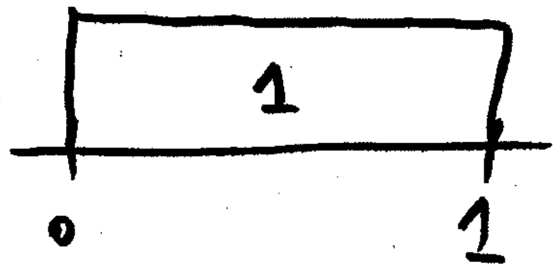
handwritten notes ①
1 Feb

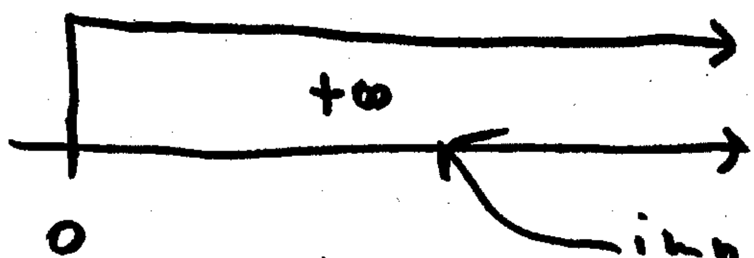
part 1

diffuse



$p(\theta) = U(0, 1)$

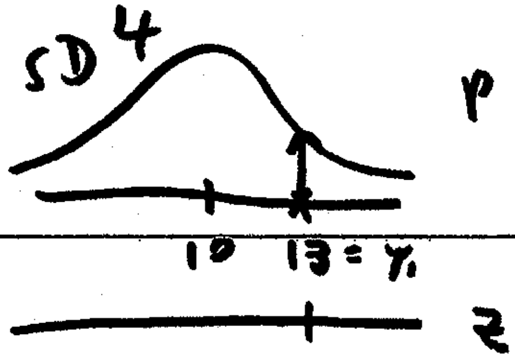
diffuse



$p(\lambda)$

improper prior (sometimes <u>incoherent</u>)

$U(0, \infty)$



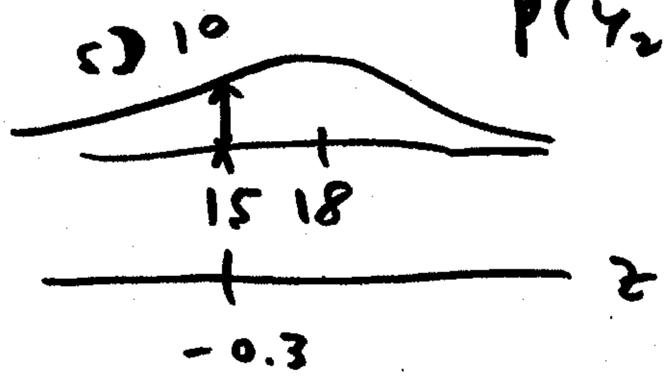
$$\frac{13-10}{4} \approx +.75$$

$p(y_1 \mid y_{-1})$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$y_{-1} = (y_2, \ldots, y_n)$

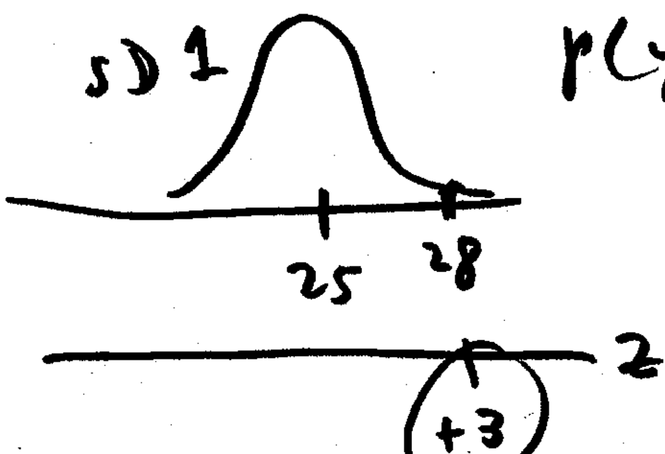
(cross-validation)



$p(y_2 \mid y_{-2})$

(log score) criterion

$\vdots$



$p(y_n \mid y_{-n})$

$$\left.\begin{array}{c} 0 \\ \vdots \\ 0 \end{array}\right\} \beta \qquad \left.\begin{array}{c} 1 \\ \vdots \\ 1 \end{array}\right\} \alpha \qquad \updownarrow \; (\alpha+\beta) = \underline{30}$$

prior sample size

prior like a data set (*)

mean $\dfrac{\alpha}{\alpha+\beta} = .15$

---

data (**)

$$\left.\begin{array}{c} 0 \\ \vdots \\ 0 \end{array}\right\} \underline{(n-s)} \qquad \left.\begin{array}{c} 1 \\ \vdots \\ 1 \end{array}\right\} \underline{(s)} \qquad \updownarrow \; n = \underline{400}$$

