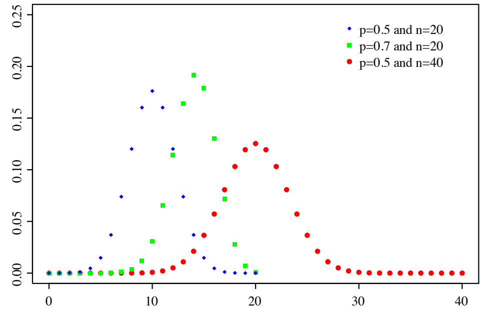
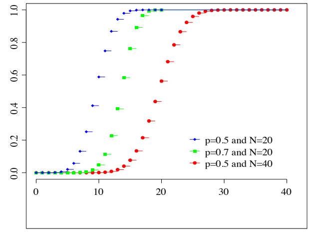
Royal Institute of Technology

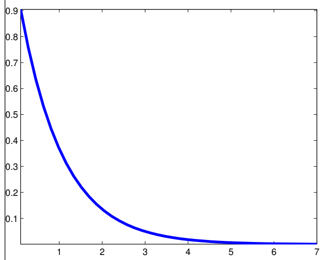
DD2447 STAT. METH. IN CS HT 2014

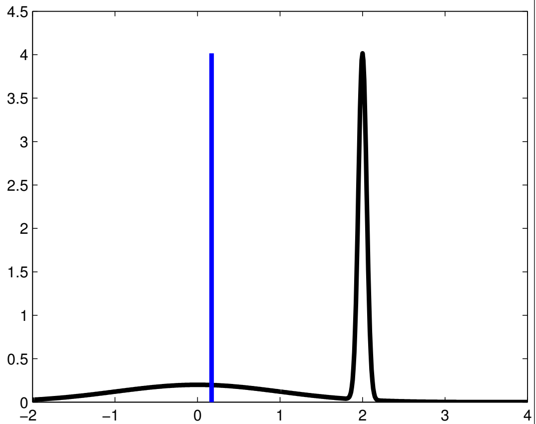
Lecture 4 - Ch. 5, Bayesian concepts



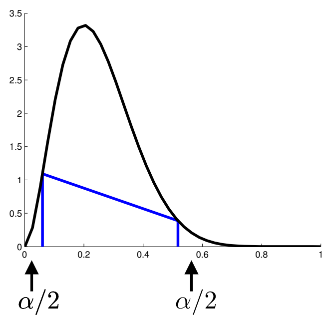


BINOMIAL DISTRIBUTION

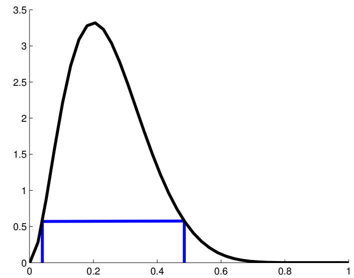




MAP

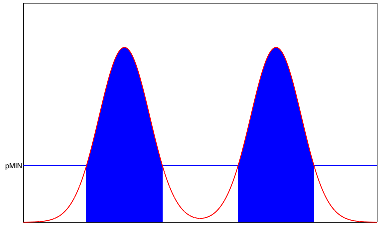


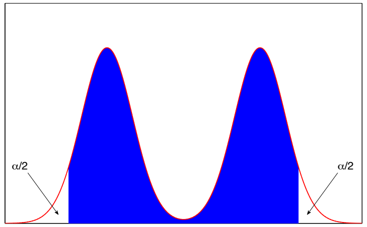
$(1-\alpha)$central intervall



$(1-\alpha)$ highest posterior interval

CREDIBLE INTERVALS





highest posterior intervall

central intervall

CREDIBLE INTERVALS

AMAZON SELLERS

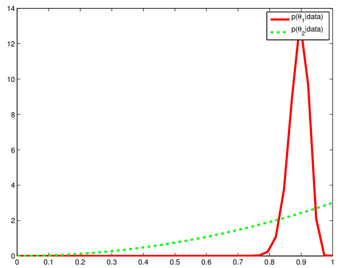
- Seller 1 – 90 positive, 10 negative
- Seller 2 – 2 positive, 0 negative
- θ1 and θ2 reliabilities with prior Beta(1,1)
- Uniform on sellers

