatly
default but need to be changed to reduce thes max uncertainty #
>>urbanPrior <- set_prior("normal()")  # unless 100% Certain set
   a prior mean 0 but can chang σ :. #
>> urban prior <- set_prior("normal(0,3)")  #
>>urbanPrior <- set_prior("normal(0,3)", class="b", coef="urban")
>>wineprior <- set_prior("normal(0,3)", class"b", coef="wine")
>> intercep <- set_prior("normal(0,100)", class="Intercept")
>> sdprior <- set_prior("normal(0,10)", class = sigma")
# can use normal, beta, gamma, t with 3 degrees of free #
Mod2 <- brm(cirrhosis ~ ... , prior = c(urbanprior, wineprior, ..., ~~interBprior~~
   interceptprior, sdprior)
```





Week 7

$y = X\beta + \varepsilon$ $\beta_1, \dots, \beta_p$ $\varepsilon \sim N(0, \sigma^2 I)$ $y|\beta, \sigma^2 \sim N(X\beta, \sigma^2 I)$

$\pi(\beta, \sigma^2)$ objective prior $\pi(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$ $(\beta \sim Unif(-\infty, \infty))$

$\pi(\beta, \sigma^2 | y) \propto \pi(\beta | \sigma^2, y)\, \pi(\sigma^2 | y)$ $\beta|\sigma^2, y \sim N(\underbrace{(X^TX)^{-1}X^Ty}_{\hat\beta}, (X^TX)^{-1})$

$\sigma^2 | y \sim Inv\text{-}\chi^2(n-k, s^2)$ $s^2 = \frac{1}{n-k}(y - X\hat\beta)^T(y - X\hat\beta)$

mle β is $\hat\beta$ mle σ^2 is s^2 $E[\sigma^2 | y] = \frac{(n-k)s^2}{n-k-2}$ $\beta|\sigma^2 \sim N(\beta_0, V_0)$

$y_i \in \{0, 1\}$ $y_i | X_i, \theta \sim Ber(\eta(X_i, \theta))$

$\eta(X_i, \theta) = Logit(\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p)$ $Logit(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1}$

$\pi(\beta | y) \propto \pi(\beta)\, p(y|\beta) \propto \pi(\beta) \prod_{i=1}^{n} \left(\frac{e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}}{1+e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}}\right)^{y_i} \left(\frac{1}{1+e^{\beta_0 + \beta_1 X_1 + \dots}}\right)^{1-y_i}$

Bayesian Hierarchical Models $p(y|\theta)$, $\pi(\theta)$

Data layer: $p(y|\theta, \phi)$ (likelihood)

process layer: $p(\theta|\phi)$ Model for params

prior layer $p(\phi)$

Ex3.10 Normal Hierarchical Model: J groups, measurements on individuals; in each group. (treatments in J hospitals)

Data layer is $y_{ij}|\theta_j, \sigma^2 \sim N(\theta_j, \sigma^2)$

process layer $\theta_j|\mu, \tau^2 \sim N(\mu, \tau^2)$

prior layer $\pi(\mu, \tau^2, \sigma^2)$ $\pi(\sigma^2) \propto \frac{1}{\sigma^2}$ $\pi(\mu|\tau^2) \propto 1$ $\pi(\tau^2) \propto \frac{1}{\tau^2}$

$\pi(\sigma^2)\pi(\mu|\tau^2)\pi(\tau^2)$

$\theta = \theta_1, \dots, \theta_J$ $\pi(\theta, \mu, \sigma^2, \tau^2|y) = \pi(\mu, \tau^2|\theta, \sigma^2, y)\pi(\theta, \sigma^2|y)$

$= \pi(\mu, \tau^2|\theta)\pi(\theta|\sigma^2, y)\pi(\sigma^2|y) =$

$\pi(\mu|\tau^2, \theta)\pi(\tau^2|\theta)\pi(\theta|\sigma^2, y)\pi(\sigma^2|y)$

$\pi(\theta|\sigma^2, y) \overset{\text{Bayes}}{\propto} \pi(\theta|\sigma^2)p(y|\theta, \sigma^2) \propto$

$\pi(\theta|\sigma^2)\prod_{j=1}^{J}\prod_{i=1}^{n_j} p(y_{ij}|\theta_j, \sigma^2) \propto \int \pi(\theta, \mu, \tau^2)d\mu\, d\tau^2 \prod_{j=1}^{J}\prod_{i=1}^{n_j} p(y_{ij}|\theta, \sigma^2)$

$\propto \int \pi(\theta|\mu, \tau^2)\pi(\mu, \tau^2)d\mu\, d\tau^2$

Def 3.2 A Markov chain is a sequence of r.v.s for which, for any t, the distri $\theta^t|\theta^{t-1}, \theta^{t-2}, \dots, \theta^1$ only depends on θ^{t-1}

$p(\theta^t|\theta^{t-1}, \dots, \theta^1) = p(\theta^t|\theta^{t-1})$

Let X be a r.v. (multidimensional) Markov chain $X^{(1)} \to X^{(2)} \to \dots$

with transition density $\psi(y|x) \equiv p_{X^{(t+1)}|X^{(t)}}(y|x)$

A equilibrium distri for a Markov chain is a density $\pi(x)$, for which $\pi(y) = \int \psi(y|x)\pi(x)dx$

State $X^{(t)}$ $p_{X^{(t)}}(x) = ?$, $p_{X^{(t+1)}}(y) = ?$

$p_{X^{(t+1)}}(y) = \int p_{X^{(t+1)}, X^{(t)}}(y, x)\,dx = \int p_{X^{(t+1)}|X^{(t)}}(y|x)\, p_{X^{(t)}}(x)dx$

$= \int \psi(y|x) p_{X^{(t)}}(x)\,dx$

$p_{X^{(t+1)}}(\cdot) = p_{X^{(t)}}(\cdot) = \pi(\cdot)$

is $\pi(x)$ is our equilibrium distri & is our target density

(we want Σ posterior) & if $X^{(1)}$ has marginal $\pi(x)$, then $X^{(2)}, \dots, X^{(N)}$ all have distri $\pi(\cdot)$

Need a $\psi(y|x)$ st Σ posterior is $\pi(x)$ (equilibrium distri). & $X^{(1)}$ from $\pi(\cdot)$. Under weak conditions on $\psi(y|x)$: $P(X^{(t)}|X^{(1)}=x) \xrightarrow{t\to\infty} \pi(x)$ ∴ MCMC is about-

Choosing $\psi(y|x)$ st Σ posterior is $\pi(x)$.

Choosing $\psi(y|x)$ st ~~$X^{(t)}$~~ $P_{x^{(t)}|x^{(1)}}$ converges to $\pi(x)$ (quickly)

Checking convergence has happened.

\Def 3.3/ Let $\psi(y|x) = P_{x^{(t)}|x^{(t-1)}}(y|x)$ be Σ transition density for a Markov chain. $\psi(\cdot|\cdot)$ is said to satisfy "detailed balance" for a distri $P(\cdot)$ if $\psi(y|x)P(x) = \psi(x|y)P(y)$

\Thm 3.1/ if $\psi(\cdot|\cdot)$ satisfies detailed balance for $P(\cdot)$, then $P(\cdot)$ is an equilibrium distri for Σ Markov chain.

\proof/ need to show $\int \psi(y|x)P(x)dx = P(y)$. Given detailed balance $\int \psi(y|x)P(x)dx = \int \psi(x|y)P(y)dx = P(y)\int \psi(x|y)dx = P(y)\cdot 1 = P(y)$ □

if an MCMC algorithm ($\sim \psi(\cdot|\cdot)$) satisfies detailed balance for Σ target (posterior distri) then Σ Markov chain is a sequence of correlated samples from Σ target.

