- Suppose the prior distribution of K is Binomial$(21, p)$. Then

r.v.

probability of generating a defective item. assume known

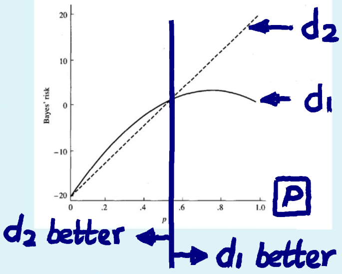
$$E(K) = 21p, \quad Var(K) = 21p(1-p), \quad E(K^2) = 21p(1-p) + (21p)^2.$$

- The Bayes risks of d_1 and d_2 are

$$
\begin{aligned}
B(d_1) &= -20 + 3\cancel{0} \times E(K) - (2/21)E(K^2) \\
&= -20 + 3 \times 21p - (2/21)[21p(1-p) + (21p)^2] \\
&= -20 + 61p - 40p^2 \quad \leftarrow \text{quadratic polynomial} \\
B(d_2) &= -20 + (40/21) \times 21p \\
&= 40p - 20 \quad \leftarrow \text{linear polynomial}
\end{aligned}
$$

θ is gone

FIGURE 15.2
Bayes risk as a function of p: The solid line corresponds to d_1 and the dotted line to d_2.



- **Figure 15.2**: Bayes risks versus p, d_2 has a smaller Bayes risk as long as $p \le 0.5$. (If the product is fairly reliable, may prefer d_2.)

❖ **Reading**: textbook (2nd ed.), 15.1, 15.2, 15.2.1 ← 6/2

• Posterior Analysis --- A simple method for finding Bayes rule

Definition 10.3 (Posterior Distribution and Posterior Risk, 2nd Ed., TBp.578-579)

- In Bayesian procedures, we have
  - Θ: a random variable with a pdf/pmf $g_\Theta(\theta)$
  - $g_\Theta(\theta)$: prior distribution of Θ ← 事前分配
  - $f_{X|\Theta}(x|\theta)$: pdf/pmf of X, conditional on the value θ of Θ

θ a fixed value

In estimation (CH8) and testing (CH9), the joint pdf/pmf of X (data) not observed

Conditioned on Θ = θ
random variable Θ
→ a fixed value θ

- Joint distribution of X and Θ is

$$f_{X,\Theta}(x,\theta) \overset{\uparrow}{=} f_{X|\Theta}(x|\theta)\, g_\Theta(\theta)$$

multiplication law (LN, CH1~6, p.23)

$= \dfrac{f_{X,\Theta}(x,\theta)}{g_\Theta(\theta)}$ ← the conditional distribution under Bayesian approach (Θ: random) is the joint pdf/pmf of X discussed in CH8 & 9 (θ: fixed) cf.

- Marginal distribution of X is

$$f_X(x) \overset{\downarrow}{=} \begin{cases} \int f_{X|\Theta}(x|\theta) g_\Theta(\theta)\, d\theta, & \text{if } \Theta \text{ is continuous} \\ \sum_{\theta_i} f(x\,|\,\theta_i) g(\theta_i), & \text{if } \Theta \text{ is discrete} \end{cases}$$

law of total probability (LN, CH1~6, p.23)

cf distribution of data

- Conditional distribution of Θ given $X = x$ is

$$h_{\Theta|X}(\theta|x) \overset{\uparrow}{=} \frac{f_{X,\Theta}(x,\theta)}{f_X(x)} = \begin{cases} \boxed{\dfrac{f_{X|\Theta}(x|\theta) g_\Theta(\theta)}{\int f_{X|\Theta}(x|\theta) g_\Theta(\theta)\, d\theta}}, & \text{if } \Theta \text{ is continuous} \\ \boxed{\dfrac{f_{X|\Theta}(x|\theta) g_\Theta(\theta)}{\sum_\theta f_{X|\Theta}(x|\theta) g_\Theta(\theta)}}, & \text{if } \Theta \text{ is discrete} \end{cases}$$

• Risk function (LNp.3)
• Bayes risk (LNp.6)

new infomation

Bayes' rule (LN, CH1~6, p.23) ← update g_Θ to $h_{\Theta|X}$

core component in Bayesian inference

which is also called the **posterior distribution** of Θ. ← 事後分配

cf.