$$\delta = \theta_1 - \theta_2$$

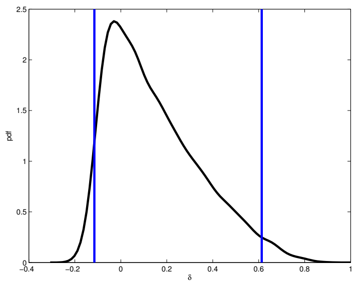
$$p(\delta > 0|\mathcal{D}) = \int_0^1 \int_0^1 \mathbb{I}(\theta_1 > \theta_2)\text{Beta}(\theta_1|y_1+1, N_1-y_1+1) \text{Beta}(\theta_2|y_2+1, N_2-y_2+1)d\theta_1 d\theta_2$$

$$p(\delta > 0|\mathcal{D}) = 0.710,$$

AMAZON SELLERS



posteriors



95% central interval

5.3 BAYESIAN MODEL SELECTION

- Trade off: low model complexity & underfit vs high model complexity & overfit
- How to find sweet spot
  - Cross-validation
  - Posterior probability of models
    $$p(\mathcal{M}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{M})p(\mathcal{M})}{p(\mathcal{D})}$$
- Bayesian: pick MAP model
  $$\hat{\mathcal{M}} = \text{argmax}_{\mathcal{M}} p(\mathcal{M}|\mathcal{D})$$
- Or don't pick one average
- If uniform prior, this is ML of marginal

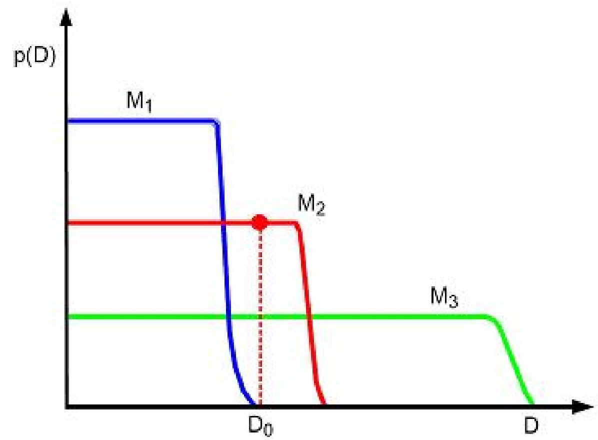
$$\text{argmax}_{\mathcal{M}} = \int_{\mathcal{M}} p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M})p(\boldsymbol{\theta}|\mathcal{M})$$

5.3.1 BAYESIAN OCCAM'S RAZOR

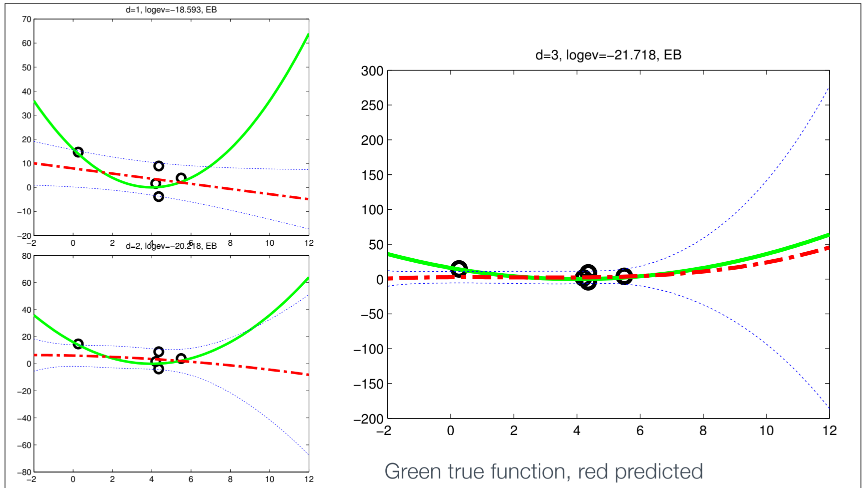
- Marginalizing protects against overfitting
- If M'⊂ M (nested models), then

$$p(\mathcal{D}|\boldsymbol{\theta}^{\mathcal{M}}_{\mathrm{ML}}) \geq p(\mathcal{D}|\boldsymbol{\theta}^{\mathcal{M}'}_{\mathrm{ML}})$$