\Gibbs Sampler/ $\theta = (\theta_1, \dots, \theta_d)$, data y. $\pi(\theta|y)$

$\theta^{t-1} = (\theta_1^{t-1}, \dots, \theta_d^{t-1})$ ($\psi(y|x)$)

\Algorithm/ For $j = 1, \dots, d$ Sample θ_j^t from $\pi(\theta_j|\theta_{-j}^{t-1}, y)$

$\theta_{-j}^{t-1} := (\theta_1^t, \dots, \theta_{j-1}^t, \theta_{j+1}^{t-1}, \dots, \theta_d^{t-1})$

$\theta_1^{(t)}$ from $\pi(\theta_1|\theta_2^{t-1}, \dots, \theta_d^{t-1}, y)$

θ_2^t from $\pi(\theta_2|\theta_1^t, \theta_3^{t-1}, \dots, \theta_d^{t-1}, y)$

θ_3^t from $\pi(\theta_3|\theta_1^t, \theta_2^t, \theta_4^t, \dots, \theta_d^{t-1}$

$(\theta^t|\theta^{t-1})$ $\pi(\theta|y)$ all d 'full conditionals'

Ex/ $\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} = \Sigma\right)$

$\pi(\theta) \propto 1$

$\theta | y \propto \exp\left\{-\frac{1}{2}\left(\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} - \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}\right)^T \Sigma^{-1} \left(\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} - \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}\right)\right\}$

$\theta | y \sim N\left(\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$

$\pi(\theta_1 | \theta_2, y) \quad \pi(\theta_2 | \theta_1, y) \quad \theta_1 | \theta_2, y \sim N(y_1 + \rho(\theta_2 - y_2), 1 - \rho^2)$

$\theta_2 | \theta_1, y \sim N(y_2 + \rho(\theta_1 - y_1), 1 - \rho^2)$

Gibbs: Start $(\theta_1^0, \theta_2^0) \quad (0, 0)$.

$\theta \sim \theta^1, \theta^2, \theta^3, \ldots \theta^n$ via 2 Gibbs sampler

$(\theta_1^1 | \theta_2^0, y \to \theta_2^1 | \theta_1^1, y \to \theta_1^2 | \theta_2^1, y \ldots)$

>> # Gibbs Sampler Code #

Thm 3.2/ Each step of 2 Gibbs Sampler satisfies detailed balance (for 2 posterior).

Proof/ Use notation $\psi(y|x) \quad (y \to \theta^t \quad x \to \theta^{t-1})$

Updating a single elem (single Gibbs step) has transition probab $\psi(y|x) = P_j(y_j | x_{-j}) \prod_{i \neq j} \delta(y_i - x_i)$

$\delta(\cdot)$ dirac delta, P_j is j^{th} full conditional

$(\pi(\theta_j^t) \theta_{-j}^{t-1}, y)) \quad \psi(y|x) P(x) = P_j(y_j | x_j) \prod_{i \neq j} \delta(y_i - x_i) P(x) =$

$(P(x_1, \ldots, x_{j-1}, y_j, x_{j+1}, \ldots, x_d) / \int P(x) dx_j) \prod_{i \neq j} \delta(y_i - x_i) P(x)$

$= (P(y_1, \ldots, y_d) / \int P(y) dy_j) \prod_{j \neq i} \delta(y_i - x_i) P(y_1, \ldots, y_{j-1}, x_j, y_{j+1}, \ldots, y_d)$

$= P(y) \psi(x|y) \quad \square$

NHM (normal hierarchical model) $y_{ij} | \theta_j, \sigma^2, \mu, \tau^2 \sim (\theta_j, \sigma^2) \quad i = 1, \ldots, n_j \quad j = 1, \ldots, J$

$\theta_j | \mu, \tau^2 \sim N(\mu, \tau^2) \quad \pi(\mu | \tau^2) \propto 1 \quad \pi(\sigma^2) \propto \frac{1}{\sigma^2} \quad \pi(\tau^2) \propto \frac{1}{\tau}$

$\forall j \quad \theta_j | \sigma^2, \mu, \tau^2, \theta_{-j}, y \sim N(\hat{\theta}_j, \hat{\sigma}_j^2)$

$\hat{\theta}_j = \left(\frac{\mu}{\tau^2} + \frac{n_j}{\sigma^2} \bar{y}_{\cdot j}\right) / \left(\frac{1}{\tau^2} + \frac{n_j}{\sigma^2}\right) \quad , \hat{\sigma}_j^2 = 1 / \left(\frac{1}{\tau^2} + \frac{n_j}{\sigma^2}\right)$

$\bar{y}_{\cdot j} = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij} \qquad \mu | \tau^2, \sigma^2, \bar{\theta}, y \sim N(\bar{\theta}, \tau^2 / J)$

$\pi(\mu | \tau^2, \sigma^2, \theta, y) \propto \pi(\mu | \tau^2, \sigma^2, \theta) P(y | \mu, \tau^2, \theta, \sigma^2)$

$\propto \pi(\mu | \tau^2, \sigma^2) \prod_{j=1}^{J} \pi(\theta | \mu, \tau^2, \sigma^2) \quad \therefore \psi$ has scaled inverse Chi-

squared distn with params ν & s^2 $\quad \sigma^2 \sim \text{Inv-}\chi^2(\nu, s^2)$ is:

$$p(\sigma^2 \mid \nu, s^2) = \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} s^{\nu} (\sigma^2)^{-(\nu/2+1)} e^{-\nu s^2/2\sigma^2}$$

$\sigma^2 \mid \mu, \tau^2, \theta, y \sim \text{Inv-}\chi^2(n, \hat{\sigma}^2)$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^{J} \sum_{i=1}^{n_j} (y_{ij} - \theta_j)^2 \qquad \tau^2 \mid \mu, \sigma^2, \theta, y \sim \text{Inv-}\chi^2(J-1, \hat{\tau}^2)$$

$$\hat{\tau}^2 = \frac{1}{J-1} \sum_{j=1}^{J} (\theta_j - \mu)^2$$

Metropolis-Hastings (i) Draw a starting pt θ^0 s.t. $\pi(\theta^0 \mid y) > 0$, from (any) distn $p_0(\theta)$

(ii) At $t = 1, 2, \ldots$ Suppose θ^{t-1} is known: sample θ^* from a proposal distn $q(\theta^* \mid \theta^{t-1})$

(iii) Compute 2 Metropolis ratio $r(\theta^{t-1}, \theta^*) = \dfrac{\pi(\theta^* \mid y) / q(\theta^* \mid \theta^{t-1})}{\pi(\theta^{t-1} \mid y) / q(\theta^{t-1} \mid \theta^*)}$

(iv) Set $\theta^t = \begin{cases} \theta^* & \text{prob } p^* = \min(1, r) \\ \theta^{t-1} & \text{otherwise} \end{cases}$

generate $u \sim \text{Unif}(0,1)$ Set θ^* if $u \le p^*$

$$\pi(\theta \mid y) \propto \pi(\theta) p(y \mid \theta) = \pi(\theta) p(y \mid \theta)$$

$$= \frac{\pi(\theta) p(y \mid \theta)}{\int p(y \mid \theta) \pi(\theta) d\theta = p(y)} \qquad \left\{ \int p(y \mid \theta) \pi(\theta) d\theta = p(y) \right\}$$

$$\frac{\pi(\theta^* \mid y)}{\pi(\theta^{t-1} \mid y)} = \frac{\pi(\theta^*) p(y \mid \theta^*) / p(y)}{\pi(\theta^{t-1}) p(y \mid \theta^{t-1}) / p(y)} = \frac{\pi(\theta^*) p(y \mid \theta^*)}{\pi(\theta^{t-1}) p(y \mid \theta^{t-1})} \qquad \left\{ \therefore \text{only need } \pi(\theta) p(y \mid \theta) \right\}$$

Comment $q(\cdot \mid \cdot)$ is symmetric $q(\theta^* \mid \theta^{t-1}) = q(\theta^{t-1} \mid \theta^*)$

$r = \dfrac{\pi(\theta^* \mid y)}{\pi(\theta^{t-1} \mid y)}$ (Metropolis Algorithm)

$\theta^* = \theta^{t-1} + D \qquad D \sim MVN(0, \Sigma)$ ← choose

Random walk proposal

Random walk metropolis

$\pi(\theta|y)$ $\underline{\theta}^{(1)} \to \underline{\theta}^{(2)} \to \ldots \underline{\theta}^{(N)}$ has my MCMC converged? $\hat{R}$

$\Psi = g(\theta|y)$ $E(\theta_i)$ $\hat{R}(\theta_i)$ near 1 demonstrates convergence

$\Psi = g(\theta|y)$ ∴ Do I have enough samples?