data sample size

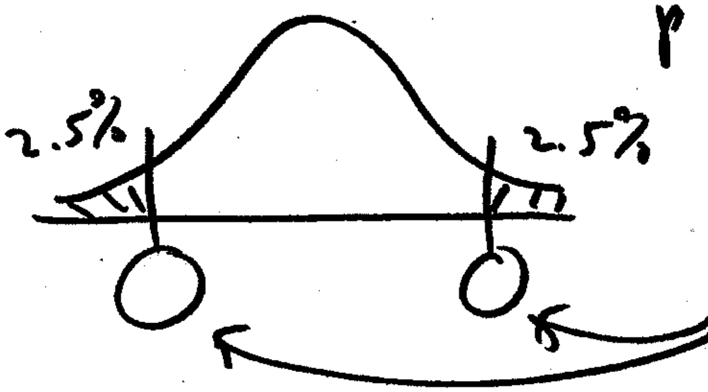
mean $\bar{y} = \dfrac{s}{n} = .18$

---

the prior is like a data set: imagine merging prior, sample datasets & feed merged data to the likelihood machinery: equivalent to Bayesian analysis with this (*) prior, this (**) sample.

$$\left.\begin{array}{c} 0 \\ \vdots \\ 0 \end{array}\right\} (\beta+n-s) \qquad \left.\begin{array}{c} 1 \\ \vdots \\ 1 \end{array}\right\} (\alpha+s)$$

$p(\theta \mid y)$



95% central Bayesian post. interval

$$\text{post SD} = \sqrt{\frac{\alpha^* \beta^*}{(\alpha^* + \beta^*)^2 (\alpha^* + \beta^* + 1)}}$$

$\alpha^* = \alpha + s$

$\beta^* = \beta + n - s$

$$\doteq \sqrt{\left(\frac{\alpha^*}{\alpha^* + \beta^*}\right)\left(\frac{\beta^*}{\alpha^* + \beta^*}\right)\left(\frac{1}{\alpha^* + \beta^*}\right)}$$

$$= \sqrt{\frac{\hat{\theta}_B (1 - \hat{\theta}_B)}{n_B}}$$

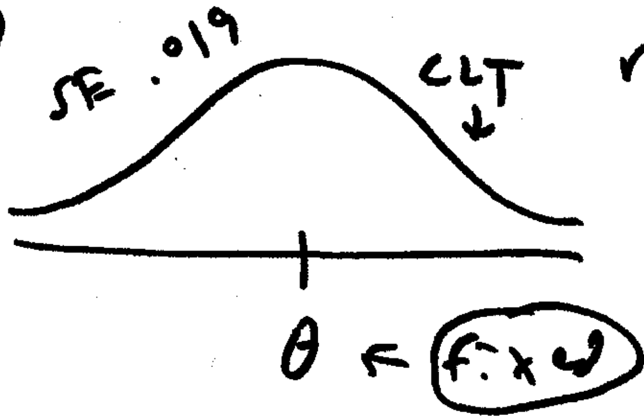
$n_B \leftarrow$ post. sample size

$\frac{\alpha^*}{\alpha^* + \beta^*}$ = post. mean = $\hat{\theta}_B$ = Bayesian point estimate

$\frac{s}{n} = \hat{\theta}_{MLE}$ = frequentist point est.

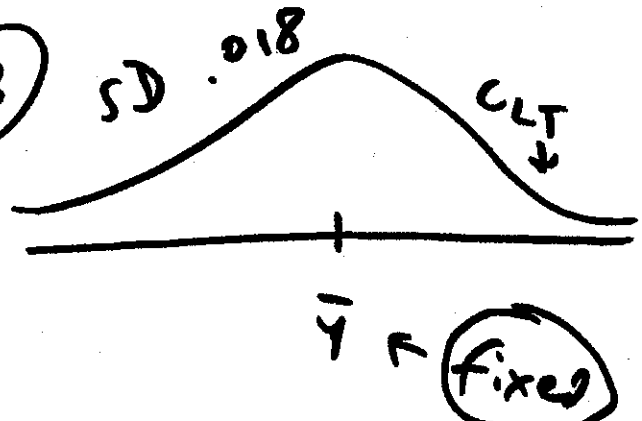
$\hat{SE}(\hat{\theta}_{MLE}) = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$

(F)



repeated-sampling distribution of $\bar{Y}$ ← random

(B)