- Also true for MAP with uniform prior (& others)
- More possible parameters choices for M, than M'.
- By marginalized likelihood, these are taken into account

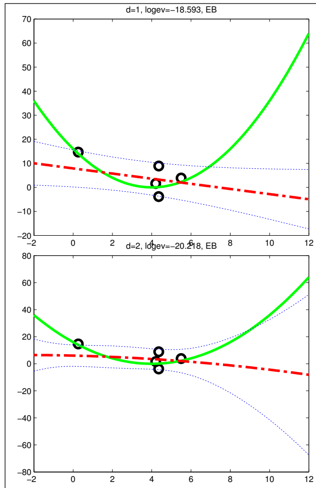


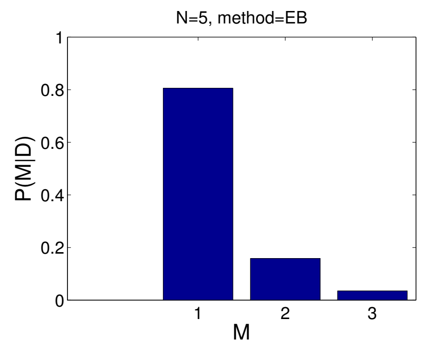
MODELS – BENEFITS OF COMPLEX VS. LESS SO



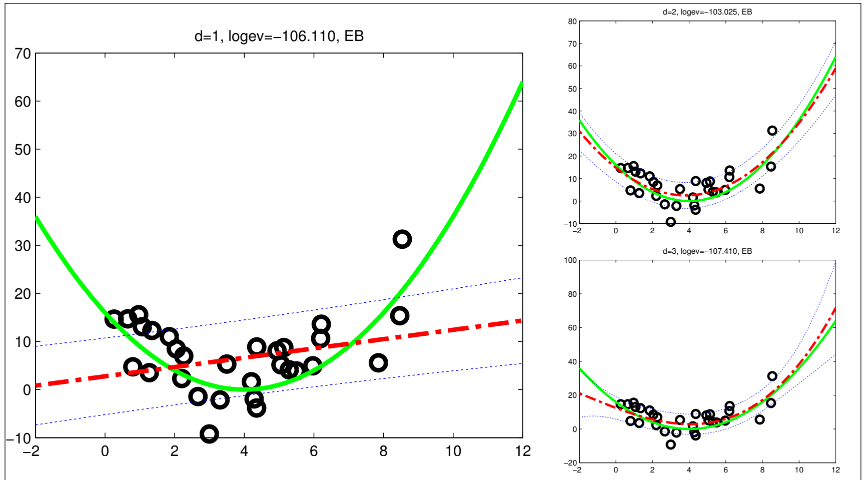
Green true function, red predicted

- Fitting degree 1-3 to 5 points.

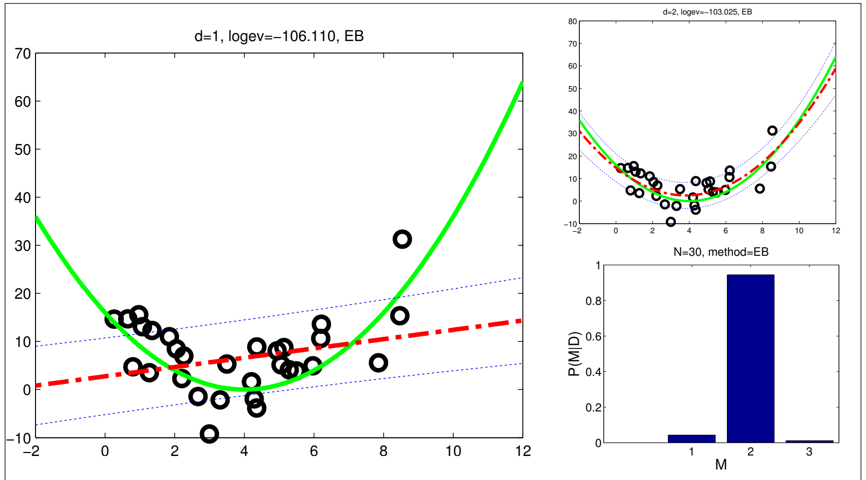




- Fitting degree 1-3 to 5 points & posterior



30 POINTS



30 POINTS

DO 5.3.2, IN YOUR OWN WAY

BAYESIAN INFORMATION CRITERIA (BIC)

