Statistical Decision Theory
Bayesian and Quasi-Bayesian estimators

Giselle Montamat

Harvard University

Spring 2020

Statistical Decision Theory

Framework to make a decision based on data (e.g., find the “best” estimator under some criteria for what “best” means; decide whether to retain/reward a teacher based on observed teacher value added estimates); criteria to decide what a good decision (e.g., a good estimator; whether to retain/reward a teacher) is.

Ingredients:

- Data: “X”
- Statistical decision: “a”
- Decision function: “$\delta(X)$”
- State of the world: “θ”
- Loss function: “$L(a, \theta)$”
- Statistical model (likelihood): “$f(X|\theta)$”
- Risk function (aka expected loss):

$$R(\delta, \theta) = E_{f(X|\theta)}[L(\delta(X), \theta)] = \int L(\delta(X), \theta) f(X|\theta) dX$$

Statistical Decision Theory

Objective: estimate $\mu(\theta)$ (could be $\mu(\theta) = \theta$) using data X via $\delta(X)$.
(Note: here the decision is to choose an estimator; we'll see another example where the decision is a binary choice).

Loss function $L(a, \theta)$: describes loss that we incur in if we take action a when true parameter value is θ. Note that estimation (“decision”) will be based on data via $\delta(X) = a$, so loss is a function of the data and the true parameter, ie, $L(\delta(X), \theta)$.

Criteria for what makes a “good” $\delta(X)$, for a given θ: the *expected* loss (aka, the risk) has to be small, where the expectation is taken over X given model $f(X|\theta)$ for a given θ.

Quadratic loss

One example of a **loss function** is quadratic loss:

$$L(\delta(X), \theta) = (\delta(X) - \mu(\theta))^2$$

For quadratic loss, the **risk function** is the MSE (mean squared error) and can be expressed in terms of variance and bias:

$$R(\delta, \theta) = E_{f(X|\theta)}\left[(\delta(X) - \mu(\theta))^2\right]$$

$$R(\delta, \theta) = Var_{f(X|\theta)}(\delta(X)) + \left(E_{f(X|\theta)}\left[(\delta(X)\right] - \mu(\theta)\right)^2$$

Exercise: Show that the MSE can be expressed as the sum of variance plus the square of the bias.

Risk function

Exercises: for each of the following exercises, compute the risk function (aka expected loss) based on squared error loss of the suggested decision functions (estimators). Provide intuition about what each decision function is doing and compare them based on this risk.

1) You observe a single data point $X \sim N(\mu, 1)$ and your objective is to estimate μ. You consider the following (family of) decision functions: $\delta(X) = \alpha + \beta X$.
2) You observe N iid data points: $X_i \sim N(\mu, 1)$ and your objective is to estimate μ. You consider the following two decision functions: $\delta_1(X) = X_1$ and $\delta_2(X) = \bar{X}$.

Risk function

3) You observe a single data point $X \sim N(\mu, \sigma)$ where σ is known and your objective is to estimate μ. You consider the following decision functions:

1. $\delta_1(X) = \alpha + \beta X$
2. $\delta_2(X) = 1(X < -\lambda)(X + \lambda) + 1(X > \lambda)(X - \lambda)$
3. $\delta_3(X) = 1(|X| > \kappa)X$

4) You observe a single data point $X \sim N(\mu, I)$ where X and μ are $k \times 1$ and your objective is to estimate μ. You consider the following decision functions:

1. $\delta_1(X) = X$
2. $\delta_2(X) = \left(1 - \frac{k-2}{\sum_{j=1}^{k} X_j^2}\right) X$
3. $\delta_3(X) = max \left\{1 - \frac{k-2}{\sum_{j=1}^{k} X_j^2}, 0\right\} X$

Optimality criteria

Because $R(\delta, \theta)$ depends on θ, we need a criteria to compare decision functions $\delta(X)$ across all possible θ (ie, what decision function generates the "smallest" expected loss across possible values for θ, aka, how to choose an overall "good" rule?).