$E[\Psi] = \frac{1}{N}\sum_{i=1}^{N}\Psi_i$ Monte Carlo Ψ_i from $\underline{\theta}^{(i)}$ (Discarding burn in samples)

$\frac{1}{\sqrt{N}}Se(\Psi)$ is wrong! Effective Sample Size $Ness$. is st

$MCerror = \frac{1}{\sqrt{Ness}}Sd(\Psi)$

$Ness$ estimated by our software from your $MCMC \leq N$ ($<<N$)

{sometimes not less than N ∴ $MCMC << N$ sometimes}

good MCMC has $Ness$ close to N (but not N)

{use hamiltonian Monte Carlo for CW?}

Generate θ^* from $q(\theta^*|\theta^{t-1})$ acceptance rate (based on average r) is too low, chain gets stuck.

$q(\theta^*|\theta^{t-1})$ '$\sim$' $N(\theta^{t-1}, \Sigma)$

idea is 'tune' Σ (automatically) st acceptance rates are 'good' (in 1 dimensional param θ want ≈ 0.44, in high dimensions ≈ 0.23)

Adaptive MCMC: (1) set $\Sigma^* = \Sigma_0$, $k = 1$

(2) run Metropolis Hastings for M steps with proposal $q(\theta^*|\theta^{t-1}) = N(\theta^{t-1}, k\Sigma^*)$

(3) set $\Sigma^* = Var[\theta^1, \ldots, \theta^m]$ & choose k st % acceptance rate is close to 'optimal'

(4) Repeat 2 & 3 a fixed number of times until % acceptance rate is optimal. Check $\Sigma \approx \Sigma^*$ has converged.

(5) Run standard N step M-H with $q(\theta^*|\theta^{t-1}) = N(\theta^{t-1}, k\Sigma^*)$

HMC (Hamiltonian Monte Carlo) is 'Auxiliary variable MCMC'

Desire 'Auxiliary variables', ϕ st $p(\theta, \phi | y) = p(\theta|y)p(\phi)$

target becomes $p(\theta, \phi, y)$ keeping only θ

Samples gives samples from $\pi(\theta|y)$

Require ϕ_j for each θ, $p(\phi)$ is called 'Momentum'

we specify: 'Mass' matrix M ($d \times d$)

- Scalar ε
- integer number of leapfrog steps, L

(1) Draw ϕ^{t-1} from $\phi \sim N(0, M)$ set $\phi^* = \phi^{t-1}$, $\theta^* = \theta^{t-1}$

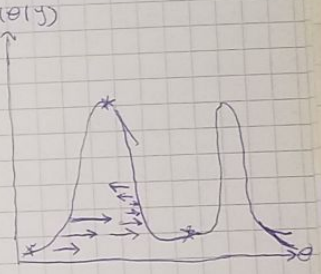
(2) Repeat L leapfrog steps: (a) $\phi^* = \phi^* + \frac{1}{2}\varepsilon \frac{d \log \pi(\theta^*|y)}{d\theta^*}$

(b) Update θ position vec using momentum $\theta^* = \theta^* + \varepsilon M^{-1} \phi^*$

(c) $\phi^* = \phi^* + \frac{1}{2}\varepsilon \frac{d \log \pi(\theta^*|y)}{d\theta^*}$

(3) Metropolis ratio $r = \frac{\pi(\theta^*|y)p(\phi^*)}{\pi(\theta^{t-1}|y)p(\phi^{t-1})}$

(4) $\theta^t = \begin{cases} \theta^* & \text{probab } \min(1, r) \\ \theta^{t-1} & \text{otherwise} \end{cases}$



Intuition: • is θ^* in flat area of $\pi(\theta|y)$, $\frac{d\pi(\theta|y)}{d\theta} = 0$ leapfrog steps 'skate' along θ space with const velocity.

• Moving in direction of decreasing posterior, $\frac{d \log \pi(\theta^*|y)}{d\theta^*} < 0$, take smaller & smaller steps until change direction

• Move in direction of increasing posterior $\pi(\theta^*|y)' > 0$

Momentum increases, take bigger steps

```
>> options(mc.cores = parallel::detectCores())
>> murders      >> library(brms)   library(bayesplot)   library(reshape)
>> ggplot
>> fit1 <- brm(murders ~ low_income + unemployed + pop + I(pop^2), data = murders)
```

$y_i | \beta, \sigma^2 \sim N(\beta_0 + \beta_1 x_{1i} + \cdots, \sigma^2)$ $\pi(\beta, \sigma^2)$ $p(y|\text{data})$

$p(y | \beta = \hat{\beta}, \sigma = \hat{\sigma})$

```
>> summary(fit1)   # rhat is not 1 but biggest #  look at effective
```

sample size (ESS) for each variable

```
>> plot(fit1)      ?brm      >> brm(..., inits = "0")
```

can run for longer

$y_i | \underline{x}_i, \beta_i, \sigma^2) \sim N(\beta_0 + \beta_{pop} pop + \beta_{low} low + \beta_{unem} Unem, \sigma^2)$

$y \in [0, 40]$ $low = O(10)$ $\therefore \beta_{low} \sim O(1)$

$Unem = O(1)$, $\beta_{unem} \sim O(10)$

$\beta_{low} \in [-2, 2]$ $\beta_{unem} \in [-2, 2]$

$pop = C(10^6)$ $\therefore \beta_{pop} \sim O(10^{-5})$ $N(0, 2 \times 10^{-5})$

pop_prior <- set_prior("normal(0, 0.00002)", class = "b", coef = "pop")

~~fit~~ >> fit2 <- brm(..., prior = pop_prior) >> summary(fit2) >> plot(fit2)

>> Murders$pop <- Murders$pop / 10^6

>> pop_prior <- set_prior("normal(0, 10)", class = "b", coef = "pop")

rhats are 1 & trace plots look good

>> preds <- predict(fit2, newdata = murders) >> preds

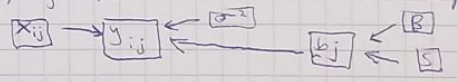
>> posterior <- samples(fit2)

>> samples <- posterior_samples(fit2)

>> sum(samples$ b_pop < 0) / 4000 ⇒ 0.127 #probab#

Menti /1/ $y_{ij} | b_j, \sigma \sim N(b_j^T X, \sigma^2)$, $b_j \sim N(B, S^2)$ & a valid prior comprises what type of model? Bayesian hierarchical regression model

/2/ $y_{ij} | b_j, \sigma \sim N(b_j^T X, \sigma^2)$, $b_j \sim N(B, S^2)$ what about 2 prior distri be? $p(B, S, \sigma)$



/3/ $y_{ij} | b_j, \sigma \sim N(b_j^T X, \sigma^2)$, $b_j \sim N(B, S^2)$, $\pi(\sigma, B, S)$ what is 2 Model & b_j called? process layer

/4/ ... $\pi(\sigma, B, S)$ what call 2 variables in 2 data identified by j?