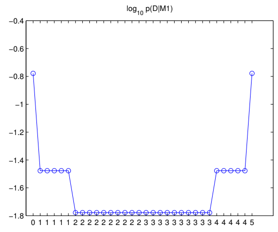
$$\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$$

$$\mathrm{BIC}(\mathcal{D}, \mathcal{M}) = \underbrace{\log p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{ML}}^{\mathcal{M}})}_{\propto N} - \underbrace{\mathrm{dof}(\mathcal{M}) \log(N)}_{\propto \log N}$$

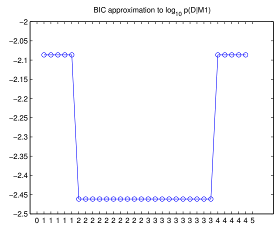
$$\log p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{ML}}^{\mathcal{M}}) = \sum_{n \in N} \log p(\boldsymbol{x}_n|\boldsymbol{\theta}_{\mathrm{ML}}^{\mathcal{M}})$$

- Computing the marginal likelihood often hard
- BIC score is an approximation of it
- dof - degrees of freedom ≈ number or parameters
- Popular approach
- Above example dof=1,2, 3

TESTING A COINS FAIRNESS



$N_1 = \#$ heads
$N_0 = \#$ tails



- Two models (hypothesis): M_0 fair coin Θ=1/2, M_1 Θ is Beta(1,1)
  - M_0 likelihood $p(\mathcal{D}|\mathcal{M}_0) = 2^{-N}$
  - M_1 marginal likelihood

$$p(\mathcal{D}|\mathcal{M}_1) = \int_{\boldsymbol{\theta}} p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = B(N_1+1, N_0+1)/B(1,1)$$

HOW TO CHOSE HYPER PARAMETERS

$$p(\mathcal{D}|\mathcal{M}) = \int\int p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{\alpha},\mathcal{M})p(\boldsymbol{\alpha}|\mathcal{M})\mathrm{d}\boldsymbol{w}\mathrm{d}\boldsymbol{\theta}$$

$$\begin{aligned}\hat{\boldsymbol{\alpha}} &= \mathrm{argmax}_{\boldsymbol{\alpha}} p(\mathcal{D}|\boldsymbol{\alpha},\mathcal{M}) \\ &= \mathrm{argmax}_{\boldsymbol{\alpha}} \int_{\boldsymbol{w}} p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{\alpha},\mathcal{M})\mathrm{d}\boldsymbol{w}\end{aligned}$$

$$p(\mathcal{D}|\mathcal{M}) \approx \int p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w}|\hat{\boldsymbol{\alpha}},\mathcal{M})\mathrm{d}\boldsymbol{w}$$

- You chose
- Prior on the prior
- The higher in hierarchy, the less effect of parameters
- Empirical Bayes: estimate the level 2 parameters

LEVELS OF BAYESIANISM

Method	Definition
Maximum likelihood	$\hat{\boldsymbol{\theta}} = \mathrm{argmax}_{\boldsymbol{\theta}}\, p(\mathcal{D}|\boldsymbol{\theta})$
MAP estimation	$\hat{\boldsymbol{\theta}} = \mathrm{argmax}_{\boldsymbol{\theta}}\, p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\eta})$
ML-II (Empirical Bayes)	$\hat{\boldsymbol{\eta}} = \mathrm{argmax}_{\boldsymbol{\eta}} \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\eta})d\boldsymbol{\theta} = \mathrm{argmax}_{\boldsymbol{\eta}}\, p(\mathcal{D}|\boldsymbol{\eta})$
MAP-II	$\hat{\boldsymbol{\eta}} = \mathrm{argmax}_{\boldsymbol{\eta}} \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\eta})p(\boldsymbol{\eta})d\boldsymbol{\theta} = \mathrm{argmax}_{\boldsymbol{\eta}}\, p(\mathcal{D}|\boldsymbol{\eta})p(\boldsymbol{\eta})$
Full Bayes	$p(\boldsymbol{\theta},\boldsymbol{\eta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\eta})p(\boldsymbol{\eta})$