- Given observed $X = x$, define **posterior risk** of an action a $(= d(x))$ as

$$E_{\Theta|X=x}\left[l(\Theta, \overset{d(x)}{a})\right] = \begin{cases} \int l(\theta, a)\, h(\theta|x)\, d\theta, & \text{if } \Theta \text{ is continuous} \\ \sum_{\theta_i} l(\theta_i, a)\, h(\theta_i|x), & \text{if } \Theta \text{ is discrete} \end{cases}$$

a function of a & x only

average loss w.r.t. Θ
w.r.t. x cf.

In Bayesian approach, the understanding about θ is always presented by distribution of θ

Information about the parameter

contains prior information about the parameter

before data is observed
$g_\Theta(\theta)$

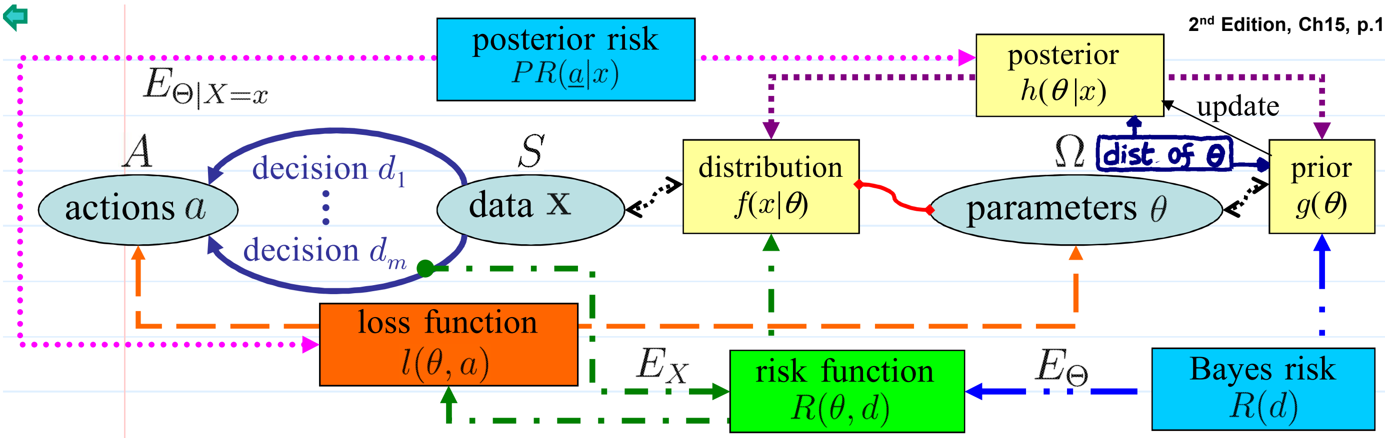
X = x
observe data x

update

(x brings in some information about the parameter)

after data is observed
$h_{\Theta|X}(\theta|x)$

contains prior information and the information from x



posterior risk $PR(\underline{a}|x)$

posterior $h(\theta|x)$

$E_{\Theta|X=x}$

A actions a

decision d_1 ⋮ decision d_m

S data X

distribution $f(x|\theta)$

Ω parameters θ

dist. of θ

update

prior $g(\theta)$

loss function $l(\theta, a)$

E_X

risk function $R(\theta, d)$

E_Θ

Bayes risk $R(d)$

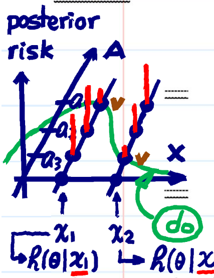
Theorem 10.1 (2nd Ed., TBp.579)

Suppose that there is a decision function $\underset{a}{\underset{\parallel}{d_0(x)}}$ that minimizes the posterior risk for each x. Then $d_0(x)$ is a Bayes rule.