Grouping variables

/5/ what need to be aware of fitting $b_j \sim N(B, S^2)$, $\pi(\sigma, B, S)$ with brms? ~~B~~ b_j must be distri $N(0, var)$ process layer hastes distri with mean 0 $\hat{B}_j \sim N(0, \Sigma)$ $N(BX + \beta_j X)$ $B_j = B + \hat{\beta}_j$

/6/ $y_{ij} | b_j, \varnothing \sim D(g^{-1}(\eta_{ij}), \varnothing)$, $\eta_{ij} = b_j^T X_{ij}$, $b_j \sim N(B, S^2)$, $\pi(\varnothing, B, S)$ what

is η? ≥ linear predictor

\2/ what is D ? ≥ family

\3/ what is $g^{-1}(\eta_{ij})$? ≥ mean of D

\4/ why do we need a link? g is link.: to make sure ≥ mean of D is within ≥ support of ≥ distri

make sure ≥ restrictions on ≥ mean of D are met

>> View (verbagg) >> dim(verbagg)

(sit2) >> verbsit1 ← brm(r2 ~ Anger + Gender + btype + situ + (1|id) + (1|item), data = verbagg.tr, family = bernoulli())

(· | ·) ~~[illegible]~~ (predictors to group | Group variable)

$r_2 | \beta \ldots \sim Ber(\phi(\beta, x))$ $p := g^{-1}$ $g(\eta) = logit(\eta) = \frac{1}{1+e^{-\eta}}$

$logit(\beta, x) = \beta_0 + \beta_1 x_1 + \cdots + = b_0 + b_1 x_1 + \cdots + b_p x_p + \beta_{0j} + \beta_{1j} x_{ij} + \ldots + \beta_{pj} x_{pj}$

$\beta_j \sim N(0, \Sigma)$ $\pi(b, \Sigma)$

$\beta_{0,id} \sim N(0, \Sigma_{id})$ $\beta_{item} \sim N(0, \Sigma_{item})$

x_1 = Anger has Coefs of b_1 x_2 Gender Coefs b_2

x_3 btype coefs b_3 x_4 situ Coefs b_4

>> verb.sit1 ← brm(... + (Anger | item), ..., ...)

#or# >> brm(... + (0 + Anger | item), ..., ...)

>> plot(verb.sit1, ask = FALSE) >> summary(verb.sit1)

>> preds ← predict(verb.sit1, newdata = verbagg.pred)

~~>> class~~ >> a_classifier ← preds[, "Estimate"] > 0.5

>> Conmat ← ConfusionMatrix(..., ...)

>> Conmat >> sum(diag(Conmat))/sum(Conmat)

>> Tomsit ← brm(TC. Co. Star. Age. Difference ~ (1|TC.age), data = Tom(r)

>> Summary(Tomsit) >> plot(Tomsit)

>> mu.prior ← set_prior("normal(0, 10)", class = "Intercept")

>> Tomsit1 ← brm(..., prior = mu.prior) >> summary(Tomsit1) >> plot(Tomsit1)

>> tau.prior ←

We tried to improve the first model by factoring in recent form, that is weighting the average goals in the previous 5 games more heavily than less recent results. We calculated recent form as a teams average goals from the last 5 games, ignoring whether they were played at home or away.
We tried weighting form at 25, 50 and 75. However all 3 attempts had worse results than the orginal model (where W=0).
We believe this is due to the fact that ignoring whether the team was playing home or away when calculting recent form is adding more noise than accuracy to the predictions.

We next tried calculating recent form seperately for home and away games, using a similar formula as before, but considering only home or away games
We again tested for 25, 50 and 75% weighting of recent form and found that 25% gave the best predictions, and was an improvement on the first model.
Calculating form separately for home and away removes the noise induced when we calculated it as a single variable.

We tried again to improve our model further by including conceded goals by the opponent to our 3rd model.
We attempted this by adding goals conceded by the away team into the home teams expected goal number, and vice versa.
This would then mean our model took into account the opponent the teams were facing. We added in two factors for conceded goals, conceded average goals and conceded form, both relative to home and away matches

As previously stated, we added the conceded goals factor to our previous model, leading to our expected goal numbers calculated like so. We weighted all 4 factors equally, and also 12.5% weighting on form's as that worked well with Model 3. The results showed that although each weighting ~~was~~ and adding this factor did not change things as much

$y_i|\beta,\sigma^2 \sim N(\beta x, \sigma^2)$ $y_i|\beta,\phi \sim D(g^{-1}(\eta), \phi)$ $\eta = \beta^T x$ $\pi(\phi,\beta)$

$\pi(\beta,\sigma^2)$ $\pi(\alpha,\beta)$ brm(y~..., data)

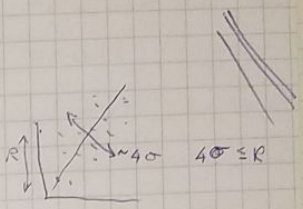
brm(bf(y~..., sigma~...), data)

formula sheet only need normals table

must remember normal formula

$N(\mu,\beta x,\sigma^2) \Rightarrow y = \beta x + \varepsilon$ $\varepsilon \sim N(0,\sigma^2)$ $\therefore y - \beta x$

$var(y) = \sigma^2$ R is range of data $\therefore \sigma \leq \frac{R}{4}$



let $\sigma \sim N(0,(\frac{R}{12})^2)$

rank years prior $\sim N(1100, 500)$

degree year prior $\sim N(750, 400)$

$\leq \sim N(\beta_0 + \beta_1 \cdot rank + \beta_2 \cdot rank.years + \beta_3 degree + \beta_4 degree.years +$

$b_{0j} + b_{1j} Rank_j + \ldots b_{4j} degree.years, \sigma^2)$ $b_j \sim N(0, \Sigma$

① give Monte Carlo est + error from + param

② give MC est + errors "Some prediction

③ give MC... or Pr(parameter thing prediction thing < 97.5)

Est. Error @ 2.5

~~E~~ $E[\text{param}_{\text{samples}}]$ $Sd(\text{param}_{\text{samples}})$ $P(\text{Params} \leq 0.025) \leftarrow 0.975$

$E[g(\theta)|y] = \int g(\theta)\pi(\theta|y)d\theta \rightarrow \theta_1, \ldots, \theta_N$ from $\pi(\theta|y)$

$\approx \frac{1}{N}\sum_{i=1}^{k} g(\theta_i)$ $E[g(y)|y] = \int\int g(y)p(y|\theta)\pi(\theta|y)d\theta = E[g(y(x))|y] =$

$\int\int g(y(x))P(y(x)|\theta)\pi(\theta|y)d\theta$ new data $= \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ $\theta_1, \ldots, \theta_n \sim \pi(\theta|y)$

$y_1(x_1)|\theta_1, \ldots, y_n(x_1)|\theta_n$

$y_1(x_n)|\theta_1, \ldots, y_n(x_n)|\theta_n$ Estimated $E(y(x))$

$E(y(x_{woman,pros}) - y(x_{woman,pros}))$

① $P(\theta > 0|y)$ or ② $P(y_1 > y_2|y)$ ①: $g(\theta) = \mathbb{1}(\theta > 0)$

$P(\theta > 0|y) = E[\mathbb{1}(\theta > 0)] = \int \mathbb{1}(\theta > 0)\pi(\theta|y)d\theta \approx \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}(\theta_i > 0)$

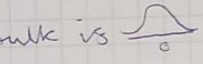
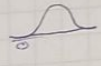
②: $P(y_1 > y_2|y)$ $\therefore$ $P(y(x_1) > y(x_2)|y) = \int\int \mathbb{1}(y_1 > y_2)P(y(\theta)\pi(\theta|y)d\theta dy$

$\frac{1}{N}\sum \mathbb{1}(y_1 > y_n)$

$Sd(MC\ esti) = \frac{sd(g(\theta_i))}{\sqrt{Ness}}$ BulkESS >> TailESS

for $E[param]$ use bulkESS

if want $P(param > 97.5)$ use tailESS

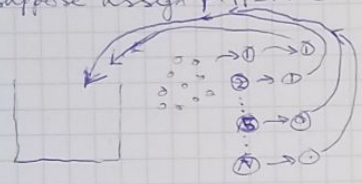
if want $P(\theta > 0)$ use bulk is  & tail is 

use smallest bulk ess

let $P(\theta > 0 | y) = \hat{p}$ ∴ MC $MC = \sqrt{\hat{p}(1-\hat{p})/Ness}$ = MC error

Week3 proof of Z representation theorem / $Y_1, Y_2, \ldots$ exchangeable binary RQ's ∴ $Y_i = 1$ is heads, 0 is tails

Suppose assign $P(Y_i = 1) = \theta$ exchangeability $\Rightarrow P(Y_i = 1) = \theta$ $i = 1, \ldots, N$



let Z be rvl as a random counter in Z bucket $P(Z=1) = \theta$ (by exchangeability)

say θ_k is Z proportion of 1's in Z bucket $P(Z=1|\theta_k) = \theta_k$ (don't know θ_k) possible vals are $\theta_i = i/N$ $i = 0, 1, \ldots, N$ Suppose set $P(\theta_k = \theta_i) = p_i$ LOTP $\Rightarrow P(Z=1) = \sum_{i=0}^{N} P(Z=1|\theta_k) P(\theta_k = \theta_i) = \sum_{i=0}^{N} \theta_i p_i =$ $\int_0^1 \theta_k \, dF_N(\theta_k)$ F_N is cdf giving probab p_i to $\theta_k = i/N$

Suppose n tosses, let X_n be number of H in first n tosses.