Optimality criteria:

1. **Admissibility**: $\delta(X)$ is not dominated by another decision rule
2. **Minimax**: $\delta(X)$ has the best possible worst-case performance
3. **Bayes criteria**: $\delta(X)$ has the lowest weighted average risk

Bayes criteria

A δ is better relative to another if a *weighted average of risk over* θ (aka weighted average of expected loss over θ, aka *integrated risk*) is lower. That is, assign weights to different values of θ according to a prior and average the risk across θ using these weights.

Prior: $\pi(\theta)$

Posterior: $f(\theta|X)$

Integrated risk:

$$R(\delta, \pi) = \int R(\delta, \theta)\pi(\theta)d\theta \quad \textbf{(A)}$$

$$R(\delta, \pi) = \int \underbrace{R(\delta, \pi|X)}_{\substack{\text{Posterior}\\ \text{expected loss:}\\ E_{f(\theta|X)}[L(\delta(X),\theta)|X]=\\ \int L(\delta(X),\theta)f(\theta|X)d\theta}} f(X)dX \quad \textbf{(B)}$$

Bayes criteria

The usefulness of being able to express (A) as (B) is that minimizing the integrated risk boils down to minimizing the posterior expected loss by choosing $\delta(X)$ (for a given X). Note that in order to do this, we need to find the posterior $f(\theta|X)$.

Exercise: show that you can express (A) as (B).

Bayes criteria+Quadratic loss

- The optimal decision function $\delta(X)$ under quadratic loss based on Bayes criteria is...the posterior mean!

$$\begin{aligned} R(\delta,\pi|X) &= E_{f(\theta|X)}[L(\delta(X),\theta)|X] \\ &= E_{f(\theta|X)}[(\delta(X)-\mu(\theta))^2|X] \\ &= Var_{f(\theta|X)}[\mu(\theta)|X] + \left[\delta(X) - E_{f(\theta|X)}[\mu(\theta)|X]\right]^2 \end{aligned}$$

$$\Rightarrow \delta^*(X) = \underset{\delta(X)}{arg\,min}\ R(\delta,\pi|X) = E_{f(\theta|X)}[\mu(\theta)|X]$$

Note: under $\delta^*(X)$, we then have $R(\delta,\pi|X) = Var_{f(\theta|X)}[\mu(\theta)|X]$.

Bayes criteria+Quadratic loss

Exercise: you observe a single data point $X \sim N(\theta, 1)$, you have a prior $\theta \sim N(0, \tau^2)$ and you're asked to find the optimal estimator for θ under Bayes criteria with quadratic loss. How does this estimator depend on the variance of the prior?

Hint: Remember normal conjugate distributions:

$$\text{Likelihood: } D|\theta \sim N(\theta, \Sigma)$$

$$\text{Prior: } \theta \sim N(\mu, \Omega)$$

$$\Rightarrow \text{Posterior: } \theta|D = d \sim N(\mu + \Omega(\Sigma + \Omega)^{-1}(d - \mu), \Omega - \Omega(\Omega + \Sigma)^{-1}\Omega)$$

Bayes criteria+Quadratic loss

Example:

- Data: $Z_1, ..., Z_{50}$
- Bernoulli likelihood: $f(Z_i|\theta) = Be(\theta)$ (ie, $P(Z_i = 1|\theta) = \theta$)
- Beta prior: $\pi(\theta) = Beta(\alpha, \beta)$
- Beta posterior:

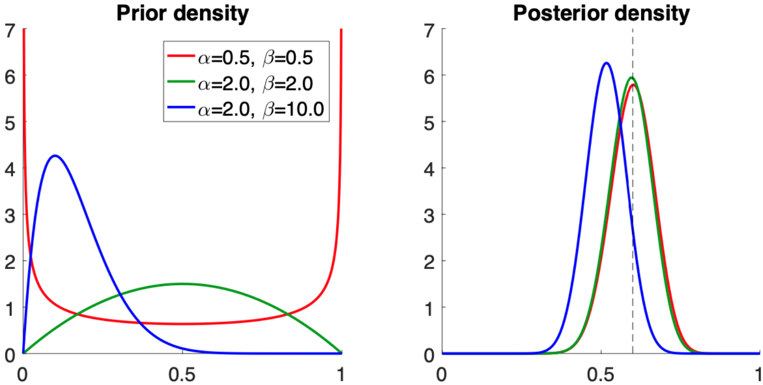
$$f(\theta|Z) = Beta(\alpha + \sum_i Z_i, \beta + \sum_i (1 - Z_i))$$

(See exercise on Conjugate Priors from Section 6).