↑ PR(a|x)

d(x)

Proof. (for continuous case) The Bayes risk of a decision function d is



$$\begin{aligned}
B(d) &= \underbrace{E_\Theta}_{\text{prior } g_\Theta(\theta)}[R(\Theta,d)] = E_\Theta\left\{\underbrace{E_{X|\Theta}}_{f_{X|\Theta}(x|\theta)}\left[l(\Theta, d(X)) \,\middle|\, \Theta\right]\right\} \\
&= \int\left[\int l(\theta, d(x))\, f_{X|\Theta}(x|\theta)\,dx\right] g_\Theta(\theta)\,d\theta = \int\int l(\theta, d(x))\, f_{X,\Theta}(x,\theta)\,dx\,d\theta \\
&= \int\left[\int l(\theta, d(x))\, \underbrace{h_{\Theta|X}(\theta|x)}_{\text{posterior dist.}}\,d\theta\right] f_X(x)\,dx = \int \underbrace{E_{\Theta|X=x}\left[l(\Theta, \overset{a}{\overset{\parallel}{d(x)}})\right]}_{\text{posterior risk (LNp.11)}} f_X(x)\,dx
\end{aligned}$$

$f_{X,\Theta}(x,\theta) = f_{X|\Theta}(x|\theta)\,g_\Theta(\theta) = h_{\Theta|X}(\theta|x)\,f_X(x)$

multiplication law

a ← an action

Since $f_X(x)$ is nonnegative, $B(d)$ is minimized by choosing $d(x) = d_0(x)$.

Algorithm for finding the Bayes rule (2nd Ed., TBp.579-580)

Step 1 : Calculate posterior distribution $h(\theta|x)$ for each x.

Step 2 : For each x, fix $X = x$. For each action a, calculate the posterior risk:

$$\boxed{E_{\Theta|X=x}}\left[l(\Theta, a)\right] = \begin{cases} \int l(\theta, a)\, h(\theta|x)\,d\theta, & \text{in the continuous case} \\ \sum_{\theta_i} l(\theta_i, a)\, h(\theta_i|x), & \text{in the discrete case} \end{cases}$$

Step 3 : The action $a^*(x)$ that minimizes the posterior risk is the Bayes rule.

Example 10.6 (steel section (cont.), 2nd Ed., TBp. 580, LNp.6~8)

- prior distribution: $g(\theta_1) = 0.8$, $g(\theta_2) = 0.2$ — update
- Suppose that we observe $X = x_2 = 45$, the posterior distribution is

$$\begin{aligned}
h(\theta_1|x_2) &= \frac{f(x_2|\theta_1)g(\theta_1)}{\underbrace{\sum_{i=1}^{2} f(x_2|\theta_i)g(\theta_i)}_{f_X(x_2)=}} = \frac{0.3\times 0.8}{0.3\times 0.8 + 0.2\times 0.2} = 0.86 \\
h(\theta_2|x_2) &= 1 - h(\theta_1|x_2) = 0.14
\end{aligned}$$

- The posterior risk (PR) for a_1 and a_2 are

$$\begin{aligned}
\mathrm{PR}(a_1|x_2) &= l(\theta_1, a_1)h(\theta_1|x_2) + l(\theta_2, a_1)h(\theta_2|x_2) = 0 + 400\times 0.14 = 56 \\
\mathrm{PR}(a_2|x_2) &= l(\theta_1, a_2)h(\theta_1|x_2) + l(\theta_2, a_2)h(\theta_2|x_2) = 100\times 0.86 + 0 = 86
\end{aligned}$$

cf.

- a_1 has the smallest posterior risk, and is the Bayes rule for x_2.
- Apply the procedure to any other possible observed data ($\overset{=40}{x_1}$ and $\overset{=50}{x_3}$) ← (exercise)

Check the d₃ in LNp.8 — a₁ ← a₂ ←

❖ **Reading**: textbook (2nd ed.), 15.2.2

• Application of Decision Theory: Estimation

← Recall Ex10.3 (LNp.4)

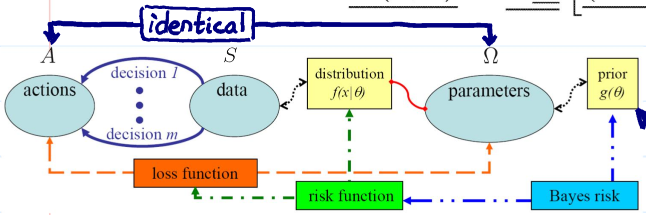
Estimation theory can be cast in a decision theoretic framework.