Exchangeability $\Rightarrow P(X_n = k)$ is Z same for observing k heads in any sample of n tosses $P(X_n = k)$ must be Z same as our probab for drawing k counters labelled 1 from our bucket when we take n out at random still have $F_N(\theta)$ giving probab p_i to i/N 1's in Z bucket. is we knew θ: $P(X_n = k|\theta) =$ ~~$\binom{N\theta}{k}\binom{N(1-\theta)}{n-k} / \binom{N}{n}$~~ $\left[\binom{N\theta}{k}\binom{N(1-\theta)}{n-k}\right] / \binom{N}{n}$ LOTP $\Rightarrow P(X_n = k) = \int_0^1 \frac{\binom{N\theta}{k}\binom{N(1-\theta)}{n-k}}{\binom{N}{n}} dF_N(\theta)$

when N gets large sampling with/without replacement become Z same as $N \to \infty$ $P(X_n = k) = \int_0^1 \binom{n}{k} \theta^k (1-\theta)^{n-1} \, dF_n(\theta)$

Helly's thm $\Rightarrow$ any sequence F_n such as this (i/N) converges to a unique limit F $P(X_n = k) = \int_0^1 \binom{n}{k} \theta^k (1-\theta)^{n-k} \, dF(\theta)$

$= \int_0^1 \binom{n}{k} \theta^k (1-\theta)^{n-k} f(\theta) d\theta$

Week 7 Full Conditional for μ / $\pi(\mu|\tau^2,\sigma^2,\underline{\theta},\underline{y}) \propto$

$\pi(\mu|\tau^2,\sigma^2,\theta)p(\underline{y}|\mu,\tau^2,\sigma^2,\theta) \propto \pi(\mu,\tau^2,\sigma^2,\underline{\theta}) \propto$

$\pi(\mu|\tau^2,\sigma^2)\pi(\theta|\mu,\tau^2,\sigma^2) \propto \prod_{j=1}^{J} \exp\left\{-\frac{1}{2\tau^2}(\theta_j-\mu)^2\right\} \propto$

$\exp\left\{-\frac{1}{2\tau^2}\sum_{j=1}^{J}(\theta_j-\mu)^2\right\} \propto \exp\left\{-\frac{1}{2\tau^2}\sum_{j=1}^{J}(\mu^2-2\theta_j\mu)\right\} \propto$

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

$\mu|\tau^2,\sigma^2,\theta,\underline{y} \sim N(\bar{\theta}, \tau^2/J)$

Full conditional for σ^2

Def 4.3 A random quantity ψ, has scaled inverse chi-squared distri with params ν, s^2 $\therefore \psi|\nu,s^2 \sim \text{Inv}\chi^2(\nu,s^2)$ iff

$p(\psi|\nu,s^2) = \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} s^{\nu}\psi^{-(\nu/2+1)} e^{-\frac{\nu s^2}{2\psi}}$

$\pi(\sigma^2|\underline{\theta},\mu,\tau^2,\underline{y}) \propto \pi(\sigma^2|\theta,\mu,\tau^2)\,p(\underline{y}|\theta,\mu,\tau^2,\sigma^2) \propto$

$\frac{1}{\sigma^2}\prod_{j=1}^{J}\prod_{i=1}^{n_j}\exp\left\{-\frac{1}{2\sigma^2}(y_{ij}-\theta_j)^2\right\}\frac{1}{\sigma^2} \propto \frac{1}{\sigma^2}\prod_{j=1}^{J}\frac{1}{\sigma^{n_j}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n_j}(y_{ij}-\theta_j)^2\right\} \propto$

$\frac{1}{\sigma^2}\frac{1}{\sigma^{\sum_{j=1}^{J}n_j}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij}-\theta_j)^2\right\}$ $\quad n=\sum_{j=1}^{J}n_j$

$\pi(\sigma^2|\mu,\tau^2,\theta,\underline{y}) \propto (\sigma^2)^{(-1-\frac{n}{2})} e^{-\frac{n\hat{\sigma}^2}{2\sigma^2}}$, $\quad \hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij}-\theta_j)^2$

$\sigma^2|\mu,\tau^2,\theta,\underline{y} \sim \text{Inv}\chi^2(n,\hat{\sigma}^2)$

Full conditional for τ / $\pi(\tau^2|\mu,\sigma^2,\theta,\underline{y} \propto$ {by Bayes}

$\pi(\tau^2|\mu,\sigma^2,\theta)\,p(\underline{y}|\sigma^2,\mu,\theta)$ $\therefore$ {$y \sim N(\mu,\sigma^2)$ $\therefore$ indep of τ $\therefore$}

$\propto \pi(\tau^2|\mu,\sigma^2)\,\pi(\theta|\tau^2,\mu,\sigma^2) \propto \frac{1}{\tau}\prod_{j=1}^{J}\frac{1}{\tau}\exp\left\{-\frac{1}{2\tau^2}(\theta_j-\mu)^2\right\}$

$\propto \frac{1}{\tau^{J+1}}\exp\left\{-\frac{1}{2\tau^2}\sum_{j=1}^{J}(\theta_j-\mu)^2\right\} \propto (\tau^2)^{-(\frac{J}{2}+\frac{2}{2})}\exp\left\{-\frac{1}{2\tau^2}\sum_{j=1}^{J}(\theta_j-\mu)^2\right\}$

$\tau^2|\theta,\mu,\sigma^2,\underline{y} \sim \text{Inv}\chi^2(J-1,\hat{\tau}^2)$, $\hat{\tau}^2 = \frac{1}{J-1}\sum_{j=1}^{J}(\theta_j-\mu)^2$

Metropolis-Hastings Satisfies detailed balance / Transition density $\psi(y|x)$ $(\psi(\theta^t|\theta^{t-1})$ θ^* from $q(\theta^*|\theta^{t-1})$

$\psi(y|x)\pi(x) \overset{\text{D.B detailed balance}}{=} \psi(x|y)\pi(y)$. {accept prob $r^* = \min(1,r)$}

$\psi(y|x) = p_{\theta^t|\theta^{t-1}}(y|x) = \int p(y,\theta^*|x)\,d\theta^* = \int p(y|\theta^*,x)\,q(\theta^*|x)\,d\theta^*$

$= \int(r^*\delta(y-\theta^*)+(1-r^*)\delta(y-x))\,q(\theta^*|x)\,d\theta^*$ $\quad r^* := r(x,\theta^*) = \begin{cases} 1 & \theta^* \in D_1(x) \\ r(x,\theta^*) & \theta^* \in D_0(x) \end{cases}$

with $D_1(x) = \{\theta^* : r(x,\theta^*) > 1\}$ $\quad D_0(x) = \{\theta^* : r(x,\theta^*) \le 1\}$

to satisfy detailed balance $\psi(y|x)\pi(x) = \psi(x|y)\pi(y)$ ← posterior

$\psi(y|x)\pi(x) = \int(r^*\delta(y-\theta^*)+(1-r^*)\delta(y-x))\,q(\theta^*|x)\,\pi(x)\,d\theta^* =$

$$\int_{D_0(x)} \big(r(x,\theta^*)\delta(y-\theta^*)+(1-r(x,\theta^*))\delta(y-x)\big) q_1(\theta^*|x)\pi(x)\,d\theta^* +$$

$$\int_{D_1(x)} \delta(y-\theta^*) q_1(\theta^*|x)\pi(x)\,d\theta^* =$$

$$\int_{D_0(x)} r(x,\theta^*)\delta(y-\theta^*) q_1(\theta^*|x)\pi(x)\,d\theta^* + \int_{D_0(x)} (1-r(x,\theta^*))\,\delta(y-x)\, q_1(\theta^*|x)\pi(x)\,d\theta^* +$$

$$q_1(y|x)\pi(x)\,\mathbb{1}(y\in D_1(x)) =$$

$$r(x,y) q_1(y|x)\pi(x)\,\mathbb{1}(y\in D_0(x)) + \delta(y-x)\Big[\int_{D_0(x)} q_1(\theta^*|x)\pi(x)\,d\theta^* - \int_{D_0(x)} r(x,\theta^*) q_1(\theta^*|x)\pi(\theta)\,d\theta^*\Big] + q_1(y|x)\pi(x)\,\mathbb{1}(y\in D_1(x))$$

$$\Big\{r(x,y) = \frac{q_1(x|y)\pi(y)}{q_1(y|x)\pi(x)}\Big\}$$ for

$$= \psi(y|x)\pi(x) = q_1(x|y)\pi(y)\,\mathbb{1}(y\in D_0(x)) \quad ①$$

$$\delta(y-x)\Big[\int_{D_0(x)} q_1(\theta^*|x)\pi(x)\,d\theta^* - \int_{D_0(x)} \frac{q_1(x|\theta^*)\pi(\theta^*)}{q_1(\theta^*|x)\pi(x)} q_1(\theta^*|x)\pi(x)\,d\theta^*\Big] +$$

$$q_1(y|x)\pi(x)\,\mathbb{1}(y\in D_1(x)) \quad ③$$

$$\Big\{\vdots = \psi(x|y)\pi(y)\Big\}$$ 2nd term

clearly symmetric in x,y.