BAYES FACTORS

$$\begin{aligned}\mathrm{BF}_{\mathcal{M},\mathcal{M}'} &:= \frac{p(\mathcal{D}|\mathcal{M})}{p(\mathcal{D}|\mathcal{M}')} \\ &= \frac{p(\mathcal{M}|\mathcal{D})/p(\mathcal{M})}{p(\mathcal{M}'|\mathcal{D})/p(\mathcal{M}')}\end{aligned}$$

Bayes factor $BF(1,0)$	Interpretation
$BF < \frac{1}{100}$	Decisive evidence for M_0
$BF < \frac{1}{10}$	Strong evidence for M_0
$\frac{1}{10} < BF < \frac{1}{3}$	Moderate evidence for M_0
$\frac{1}{3} < BF < 1$	Weak evidence for M_0
$1 < BF < 3$	Weak evidence for M_1
$3 < BF < 10$	Moderate evidence for M_1
$BF > 10$	Strong evidence for M_1
$BF > 100$	Decisive evidence for M_1

Table 5.1 Jeffreys' scale of evidence for interpreting Bayes factors.

- ★Ratio between marginals
  - Natural way to compare models
- ★ But what do they mean?
- ★ "However, ultimately our goal is to convert our beliefs into actions."

MIXTURE OF CONJUGATE IS CONJUGATE PRIOR

$$p(\boldsymbol{\theta}) = \frac{1}{2}\mathrm{Beta}(\boldsymbol{\theta}|20, 20) + \frac{1}{2}\mathrm{Beta}(\boldsymbol{\theta}|30, 10)$$

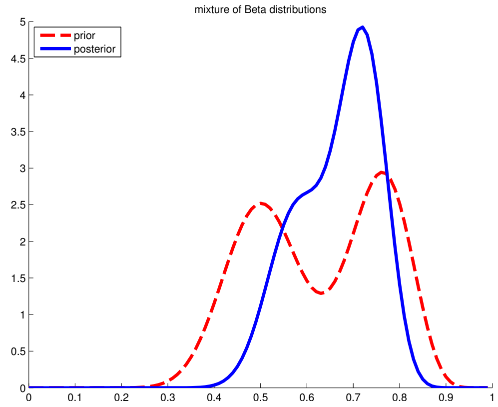
In general

$$p(\boldsymbol{\theta}) = \sum_k \underbrace{p(\boldsymbol{\theta}|z=k)}_{\text{conjugate}} \underbrace{p(z=k)}_{\text{mixing weights}}$$

Gives posterior

$$p(\boldsymbol{\theta}|\mathcal{D}) = \sum_k p(\boldsymbol{\theta}|z=k, \mathcal{D})p(z=k|\mathcal{D})$$

- Can approximate any prior
- Say, likelihood Ber(Θ)
- Mixing with prior weights gives posterior weights



MIXTURE OF 2 BETAS

CONSERVED DNA SEQUENCES

```
pap   CATTTAGACGATCTTTTATGCTGT-AAATTCAATTTGCCATGATGTTTTTATCTGAGTACCCTCTTGCTATTAGTGTTTT
foo   CATTTAGACGATCTTTTATGCTGT-AAATTCAATTCACCATGATGTTTTTATCTGAGTGTATTCTTGTTGTTTGTGTTTT
sfa   GATTTTAACGATCTTTTATACTGA-ATATTCATGCTTATACAGTATTAATAACTAAAACGCCAATCCACTGCGAGATATA
afa   GATTATAACGATCTTTTATTCTAC-ACATGAATAATATCCCGGTTATATATTCTGATTGTATTCTTTTTTGTGTTATCTG
daa   AATTATAACGATCTTTTATTCTGC-ATATGAATAATATTCCAGTCATATATTCTGATTGTATTCTTTTTTGTGTTATATG
clp   GTTTTAGACGATCTTTTTATCTGTGATTTTTGTGTTTTTTTGGTGGCTTTTGCTGTTGTTTGTTGTTTTTGTGGTTTTAT
fae   TGAAATAGCGATCTTTTATTTGTGTATTTTTTTCTGTTTTTGGTTGTTTGTGCTGTGCTTTTGTGTTGTTGTGGTTTTTT
pef   TGCTATAGCGATCTTTTATTCTGCATTATGTATGATAA-ATAGACGTTTAACCTGTTTTCTTGTGATTTTGTGGTGAAAC
```