- action space A = parameter space Ω. A
- a decision function $d(X)$ ($=\hat{\theta}$: $S \to \overset{A}{\overset{/\!/}{\Omega}}$) is an estimator of θ
- square error loss: $l(\theta, d(X)) = [\theta - d(X)]^2 = (\theta - \hat{\theta})^2$ ← L^2-norm ← cf.

  (other loss functions, e.g., $|\theta - \hat{\theta}|$, are allowable) ← check Thm10.3 (LNp.16)
- Then the risk is $R(\theta, d) = E_X\left[(\theta - d(X))^2\right] = E\left[(\theta - \hat{\theta})^2\right]$.= MSE

Alternative: L^p-norm $= |\theta - \hat{\theta}|^p$

minimax estimator $\left(\min_{d} \max_{\theta\in\Omega} MSE\right)$



Q: what if a prior is available?

UMVUE ← cf.

Theorem 10.2 (Bayes rule for Estimation under Squared Error Loss, 2nd Ed., TBp.584)

- The Bayes rule minimizes the posterior risk, which is

MSE = variance + (bias)² (LN, CH1~6, p42, item 5)

$$E_{\Theta|X}[(\Theta - \hat{\theta})^2 | X = x] = \underbrace{Var_{\Theta|X}(\Theta \,|\, X = x)}_{\text{irrelevant to } \hat{\theta}} + \underbrace{[E_{\Theta|X}(\Theta \,|\, X = x) - \hat{\theta}]^2}_{\text{minimized by } \hat{\theta} = E_{\Theta|X}(\Theta \,|\, X = x)}$$

posterior distribution

r.v.

a constant (action) when conditioned on X = x

decision function

conditional mean

best predictor in G3 (LN, CH1~6, p.54) ←cf.→

- Thus, Bayes rule is

mean is the best predictor of a r.v. under MSE → reasonable? →

$$\hat{\theta} = \begin{cases} \int \theta\, h(\theta \,|\, x)\, d\theta, & \text{in the continuous case} \\ \sum_{\theta_i} \theta_i\, h(\theta_i | x), & \text{in the discrete case} \end{cases}$$

- In the case of squared error loss, the Bayes estimator (i.e., Bayes rule) is the mean of the posterior distribution. 6/4

Example 10.7 (Throw a coin once, Bayes estimator, 2nd Ed., TBp. 584-585)

- A biased coin is thrown once. Estimate θ = probability of heads.
- Suppose that we have no idea how biased the coin is $\Rightarrow$ for θ, can use uniform prior: $g(\theta) = 1$, $0 \le \theta \le 1$.
- Let

a vague prior

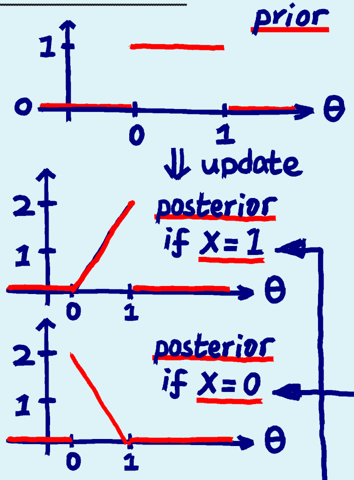
Data → $X = \begin{cases} 1, & \text{if a head appears} \\ 0, & \text{if a tail appears} \end{cases}$

Then the distribution of X given θ is Bernoulli(θ):

coditional

This is the pmf of X in CH8 & CH9

$$f(x|\theta) = \begin{cases} \theta, & x = 1 \\ 1-\theta, & x = 0 \end{cases}$$



- The posterior distribution is

conditional θ|X → joint (θ,X) → marginal X → g(θ)

$$h(\theta|x) = \frac{f(x \,|\, \theta) \times 1}{\int_0^1 f(x \,|\, \theta) \times 1\, d\theta} = \begin{cases} \frac{\theta}{\int_0^1 \theta\, d\theta} = 2\theta, & x = 1 \\ \frac{1-\theta}{\int_0^1 1-\theta\, d\theta} = 2(1-\theta), & x = 0 \end{cases}$$