$\psi(x|y)\pi(y)$: ① $q_1(y|x)\pi(x)\,\mathbb{1}(x\in D_0(y))$

③ $q_1(x|y)\pi(y)\,\mathbb{1}(x\in D_1(y))$

$y\in D_1(x) \iff r(x,y)>1 \iff \frac{1}{r(y,x)}>1$

$\iff r(y,x)\le 1 \quad x\in D_0(y)$ $\qquad y\in D_0(x) \iff x\in D_1(y)$

$\therefore \psi(x|y)\pi(y) = \psi(y|x)\pi(x)$ ☐ Metropolis Hastings has Σ equalibrium distri $\Leftrightarrow$ Σ posterior distri

/Week 8/ \Stan for Σ NHM/ $y_{ij}|\theta_j,\sigma^2 \sim N(\theta_j,\sigma^2)$ $j=1,\dots,J$

$\theta_j|\mu,\tau^2 \sim N(\mu,\tau^2)$ $\pi(\mu,\tau^2,\sigma^2)$

(x_i,y_i)  $y_i \sim N(a+bx_i,\sigma^2)$ $y_i|x_i,a,b,\sigma^2 \sim N(a+bx_i,\sigma^2)$

default priors uniform priors $\pi(a,b)\propto 1$ $\pi(\sigma^2)\propto \frac{1}{\sigma^2}$ but

with Stan: $a\sim?$ $b\sim?$ $\sigma^2\sim?$ what ever we want

instead: $y_i|x_i,\underline{\beta}_i,\sigma^2 \sim N(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\dots+\beta_p x_{ip},\sigma^2)$ is baysesian likelihood $\therefore a\sim\beta_0$ $b\sim\beta_1$ $\therefore \underline{\beta}\sim?$ & $\sigma^2\sim?$

>> rstan;;extract(...) /.do rstan;;extract to tell R which package extract lives in since it doesn't know which to choose/.

DATA layer $y_{ij} | \theta_j, \sigma^2, \mu, \tau^2 \sim N(\theta_j, \sigma^2)$
anything anything what ever params need in this distri

process layer in brms must always be $\theta_j \sim N(0, \tau^2)$, $\pi(\tau^2, \sigma^2, \text{param in data } \mu)$
i.e. $\theta_j \sim N(0, \tau^2)$ with prior $\pi(\tau^2, \sigma^2$, any params in Z mean of Z data)
∴ Model: $y_{ij} \sim N(\mu + \theta_j, \sigma^2)$ $\theta_j \sim N(0, \tau^2)$ $\pi(\mu, \tau^2, \sigma^2)$

any thing after an b is one of these population params, so b are params in Z data layer mean only, sigma are Z params in Z data layer that are not in Z mean, Sd's are standard diviation of Z group effects which have mean zero

>> plot(rats_z) >> plot(rats_z, pars="diet") %. Specify Z params by Z group, get plot of Z different group effects

>> DietEffects ← ranef(rats_z, summary=TRUE) %. gives mean estis of each diet

>> mains ← summary(rats_z) >> mains $ fixed

>> mains $ fixed ["Estimate"] + DietEffects $ diet[, "Estimate", 1] %. gives Z means had before and can do similar with samples %.

can set for Z params of Z data layer not in Z mean in this case σ (so μ & τ are in Z mean?) ∴ set a prior

choosing Class Sigma & it knows in sitting a gausian model, it auto limit it so not allowed less than zero eg choose N(0,100) it's only half normal since truncated at 0

for (x.variable | group.variable) to Z right, put variable where have groups, so variable diet, different level & like to fit our model, y func, differently for each of those groups - invokes Z hierarchical model. Z params in x.variable, like intercept params each get Z normal distri for each group with mean zero & some common variance, common var has a prior can set via 1) Z "sd" class & there becomes a global mean term for x.variable, that's estied & takes role of our μ in Z normal hierarchical model

$y_i \mid \quad \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

$y_i \mid \beta, \sigma^2 \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ ← Data layer

prior layer $\pi(\beta, \sigma^2)$

① ~~Start~~ brm model: intercept 249, $\sigma^2 = 51$

each day increase by 11

reaction time longer, more sleep deprived, all by different amounts

Σ starts from different place, good sign since dont want single regression ∴ dont have a bunch of pts random about a line, seems like we have a bunch of lines

$j = 1, \ldots, J$ Data layer: $y_{ij} \mid \beta_j, x_{ij}, \sigma^2, \beta, \Sigma_\beta \sim N(\beta_{0j} + \beta_{1j} x_{ij}, \sigma^2)$

β's intercept & slope ~~are~~ from each group ∴

Process layer: $\beta \mid B, \Sigma_\beta \sim MVN(B, \Sigma_\beta)$ $\beta = (\beta_0, \beta_1)$

↑ brm always has mean zero in process layer

Prior layer: $\pi(B, \Sigma_\beta, \sigma^2)$

brm version: $y_{ij} \mid \beta_j, x_{ij}, \sigma^2, B, \Sigma_\beta \sim N(b_0 + \beta_{0j} + b_1 x_i + \beta_{1j} x_{ij}, \sigma^2)$

∴ ~~Zero~~ mean where B is, b need to come in mean ∴ b_0

β_{0j} group level deviation from it, β_{0j} has mean zero in next layer

global b regression term for x_i: b_1 (general days for a subject)

$\beta_{1j} x_{ij}$ is (group level) process level deviation from this line

∴ $\beta \mid \Sigma_\beta \sim MVN(0, \Sigma_\beta)$ prior: $\pi(\Sigma_\beta, \sigma^2, B)$ $B = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}$

say all regression params ∴ just regressed on days, all have a group level

having sd says Z rand deviation around Z thing fitted is actually a lot smaller Σ is explainable by something else, eg like Subject variability

now have group level effects, giving sigma (Σ_β), getting Sd of intercept, Sd of days is bottom right diagonal, Correlation is there too

all estied well

>> plot(fit_sleep, N=3) %. plot limited plots%

Sd params are group level params Cor-Subject (Correlation in the Matrix)

>> plot(fit-sleep, N=9, pars="Subject") %. same as before & get Γ's one for each subject & get them for slopes eg for Days is Z

Slope variable ∴ see effects for each subject

>> ranef %. random effects for us isn't random ∴ randomness is in 2 params%.

for intercept, seeing different subjects

>> b-prior %. use default N(0,10) default for regression params%.