- Functional elements are more conserved than non-functional
- Dependence between consecutive positions, which we ignore

TWO MODELS – THAT CAN BE TESTED

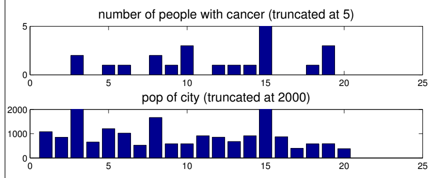
t – column of nucleotides

$\mathbf{z}_t = 1$ if conserved; $\mathbf{z}_t = 0$ if not conserved

$$p(N_t|z_t) = \int_{\mathcal{M}} p(N_t|\theta_t)p(\theta_t|z_t)\, d\theta_t$$

$$p(\theta_t|z_t = 1) = \frac{1}{4}[\mathrm{Dir}(\theta|10,1,1,1) + \cdots + \mathrm{Dir}(\theta|1,1,1,10)]]$$

$$p(\theta_t|z_t = 0) = \mathrm{Dir}(\theta|1,1,1,1)$$



HIERARCHICAL BAYES: EX. CANCER RATES

- Model Bin(x|ϴ, N)
- Alternatives:
  - cities independent
  - tie ϴ
  - common prior Beta(ϴ|a,b)
- Problem: small city, poor estimate
- Tie – pool data and use MLE
  - but we expect differences

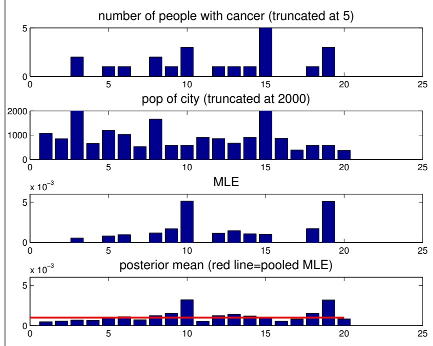
$\mathcal{D} := \{x_1, \ldots, x_S\}$

Full joint

$$p(\mathcal{D}, \boldsymbol{\theta}, \overbrace{\boldsymbol{\eta}}^{(a,b)} | \boldsymbol{N}) = p(\mathcal{D}, \boldsymbol{\theta} | \boldsymbol{\eta}, \boldsymbol{N}) p(\boldsymbol{\eta}) = p(\boldsymbol{\eta}) \prod_{i \in [S]} \mathrm{Bin}(\boldsymbol{x}_i | \boldsymbol{N}_i, \boldsymbol{\theta}_i) \mathrm{Beta}(\boldsymbol{\theta}_i | \boldsymbol{\eta})$$

EX. CANCER RATES – CONTINUED

- x_i cancer deaths in city i
- N_i population size in city i
- Common prior Beta(a,b)
- η = (a,b) must be inferred
  - otherwise (given η=(a,b)) cities are independent

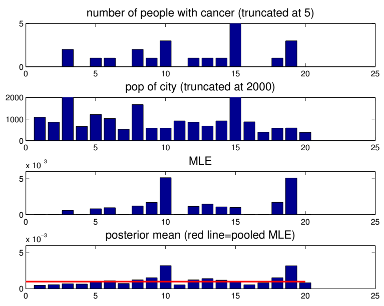


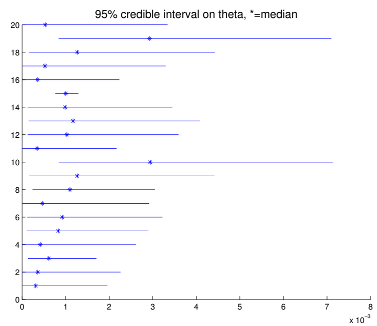
Means of marginals of joint posterior

HIERARCHICAL BAYES: EX. CANCER RATES

- Model Bin(x|ϴ, N)
- Alternatives:
  - cities independent
  - tie ϴ
  - common prior Beta(ϴ|a,b)