The optimal estimator of θ based on Bayes criteria and quadratic loss is the posterior mean:

$$E_{f(\theta|Z)}[\theta|Z] = \frac{\alpha + \sum_i Z_i}{\alpha + \sum_i Z_i + \beta + \sum_i (1 - Z_i))} = \frac{\alpha + \sum_i Z_i}{\alpha + \beta + N} = \frac{\frac{\alpha}{N} + \bar{Z}}{\frac{\alpha+\beta}{N} + 1}$$

Bayes criteria+Quadratic loss



$$P(Z = 1 \mid \theta) = \theta, \quad \theta \sim \text{Beta}(\alpha, \beta),$$

$$N = 50, \quad \bar{Z} = 60\%.$$

Bayes criteria+Quadratic loss

As an estimator of $\theta_0 = E_{f(Z_i|\theta=\theta_0)}[Z_i]$ (frequentist perspective), posterior mean can be compared to MLE estimator in terms of bias and variance:

MLE estimator of θ_0:

$$\delta_{MLE}(Z) = arg \min_{\delta(Z)} \; log\left(\Pi_i\{\theta^{Z_i}(1-\theta)^{1-Z_i}\}\right) = \bar{Z}$$

1) It is unbiased:

$$E_{f(Z_i|\theta_0)}[\bar{Z}] = \theta_0$$

2) Variance:

$$Var(\bar{Z}) = \frac{Var(Z_i)}{N} = \frac{\theta_0(1-\theta_0)}{N}$$

Bayes criteria+Quadratic loss

The posterior mean, as an estimator $\delta(Z)$ of θ_0:

$$\delta_{BC}(Z) = E_{f(\theta|Z)}[\theta|Z] = \frac{\frac{\alpha}{N} + \bar{Z}}{\frac{\alpha+\beta}{N} + 1} = \frac{\frac{\alpha}{N}}{\frac{\alpha+\beta}{N} + 1} + \frac{1}{\frac{\alpha+\beta}{N} + 1}\bar{Z}$$

1) It is biased (in finite samples) unless the prior happens to be centered around θ_0:

$$E_{f(Z|\theta_0)}[\delta_{BC}(Z)] = \frac{\frac{\alpha}{N} + E_{f(Z|\theta_0)}[\bar{Z}]}{\frac{\alpha+\beta}{N} + 1} = \frac{\frac{\alpha}{N} + E_{f(Z|\theta_0)}[Z_i]}{\frac{\alpha+\beta}{N} + 1} = \frac{\frac{\alpha}{N} + \theta_0}{\frac{\alpha+\beta}{N} + 1}$$

Note that $\frac{\frac{\alpha}{N}+\theta_0}{\frac{\alpha+\beta}{N}+1} = \theta_0$ if $\frac{\alpha}{\alpha+\beta} = \theta_0$ (ie, the prior has mean θ_0).

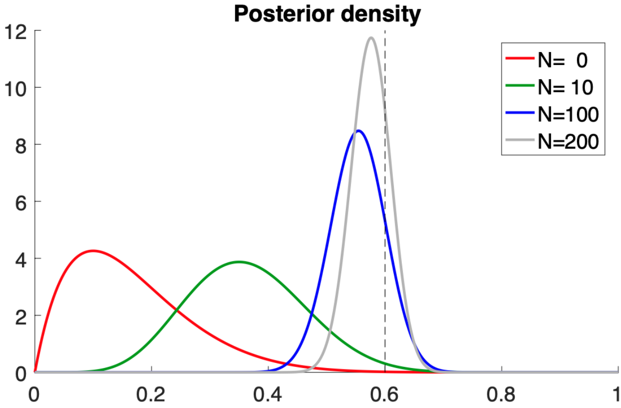
2) Its variance (in finite samples) is lower than that of the MLE estimator:

$$Var(\delta_{BC}(Z)) = \left(\frac{1}{\frac{\alpha+\beta}{N} + 1}\right)^2 \frac{Var(Z_i)}{N} = \left(\frac{1}{\frac{\alpha+\beta}{N} + 1}\right)^2 \frac{\theta_0(1-\theta_0)}{N}$$