brm has population params are class b apart from intercept (intercept does get this prior too), class sd for group level params

final prior on Sigma

~~Week 12 tutorial~~

~~Ch1~~ intro to Bayesian machinery:/

By Bayes $\underset{\text{posterior}}{\pi(\theta|y)} \propto \pi(\theta) \underset{\text{likelihood}}{p(y|\theta)}$ ∝ Z marginal likelihood (prior predictive) $p(y) = \int_{-\infty}^{\infty} p(y|\theta)\pi(\theta)d\theta$

∴ form is proportional to something we know (normal, Gamma, Beta, Binomial)

Conjugate: Form $\pi(\theta) \equiv$ for $\pi(\theta|y)$

Bayesian inference/ estimation: $E[g(\theta)|y] = \int_{-\infty}^{\infty} g(\theta)\pi(\theta|y)d\theta$

use MAP when sampling from $\pi(\theta|y)$ hard ∴ $\arg\max_\theta \pi(\theta|y)$

Bayesian Credible intervals: $[a,b]: \int_a^b \pi(\theta|y)d\theta := P(a<\theta<b|y) := \alpha$ $(100\alpha)\%$

credible interval (not confidence interval)

prediction ~~P(y)~~ $P(y_{new}|y) = \int_{-\infty}^{\infty} \underbrace{p(y_{new}|\theta)}_{\text{likelihood}} \pi(\theta|y)d\theta$

'objective priors:/ Goals improper prior. $\int \pi(\theta)d\theta \neq 1$

'uninformative' prior ⇒ $\pi(\theta) \propto 1$ $\theta \in [0,\infty]$ improper

Jeffery's prior $\pi(\theta) \propto J(\theta)^{1/2}$

Subjective probability theory:/ Event random quantity duality

Event: is TRUE/FALSE. Can be treated as 1/0 is random because we don't know it's 1/0 $A \vee B$ 'v' or $A \wedge B$ '^' and

$\tilde{A}$ '~' not $x \vee y = \max(x,y)$ $x \wedge y = \min(x,y)$ $\sim x = 1-x$

$A \wedge B = AB$ $(\min(A,B))$ $A \vee B = \sim(\tilde{A} \wedge \tilde{B}) = 1-(1-A)(1-B)$

Prove $P(A \vee B) = P(A) + P(B) - P(A|B)$

Week 9 videos / Cake & hierarchical model checking /

temperature took 1 of 6 vals ∴ literature looks at how you treat that as an ordered factor, like recipe A, B, C are factors but no natural ordering, but 175 < 185 ∴ should come into your model somewhere

is replicates ∴ 15 different experiments

>> f dont use >> (temp | recipe * replicate) ∵ since in R it's recipe + replicate + (recipe)(replicate) ∵

so use >> (temp | recipe : replicate) ∵ is correct in R it's (recipe)(replicate) ∵

45 combinations of replicate & recipe

(x.variable | group) is x variable changes with 2 group

45 levels ∴ >> recipe:replicate has 45 possible vals of this grouping variable

want to take out group level results A15, B15, C15 then predict replicate 15 from 2 fit of 2 old data

can plot 95% credible intervals so fine to have 1 pt out of 15 outside 2 error bars ∴ expect 5%.

find good model by write down & justify subjectively, write down priors can justify subjectively ∴ show through these plots, what you've done make sence for data

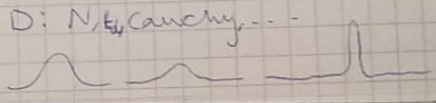
$y_i | \beta, \sigma^2 \sim N(\sum_{j=0}^{p} \beta_j X_{ij}, \sigma^2)$ $\quad \eta_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} \sim N(\eta_i, \sigma^2)$

$y_i | \beta, \theta \sim D(g^{-1}(\eta_i), \theta)$ $\quad$ D has mean $g^{-1}(\eta)$

$\{G \sim (\alpha, \beta)$ $E[G] = \frac{\alpha}{\beta}$, $G(\frac{\alpha}{\beta}, \theta(\alpha, \beta))\}$

$g(\cdot)$ is called 'link function' $\quad \eta \in [-\infty, \infty]$ ∴ role of g is to map $\eta \to$ support of D via g^{-1} $\quad g(x) = \log(x)$ is a popular link when $E[y]$ must be > 0

y: continuous $\quad$ support: $(-\infty, \infty)$ $\quad$ D: N, t, Cauchy...



y: Continuous support: $(0, \infty)$ D: G, lognormal

support: $[a, b]$ D: Beta, scaled beta

y: Discrete support $\{0, 1\}$ D: Bernoulli

$\{0, 1, \ldots, N\}$ D: Binomial

$\{0, 1, \ldots\}$ D: poisson, Negative Binomial

y: Rain $\{0, R\}$ D: Hurdle Model, Zero inflated Models
R is continuous >0

choice of D is called Z family, default is family = gaussian (normal)

>> ? brms family

$y_{ij} \mid \mu_{ij}, \theta \sim D(g^{-1}(\mu_{ij}), \theta) \quad j = 1, \ldots, J$

Data layer:
$\mu_j = b_0 + b_1 x_{1j} + \ldots + b_p x_{pj} + \beta_{0j} + \beta_{1j} x_{1j} + \ldots + \beta_{pj} x_{pj}$

$g(\cdot)$ known

process layer: $\beta_j \mid \Sigma \sim MVN(0, \Sigma)$

prior: $\pi(b, \Sigma, \theta)$

>> ? 'brms-package'

>> brm(formula=, data=, family=D, prior=??) %. if set prior = Null then uses default priors that depend on formula & family %.

%. Should always be specifying own priors %.

%. normal family is gaussian, link is "identity" is only one that makes sense

%. once get data find what type it is (eg count, binary, positive, normal, continuous) then make choice of D (discri wanted? Sort of thin tailed, poisson, negative binomial for counts) then link fund choices default or somewhere else, then what are Z other params then which predictors to put in, which variables to group on so were doing $\beta_j \mid \Sigma \sim MVN(0, \Sigma)$, whats prior on group variables, predictions, other params.) done MCMC, give samples from posterior, gives Monte carlo ests is of predictive distris or of params for inference

prove $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

$y_i \in \{0,1\}$ ~ Bernoulli ~ $Ber(p = g^{-1}(\eta))$ (probability)

$g: [0,1] \to (-\infty, \infty)$ $\eta = \text{logit}(p) = \frac{p}{1-p}$ (link) $\therefore g^{-1}(\eta) = \frac{1}{1+e^{-\eta}}$

Logistic Regression

\Week 12 videos/ good Bayesian analysis: write down & justify model structure & priors (30%-40%)

$y_i | \beta, x_{ij}, \sigma^2 \sim N(\beta_{0j} + \beta_j x_{ij} + \dots, \sigma^2)$ (normal)

$\beta_j \sim MVN(B, \Sigma)$

prior distri $\pi(B, \Sigma, \sigma^2)$

$\beta_1 \sim N(100, 0.01)$

- Convergence (10-20%)
- Model checking: posterior predictive plots for (hopefully withheld) data [out of sample - if in sample - justify (i.e. not much data)]. Are prediction intervals sensible etc
- posterior (densities) param plots, are they sensible, is Σ param doing anything, if so can we take it out & refit. Is Σ posterior different to Σ prior (at param level

- inference: $E[g(\tilde{y}) | \underline{y}]$ $E[\psi(\theta) | \underline{y}]$ $p(f(\tilde{y}) | \underline{y})$, $p(\psi(\theta) | \underline{y})$

θ Show posterior densities & comment

then sensitivity analysis

\Modelling:/ y_i = Total$_i$ $i = 1, \dots, 30$ $y_i | \underline{\theta} \sim \text{Poisson}(\lambda(x))$

(λ inverse link) $\log \lambda(x) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$

x_1 high-temp x_2 low-temp, x_3-precip x_4 snow day

(Days Grouping?)