≐ posterior distribution of θ given $\bar{Y}$ & diffuse prior info

random

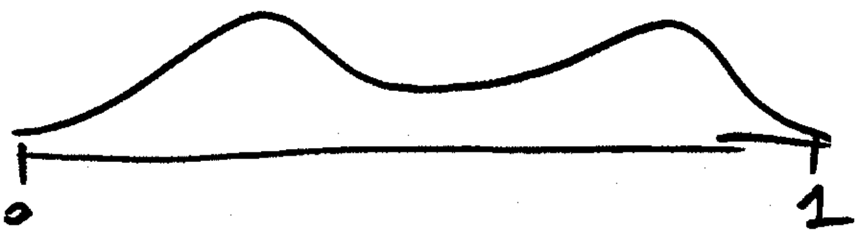
(F) $c_1 e^{-c_2(\bar{Y}-\theta)^2}$

$\quad\quad =$

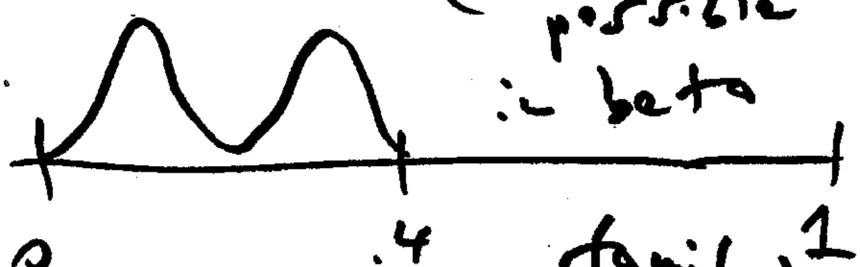
(B) $c_1 e^{-c_2(\theta-\bar{Y})^2}$

same if n large & prior info small

(Bernstein - Von Mises Theorem)



$p(\theta)$ in AMI case study



(not possible in beta family)

$$\begin{array}{c|c} Y & X \\ \hline 1\text{s} & \text{c} \\ \& & \text{o} \\ 0\text{s} & \text{n} \\ & \text{t.} \end{array} \quad \updownarrow n$$

standard logistic regression model:

$$(Y_i \mid p_i) \overset{\text{indep}}{\sim} B(p_i)$$

$$\boxed{\log\left(\frac{p_i}{1-p_i}\right)} = \beta_0 + \beta_1 X_i$$

← natural parameterization in Bernoulli likelihood steps in Bayesian conjugate machinery:

1. identify joint sampling distribution for observables

$$p_{Y_1, \ldots, Y_n}(y_1, \ldots, y_n \mid \theta)$$ will involve $\underline{\theta}$

2. think of 1 as f'n of θ for fixed $y = (y_1, \ldots, y_n) \Rightarrow$ lik f'n $\ell(\theta \mid y)$

3. ~~try to find~~ think of $\ell(\theta \mid y)$ as density in θ & try to find another density $p(\theta)$ s.t. $\ell(\theta \mid y) \cdot p(\theta)$ has same math. form as $p(\theta)$: this is conj. prior for θ (choose a member of this family)

(4) use Bayes' Thm. to do conj. updating (5) create predictive dist for future data (8)

$$\ell(\lambda|y) = \left(\frac{\lambda^{y_1} e^{-\lambda}}{y_1!}\right)\left(\frac{\lambda^{y_2} e^{-\lambda}}{y_2!}\right)\cdots\left(\frac{\lambda^{y_n} e^{-\lambda}}{y_n!}\right)\cdot c$$

$$= c\,\frac{\lambda^{y_1+\cdots+y_n}\, e^{-n\lambda}}{\cancel{\prod_{i=1}^{n} y_i!}} = c\,\lambda^{s} e^{-n\lambda} \quad \left(s = \sum_{i=1}^{n} y_i\right)$$

s is evidently <u>sufficient</u> for λ in this model

lots of suff. stat. in any given problem, e.g. y is itself suff. (but not helpful to notice this) another suff. stat is (n even) 9

$$\left(\sum_{i=1}^{n/2} y_i \; , \; \sum_{i=n/2+1}^{n} y_i \right)$$ this gets us down

from n dim. to 2 but s gets down from n to 1 & is therefore a "better" suff. stat (s is 9 <u>minimal</u> suff. stat because can't get any lower than 1 dimension with 1 parameter)

Empirical Rule (E.14) For almost any dist., if you start at mean & go

1 SD either way, you'll capture about $\frac{2}{3}$ (68%)

2 SDs ———————————— most (95%)

3 SDs ———————————— almost all (99.7%)

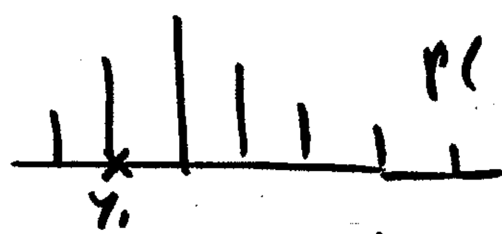
E.R. exactly followed by all Gaussian dist. & approx. true for almost all non-Gaussian dist. as well. (10)

$$V(y_{n+1} \mid y) = \left(\frac{\alpha+s}{\beta+n}\right)\left(1 + \frac{1}{\beta+n}\right)$$

uncertainty about y_{n+1} given λ

uncertainty about λ

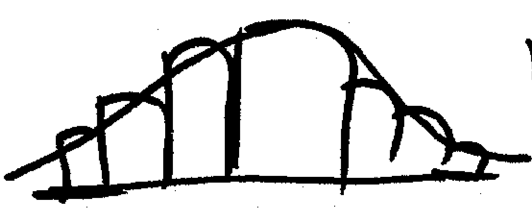
$$\begin{bmatrix} y_2 \\ \vdots \\ y_n \end{bmatrix} = y_{-1}$$



$p(y_1 \mid y_{-1})$

$\vdots$

omit each y_i one at a time in turn



histogram of z