Bayes criteria+Quadratic loss

The bias is reduced as N goes to infinity because posterior distribution allows data to swamp the prior:

$$\lim_{N\to\infty} \delta_{BC}(Z) = \lim_{N\to\infty} \bar{Z} = E_{f(Z_i|\theta_0)}[Z_i] = \theta_0$$



$P(Z = 1 \mid \theta) = \theta, \quad \theta \sim \text{Beta}(2, 10), \quad \bar{Z} = 60\%.$

Bayes criteria+Absolute loss

- The optimal decision rule $\delta(X)$ under absolute error loss based on Bayes criteria is...the posterior median!

Exercise: you observe a single data point $X \sim f(X|\theta)$, you have some prior $\pi(\theta)$ and you are asked to estimate $\mu(\theta)$ (a real-valued function of θ) under Bayes criteria based on the following loss function:

$$L(\delta(X), \theta) = (\mu(\theta) - \delta(X))(\tau - 1(\mu(\theta) \leq \delta(X)))$$

In particular, what is the estimator when $\tau = 0.5$?

Bayes criteria

Bayes criteria $\Leftrightarrow$ Admissibility criteria

$\Rightarrow$ If the risk function is continuous in θ for all δ, and the prior is everywhere positive, then the Bayes decision rule is admissible.

$\Leftarrow$ *Complete Class Theorem*: All admissible rules are Bayes decision rules (aka optimal according to Bayes criteria) under some prior, if certain conditions are satisfied.

Bayesian estimators: properties

If we take the posterior mean of $f(\theta|X)$ as estimator $\hat{\theta}$ of θ_0 (frequentist perspective), natural question is: what are the properties of this estimator?

- Small sampling distribution? Biased? Variance?

  (Note frequentist nature of this question: if one could draw data repeated times from likelihood, and true parameter is θ_0, what would $\hat{\theta}$ look like?)

  Variance-bias trade-off: the posterior mean will generally be a biased estimator (in finite samples) unless the prior happens to be centered around the truth. But the posterior mean lowers the variance, and also allows data to swamp prior (and thus eliminate bias) as N goes to infinity.

- Consistency and asymptotic distribution as $N \to \infty$?

  Berstein-von Mises: the posterior distribution concentrates around the MLE of θ_0 as $N \to \infty$.

Bayesian estimators: properties

First, a reminder of the asymptotic normality result for MLE:

$$\sqrt{N}(\hat{\theta}_{MLE} - \theta_0) \xrightarrow{d} N\left(0, H^{-1} J H^{-1}\right)$$

$$\text{Hessian} : H = E\left[\frac{\partial^2 log(f(D|\theta_0))}{\partial\theta\partial\theta'}\right]$$

$$\text{Jacobian} : J = E\left[\frac{\partial log(f(D|\theta_0))}{\partial\theta}\left(\frac{\partial log(f(D|\theta_0))}{\partial\theta}\right)'\right]$$

If likelihood correctly specified, Information Matrix Equality holds: $H = -J$.

$$\sqrt{N}(\hat{\theta}_{MLE} - \theta_0) \xrightarrow{d} N\left(0, J^{-1}\right)$$

Bayesian estimators: properties

Berstein-von Mises: the posterior distribution concentrates around the MLE of θ as $N \to \infty$:

$$\theta|\hat{\theta}_{MLE} \overset{approx}{\sim} N\left(\hat{\theta}_{MLE}, \frac{1}{N}J^{-1}\right)$$

Note: model (likelihood) must be correctly specified.

Remember the intuition from the previous exercise: the posterior mean is biased in finite samples if prior is far from the “truth” (θ_0). But! As $N \to \infty$ data predominates over prior in determining the posterior distribution; the posterior distribution centers around the MLE estimator, which is unbiased. The posterior mean is a consistent estimator for θ_0. Moreover, Bayesian 95% credible intervals are asymptotic valid frequentist confidence intervals.