$j = \{1, 2\}$ 1 - weekday, 2 weekend $n_2 = 9$, $n_1 = 21$

$y_{ij} | \beta_j, x_{ij} \sim \text{Poisson}(\lambda(\beta, x))$ $\log \lambda(\beta, x) =$ ~~$\beta_{0j} + \beta_{1j} x_{HT,j}$~~

$\beta_{0j} + \beta_{1j} x_{HT,j} + \beta_{2j} x_{LT,j} + \beta_{3j} x_{P,j} + \beta_{4j} x_{S,j}$

$\beta \sim N(B, \Sigma)$ prior $\pi(B, \Sigma)$

Estimate Sd Q2.5

r-group[..., Intercept]

r-group[..., SameX]

rare $[c(4,56), 'Estimate']

{>>stersplit}

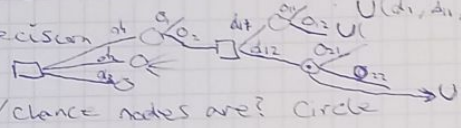
Same priors & grouping is collusion & litteraly impossible without it (& same chains)

Follow week 10 cyclists data for high marks when doing coursework

Menti /1/ which is not a necessary ingredient of a decision problem?

Data

/2/ on a decision tree, what shape are Z decision nodes? Squares

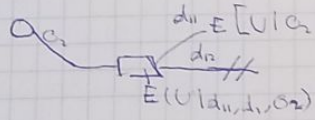


$E[U|O_2, d_1, d_{11}]$

/3/ Chance nodes are? Circle

/4/ starting at Z RHS of Z tree, when get to a chance node must:
Mark with Z conditional expectation of Z leaves to Z right given Z branch you are on

/4/ starting at Z RHS of Z tree, when get to a decision node must:
Mark with Z maximum utility to Z right



/5/ a utility func maps rewards (of any kind) to Z real line & has Z properties. what are they? Z magnitude of utility respects Z preference ordering Z utility of a gamble is equal to Z expected utility

/6/ a utility func always exists? TRUE

/7/ how unique is your utility func? it is unique up to positive linear transformation least $U=0$ $U(best)=1$

Q3/ let utility curve for money $U(x)$ be linear. Let d_1 be receiving £50, d_2 be toss a coin & receive £101 if heads else £0?

Choose d_2 $E[U(d_2)] = 50.5$ $E[U(d_1)] = 50$

Q4/ let utility over returns x, $U(x) = 2\log(1+x)$ $G_{T_1} = 0.4r_1 + 0.6r_2$, $G_{T_2} = 0.1r_3 + 0.9r_2$ with $r_1 = 3$, $r_2 = 10$, $r_3 = 15$ G_{T_1} or G_{T_2}? G_{T_3}?

Week 10 videos/ Modelling: $\log(\lambda(\beta, x_{ij})) = b_0 + B_1 x$

$b_0 + b_1 x_{HT;j} + b_2 x_{LT;j} + b_3 x_{pij} + b_4 x_{s;i;j} + \beta_{0j} + \beta_{1j} x_{HT,j} + \dots$

$F_j \sim N(0, \Sigma)$

need prior for $\pi(b, \Sigma)$

$E[y] = \lambda = e^{f_{abc} + b_1 x_1 + \dots} f_0 + \dots = e^{b_0} e^{b_1 x_1} e^{b_2 x_2} \dots$

$e^{b_0} \in [5000, 20000]$ $\therefore$ >>log(c(5000, 20000)) $\Rightarrow$ (8.5, 9.9)

$N(0, 5)$ {maybe use $N(4, 3)$ for sensitivity analysis/ let 2 end}

$b_1 \sim N(0, 0.05)$ {>>exp(0.1*c(35, 81)) $\Rightarrow$ 33, 3294 $\therefore$ $\frac{0.1}{2} = 0.05$ $\therefore$ $N(0, 0.05)$}

$\Sigma = \text{vec}(\sigma) \times \Omega_{2k} \times \text{vec}(\sigma)$ $\sigma = \sigma_1, \sigma_2, \dots, \sigma_p$ ← snowday

(intercept)

>>set_prior(class = "sd") $t_3(0, 5)$

$N(0, 0.05)$ non-intercept sds

Assessing Convergence: precipitation amount is two orders of magnitude lower $\therefore$ expect to hit it at 2 orders of magnitude higher $\therefore N(0, 5)$

{can be upto 30 seconds per chain} {very bad Rhat ≠ 1 $\therefore$ don't use} {always marks for trace plots to check can interpret} trace plots are starting in different places & getting stuck $\therefore$ hasn't converged $\therefore$ 4 chains not converging} mean alot longer to chace to warm up $\therefore$ allow stan to run for longer >>iter = 10000, warmup = 8000

{3 mins per chain} 10000 samples, 2000 can see each chain, started in different places & not converged

$\therefore$ Maybe starting in a bad place? {unless perfect Rhats, they're bad,

nonetheless but Ess}

Missing everywhere ∴ poisson has same mean & var ∴ eg at mean 25000, 2 variances is 25000

$y|\beta, x \sim D(g^{-1}(1), \phi)$ {never seen a poisson regression work}
(mean, params)

$y|\beta, x \sim NegBin(\lambda(\beta, x), \phi)$ $p(y=k|\lambda,\phi) \propto \left(\frac{\mu}{\phi+\mu}\right)^k \left(\frac{\phi}{\phi+\mu}\right)^\phi$

→ Poisson as $\phi \to \infty$

$E[y] = \lambda \geq 0$ ∴ can still use $\log\lambda(x,\beta) = \sum_{i=0}^{p} \beta_i x_i$

$var[y] \neq \lambda$ ∴ $Var[y] = \lambda(1 + \frac{\lambda}{\phi})$ for $\phi \geq 0$

$\pi(\beta, \Sigma) = \pi(\beta, \sigma)$ ∴ now need $\pi(\beta, \sigma, \phi)$

>> sqrt(vars(10)) ⇒ 1600, 7000 % is bc error is really bad

>> sqrt(vars(100)) % gives 20% error ~~is~~ to bad %

>> sqrt(vars(600)) % error bars getting tight is 3000 ± 400 %

∴ Set Mean=200, var=300 eg 200+3*300=500 ∴ set as prior

∴ trace plots look better, signs things have converged, Rhats good

∴ very good argument for convergence, trace plots excellent, no low Ess's ∴ can do Monte Carlo & inference, is 2 model is ok

>> preds % error bars are pretty big, for lower vals model is fairly skillful, everything's hitting 2 error bars ∴ model is ok, only one missing pt %

Model checking: error bars too big ∴ decrease var of shape prior {low temp is mean zero & unsure of sign ∴ chance low temp is - effectively not a thing ∴ look into taking it out potentially get rid of snow day ∴ clearly not doing anything & hasn't moved from 2 prior {in regression & Bayesian regression, shouldn't take out two variables at one time, could be they're perfectly correlated, & has 2 effect was spread betw them} {could put strong prior on snow eg mean!) negative, say people don't cycle in it}

is >> MCMC-pairs shows high correlation, its saying high-temp &

low.temp are doing 2 same sorts of thing & would be a good argument to kick one of them out. First kick out snow day then look again at this distri for low-temp to see if there's are correlation.

now it's converged dont need such a hard warmup ∴ decrease it

∴ >>iter = 6000, warmup = 5000

∴ only one pt outside, model believable, still a bit big uncertainty

gives probab got sign wrong from 2 expectation, & it has a negative on cyclists is 0.158 ∴ fairly sure of it

but for low-Temp it's 0.357 probab it's 2 other sign

∴ low-temp does see M. to have an effect just a high chance of removing it

{focus on 2 biggest change (just make one) & demonstrate 2 effect it made (staying sensible)}

>> waves (third_model) ∴ precip effect looks tiny

∴ only 1 pt missing ∴ comfortable with it

>> summary(Sens.model) ∴ Rhats are fine, were converged, a much smaller model

conclusions insensitive to main choices {document decisions made}

final model then change something to check sensitivity

∴ final model is fourth model

Inference: ∴ prob that increase in High-Temp leads to more cyclists is 1 − 0.169 ∴ 83.1%

desinetly reduce cyclists by precip increase

{learn how to plot random effects

to predict probab of an event ∴ need to do a posterior prediction,

comes from 2 samples, & comes directly from 2 pred func

eg >> predict(...) ∴ give it 2 new data you want, it gives Expectation of 2 probab, 2 standard deviation of 2 samples (Est.Error) ∴ need