CANCER RATES CREDIBLE INTERVAL





5.7 BAYESIAN DECISION THEORY

- Given x we chose an action a
- Loss L(y,a) measured compared to hidden state/ param./class y
  - misclassification L(y,a)=I(y≠a)
  - squared L(y,a)=(y-a)^2
- Economics utility U(y,a)=-L(y,a)
- Optimal decision procedure

$$\delta(\boldsymbol{x}) = \text{argmin}_a E[L(y,a)]$$

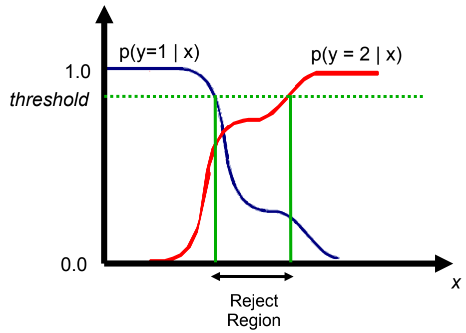
- Bayesian approach expected posterior loss

$$\rho(a|\boldsymbol{x}) = \text{argmin}_{p(y|x)} E[L(y,a)]$$
$$= \sum_y p(y|x) L(y,a)$$

if discrete

MAP MINIMIZES 0-1 LOSS

- L(y,a)=I(y≠a)
- ρ(a|x) =p(a≠y|x)=1-p(a|x)
- Hence,
  - maximizing p(a|x)
  - minimizes ρ(a|x)
- Read about reject option



POSTERIOR MEAN MINIMIZES QUADRATIC (L_2) LOSS

$$\rho(a|\boldsymbol{x}) = E[(y-a)^2|\boldsymbol{x}]$$
$$= E[y^2|\boldsymbol{x}] - 2aE[y|\boldsymbol{x}] + a^2$$

$$\frac{\partial \rho(a|\boldsymbol{x})}{\partial a} = 2a - 2E[y|\boldsymbol{x}]$$

$$\frac{\partial \rho(a|\boldsymbol{x})}{\partial a} = 0$$

⬇

$$a = E[y|\boldsymbol{x}] = \int y\, p(y|\boldsymbol{x}) \mathrm{d}$$

- Squared loss L(y,a)=(y-a)^2
- Assuming continues

TRUE AND FALSE, POSITIVE AND NEGATIVE

		Truth 1	0	Σ
Estimate	1	TP	FP	$\hat{N}_+ = TP + FP$
	0	FN	TN	$\hat{N}_- = FN + TN$
	Σ	$N_+ = TP + FN$	$N_- = FP + TN$	$N = TP + FP + FN + TN$

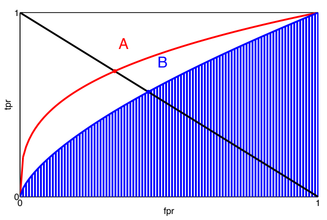
Table 5.2 Quantities derivable from a confusion matrix. N_+ is the true number of positives, $\hat{N}_+$ is the "called" number of positives, N_- is the true number of negatives, $\hat{N}_-$ is the "called" number of negatives.

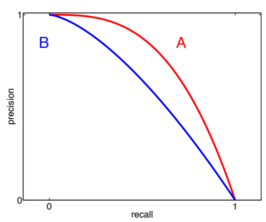
	$y=1$	$y=0$
$\hat{y}=1$	TP/N_+=TPR=sensitivity=recall	FP/N_-=FPR=type I
$\hat{y}=0$	FN/N_+=FNR=miss rate=type II	TN/N_-=TNR=specifity

Table 5.3 Estimating $p(\hat{y}|y)$ from a confusion matrix. Abbreviations: FNR = false negative rate, FPR = false positive rate, TNR = true negative rate, TPR = true positive rate.

- Binary decision problem: for x ∊ U, x ∊ C?
- Positives are the ones claimed to be in C
  - true or false depending on membership of C

RECEIVER OPERATING CHARACTERISTICS (ROC) CURVES





- True positive rate (TPR; recall): TP/(TP+FN)=TP/|C|=p(ŷ=1| y=1)
- False positive rate (FPR): FP/(TN+FP)=FP/|U\C|=p(ŷ=1| y=0)
- Precision: TP/(TP+FP)=p(y=1| ŷ=1)

The end