Side note: Quasi-Bayes

Posterior distribution: $\theta \sim f(\theta|D) = \frac{f(D|\theta)\pi(\theta)}{f(D)} = \frac{f(D|\theta)\pi(\theta)}{\int f(D|\theta)\pi(\theta)d\theta}$

We can re-write:

$$f(D|\theta) = exp\left(N\frac{1}{N}log(f(D|\theta))\right) = exp\left(\underbrace{N\frac{1}{N}\sum_i log(f(D_i|\theta))}_{l(\theta)}\right)$$

Posterior distribution: $\theta \sim f(\theta|D) = \frac{exp\left(N\frac{1}{N}\sum_i log(f(D_i|\theta))\right)\pi(\theta)}{\int exp\left(N\frac{1}{N}\sum_i log(f(D_i|\theta))\right)\pi(\theta)d\theta}$

Side note: Quasi-Bayes

Posterior distribution: $\theta \sim f(\theta|D) = \dfrac{exp\left(N\frac{1}{N}\sum_i log(f(D_i|\theta))\right)\pi(\theta)}{\int exp\left(N\frac{1}{N}\sum_i log(f(D_i|\theta))\right)\pi(\theta)d\theta}$

Quasi-posterior distribution: $\theta \sim f^Q(\theta|D) = \dfrac{exp\left(N(-\hat{Q}_N(\theta))\right)\pi(\theta)}{\int exp\left(N(-\hat{Q}_N(\theta))\right)\pi(\theta)d\theta}$

A reminder about MLE: Maximum Likelihood Estimator

A reminder of the asymptotic normality result for MLE:

$$\sqrt{N}(\hat{\theta}_{MLE} - \theta_0) \xrightarrow{d} N\left(0, H^{-1} J H^{-1}\right)$$

$$\text{Hessian} : H = E\left[\frac{\partial^2 log(f(D|\theta_0))}{\partial\theta\partial\theta'}\right]$$

$$\text{Jacobian} : J = E\left[\frac{\partial log(f(D|\theta_0))}{\partial\theta}\left(\frac{\partial log(f(D|\theta_0))}{\partial\theta}\right)'\right]$$

If likelihood correctly specified, Information Matrix Equality holds: $H = -J$.

$$\sqrt{N}(\hat{\theta}_{MLE} - \theta_0) \xrightarrow{d} N\left(0, J^{-1}\right)$$

A reminder about EE: Extremum Estimator

A reminder of the asymptotic normality result for EE:

$$\sqrt{N}(\hat{\theta}_{EE} - \theta_0) \xrightarrow{d} N\left(0, H^{-1}\Omega H^{-1}\right)$$

$$\text{Hessian : } H = \frac{\partial^2 Q(\theta_0)}{\partial\theta\partial\theta'}$$

$$\Omega = Var_{asymp}\left(\sqrt{N}\frac{\partial}{\partial\theta}\hat{Q}_n(\theta_0)\right)$$

If Generalized Information Matrix Equality holds: $H = \Omega$.

$$\sqrt{N}(\hat{\theta}_{EE} - \theta_0) \xrightarrow{d} N\left(0, H^{-1}\right)$$

(Quasi) Bayesian procedures & large sample interpretations

- **Berstein-von Mises**: under correct specification of likelihood, as $N \to \infty$:

$$f(\theta|D) \overset{approx}{\sim} N\left(\hat{\theta}_{MLE}, \frac{1}{N}J^{-1}\right)$$

  So: Bayesian procedures have a frequentist intepretation in large samples.

- **Cherenozhukov and Hong**:

$$f^{Q}(\theta|D) \overset{approx}{\sim} N\left(\hat{\theta}_{EE}, \frac{1}{N}H^{-1}\right)$$

  So: Quasi-Bayesian procedures have a frequentist interpretation in large samples.

Note: if (generalized) information equality holds, can use posterior standard deviation, multiplied by N, as estimate of asymptotic standard deviation of MLE/EE estimates (aka, the frequentist standard errors).

(Quasi) Bayesian procedures & large sample interpretations

Exercise: Pset 8 - Exercise 1 and Pset 9 - Exercise 1 asked to to implement quasi-Bayesian procedures in these two contexts:

- GMM objective function based on method of simulated moments:

$$E\left[m(Y_i) - \frac{1}{S}\sum m(Y_i^s(\theta_0))\right] = 0 \Rightarrow \frac{1}{N}\sum\left[m(Y_i) - \frac{1}{S}\sum m(Y_i^s(\theta))\right] = 0$$

$$\hat{Q}_N(\theta) = \left(\frac{1}{N}\sum\left[m(Y_i) - \frac{1}{S}\sum m(Y_i^s(\theta))\right]\right)' \hat{W} \left(\frac{1}{N}\sum\left[m(Y_i) - \frac{1}{S}\sum m(Y_i^s(\theta))\right]\right)$$

- GMM objective function based on quantile IV:

$$E[(\tau - 1\{Y_i < \alpha_\tau + \beta_\tau P_i\})Z_i] = 0 \Rightarrow \frac{1}{N}\sum(\tau - 1\{Y_i < \alpha_\tau + \beta_\tau P_i\})Z_i = 0$$

$$\hat{Q}_N(\theta) = \left(\frac{1}{N}\sum[(\tau - 1\{Y_i < \alpha_\tau + \beta_\tau P_i\})Z_i]\right)' \hat{W} \left(\frac{1}{N}\sum[(\tau - 1\{Y_i < \alpha_\tau + \beta_\tau P_i\})Z_i]\right)$$

For $\hat{W}$, you're told to use the continuously updating GMM objective function approach, so $\hat{W} = \hat{Var}(g(D_i, \theta))$.

Statistical decision theory: another example

Exercise: Pset 8, exercise 2 asks you about decision-making with regard to a teacher j's value added, θ_j. Specifically, the decision is a binary action $a_j \in \{0, 1\}$ (example: retain/don't retain, reward/don't reward, where 1 corresponds to retaining or rewarding) based on observed estimates $\hat{\theta}_j$ (ie, you make a decision based on an estimate of θ_j).

Goal: reward/retain teachers with θ_j above some threshold value θ^*.

Possible loss functions considered:

1. You count a loss of 1 if you decide $a_j = 1$ for a teacher j that has $\theta_j < \theta^*$, and if $a_j = 0$ for a teacher j that has $\theta_j > \theta^*$:

$$L(a, \theta) = \sum_j (a_j 1\{\theta_j < \theta^*\} + (1 - a_j) 1\{\theta_j > \theta^*\})$$

2. You count a loss of $\theta^* - \theta_j$ if you decide $a_j = 1$ for a teacher j that has $\theta_j < \theta^*$, and a loss of $\theta_j - \theta^*$ if $a_j = 0$ for a teacher j that has $\theta_j > \theta^*$:

$$L(a, \theta) = \sum_j (a_j max\{\theta^* - \theta_j, 0\} + (1 - a_j) max\{\theta_j - \theta^*, 0\})$$

Statistical decision theory: another example

Data: for each teacher $j = 1, ..., J$ you observe an estimate $\hat{\theta}_j$.

(Moreover, you are given observations for covariates X_j and the standard error for the estimated value added σ_j; in the notation that follows I suppress the conditioning on X_j and σ_j to simplify notation).

To make a optimal decision according to Bayes criteria, need to minimize posterior risk (aka posterior expected loss). This requires finding a posterior density for θ_j.

Ingredients:

- Likelihood (for $\hat{\theta}_j$ conditional on θ_j): $\hat{\theta}_j \sim N(\theta_j, \sigma_j) = f(\hat{\theta}_j|\theta_j)$
- Prior (for θ_j): $\theta_j \sim \pi(\theta_j|\beta)$
- Hyperprior (for β): $\beta \sim \pi(\beta)$
- Likelihood (for $\hat{\theta}_j$ conditional on β): $\hat{\theta}_j \sim f(\hat{\theta}_j|\beta)$

Statistical decision theory: another example

You are asked to find the posterior desity of β, the (joint) posterior density of both θ and hyperparameter β, and the posterior density of θ_j.

$$f(\beta|D) = \frac{f(D|\beta)\pi(\beta)}{\int f(D|\beta)\pi(\beta)d\beta}$$

$$f(\theta,\beta|D) = \frac{f(D|\beta)\pi(\theta|\beta)\pi(\beta)}{\int\int f(D|\beta)\pi(\theta|\beta)\pi(\beta)d\theta d\beta} = \frac{f(D,\theta|\beta)\pi(\beta)}{\int\int f(D,\theta|\beta)\pi(\beta)d\theta d\beta}$$

(Remember Hierarchical Bayes from past section?)

$$f(\theta_j|D) = \int\int f(\theta,\beta|D)d\beta d\theta_{-j}$$

So we need to find $f(D|\beta)$ and $f(D,\theta|\beta)$. Where in our case $D = \hat{\theta}$.

Statistical decision theory: another example

$$f(\hat{\theta}|\beta) = \Pi_j f(\hat{\theta}_j|\beta) = \Pi_j \int f(\hat{\theta}_j, \theta_j|\beta) d\theta_j = \Pi_j \int f(\hat{\theta}_j|\theta_j)\pi(\theta_j|\beta) d\theta_j$$

$$f(\hat{\theta}, \theta|\beta) = \Pi_j f(\hat{\theta}_j, \theta_j|\beta) = \Pi_j f(\hat{\theta}_j|\theta_j)\pi(\theta_j|\beta)$$

Once we've found the posterior for θ_j, let's find the optimal decision by minimizing the posterior expected loss, i.e.: $E_{f(\theta|\hat{\theta})}[L(\delta(\hat{\theta}), \theta)|\hat{\theta}]$.

- For loss function 1:

$$E_{f(\theta|\hat{\theta})}[L(\delta(\hat{\theta}), \theta)|\hat{\theta}] =$$

$$\sum_j \left(\delta_j(\hat{\theta}) E_{f(\theta|\hat{\theta})}[1\{\theta_j < \theta^*\}|\hat{\theta}] + (1 - \delta_j(\hat{\theta})) E_{f(\theta|\hat{\theta})}[1\{\theta_j > \theta^*\}|\hat{\theta}]\right)$$

$$= \sum_j \left(\delta_j(\hat{\theta}) P[\theta_j < \theta^*|\hat{\theta}] + (1 - \delta_j(\hat{\theta})) P[\theta_j > \theta^*|\hat{\theta}]\right)$$

$$\delta_j^*(\hat{\theta}) = \begin{cases} 1 \;\; if \;\; P[\theta_j < \theta^*|\hat{\theta}] < P[\theta_j > \theta^*|\hat{\theta}] \\ 0 \;\; if \;\; P[\theta_j < \theta^*|\hat{\theta}] > P[\theta_j > \theta^*|\hat{\theta}] \end{cases}$$

I.e., reward the teachers for whom the posterior probability of being above the threshold exceeds that of being below.

Statistical decision theory: another example

- For loss function 2:

$$E_{f(\theta|\hat{\theta})}[L(\delta(\hat{\theta}),\theta)|\hat{\theta}] =$$

$$\sum_j \left(\delta_j(\hat{\theta}) E_{f(\theta|\hat{\theta})}[max\{\theta^* - \theta_j, 0\}|\hat{\theta}] + (1-\delta_j(\hat{\theta})) E_{f(\theta|\hat{\theta})}[max\{\theta_j - \theta^*, 0\}|\hat{\theta}]\right)$$

$$\delta_j^*(\hat{\theta}) = \begin{cases} 1 \;\; if \;\; E_{f(\theta|\hat{\theta})}[max\{\theta^* - \theta_j, 0\}|\hat{\theta}] < E_{f(\theta|\hat{\theta})}[max\{\theta_j - \theta^*, 0\}|\hat{\theta}] \\ 0 \;\; if \;\; E_{f(\theta|\hat{\theta})}[max\{\theta^* - \theta_j, 0\}|\hat{\theta}] > E_{f(\theta|\hat{\theta})}[max\{\theta_j - \theta^*, 0\}|\hat{\theta}] \end{cases}$$

I.e, reward teachers when the posterior expected amount by which they exceed the threshold is larger than that by which they fall short.