$P(C_j) \rightsquigarrow \theta \mapsto \pi(\theta)$ prior distribution

$P(C_j|B) \rightsquigarrow \theta \mapsto \pi_{post}(\theta|Y=y)$ posterior distribution

$P(B|C_j) \rightsquigarrow y \mapsto f_Y(y|\theta)$

$\sum_K P(B|C_K)P(C_K) \rightsquigarrow y \mapsto \int f_Y(y|\theta)\pi(\theta)d\theta$

Suppose $Y = \sum_{j=1}^{N} X_j \sim$ binomial (N, p), $\theta \in [0,1]$

$\pi(p) \sim$ beta (α, β) (α, β fixed)

$$\Rightarrow \pi_{post}(p|y) = \frac{\text{binomial}(y|N,p)\,\text{beta}(p|\alpha,\beta)}{\int_0^1 \text{binomial}(y|N,s)\,\text{beta}(s|\alpha,\beta)\,ds} \quad \} ① \ / \ } ②$$

$$① = \binom{N}{y} p^y (1-p)^{N-y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

$$= \binom{N}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{y+\alpha-1}(1-p)^{N-y+\beta-1}$$

Gamma function:

$$② = \int ① \, ds = \binom{N}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(y+\alpha)\Gamma(N-y+\beta)}{\Gamma(N+\alpha+\beta)}$$

$$\Rightarrow \pi_{post}(p|y) = \frac{①}{②} = \frac{\Gamma(N+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(N-y+\beta)} p^{y+\alpha-1}(1-p)^{N-y+\beta-1}$$

$$= \text{beta}(y+\alpha, N-y+\beta)(p)$$

By taking expectation of the prior and posterior distributions

$$E_{prior}(p) = \frac{\alpha}{\alpha+\beta}$$

$$E_{post}(p) = \frac{y+\alpha}{N+\alpha+\beta} = \frac{N}{\alpha+\beta+N}\frac{y}{N} + \frac{\alpha+\beta}{\alpha+\beta+N}\boxed{\frac{\alpha}{\alpha+\beta}}$$

→ estimation on p without any prior idea

The posterior expectation is a linear combination of prior expectation and experimental expectation.

Remark: As N becoming large, the prior expectation is becoming less and less importa

Remark: The linear combination is not an accident;

⇝ notion of conjugate family of a distribution.

IV.2 Evaluating estimators

Different estimators for θ can give you different values for θ.

Which one should we choose?

2.1 Mean square error

Framework $\underline{X} = (X_1, \cdots, X_N)$ with $X_j = X \sim f(\cdot|\theta)$.

Let $W = W(\underline{X})$ be an estimator ($\equiv$ statistic) for θ.

Def. The mean square error (MSE) of W is defined by

$$E((W-\theta)^2)$$ ← still a function of θ

⚠ We made the choice of $s \mapsto s^2$. Later there will be a generation.

Observe that

$$E((W-\theta)^2) = E((W-E(W)+E(W)-\theta)^2)$$

$$= E((W-E(W))^2) + \underbrace{E(2(W-E(W))(E(W)-\theta))}_{=0} + E((E(W)-\theta)^2)$$

($E(W)-\theta$: constant; $(E(W)-\theta)^2$: constant)

$$= Var_\theta(W) - (E_\theta(W)-\theta)^2$$

Def. If W is an estimator for θ, we set

$$Bias_\theta(W) := E_\theta(W) - \theta$$

If $Bias_\theta(W) = 0$ we say that W is unbiased.

$$\Rightarrow E_\theta((W-\theta)^2) = Var_\theta(W) + (Bias_\theta(W))^2$$

Remark: In Section II.1, we saw that if $E(X) = \mu$ and $Var(X) = \sigma^2$ then

$$E(\overline{X}) = \mu \text{ and } E(S^2) = \sigma^2$$

($\overline{X}$: sample mean; S^2: sample variance)

$\Rightarrow E(\overline{X})$ and $E(S^2)$ are unbounded estimators for μ and σ^2 resp.

Now if $X \sim n(\mu, \sigma^2)$ then $E((\overline{X}-\mu)^2) = Var(\overline{X}) = \frac{\sigma^2}{N}$

and $E((S^2-\sigma^2)^2) = \frac{e}{N-1}\sigma^4$

Question: Can we find better (with a smaller MSE) estimator for μ and σ^2?

Remark: for $X \sim n(\mu, \sigma^2)$ consider

$$\widetilde{S}^2 := \frac{1}{N}\sum_{j=1}^{N}(X_j - \overline{X})^2 = \frac{N-1}{N}S^2$$

$\Rightarrow E(\widetilde{S}^2) = \frac{N-1}{N}\sigma^2$ $\Rightarrow \widetilde{S}^2$ is biased

But $E((\widetilde{S}^2-\sigma^2)^2) = \frac{2N-1}{N^2}\sigma^4 < \frac{2}{N-1}\sigma^4$

$\Rightarrow \widetilde{S}^2$ has a lower MSE, despite being biased.

$\rightsquigarrow$ Finding a better estimator is not a simple question.

We shall consider only unbiased estimators.

Def. An estimator W^* for θ (parameter with value unknown) is the best unbiased estimator for θ if $\mathrm{Var}_\theta(W^*) \leq \mathrm{Var}_\theta(W)$ for any estimator W for θ for any θ (value of parameter θ) with W^* and W unbiased.

Thm. (7.3.19) The best unbiased estimator is unique if it exists.

But it does not tell you how to find it.

Question: Can we get $\mathrm{Var}_\theta(W) = 0$?

Thm: (7.3.9 + 7.3.10) (Gramer-Rao inequality) (based on Cauchy-Schwartz inequality)

If $\frac{d}{d\theta} E_\theta(W) \equiv \frac{d}{d\theta} \int_{\mathbb{R}^N} W(\underline{x}) f_{\underline{X}}(\underline{x}|\theta)\, d\underline{x} = \int_{\mathbb{R}^N} \frac{\partial}{\partial\theta}[W(\underline{x}) f_{\underline{X}}(\underline{x}|\theta)]\, d\underline{x}$

and $\mathrm{Var}_\theta(W) < \infty$ then

$$\mathrm{Var}_\theta(W) \geq \underbrace{\frac{\left(\frac{d}{d\theta} E_\theta(W)\right)^2}{N E_\theta\left[\left(\frac{\partial}{\partial\theta} \ln f_{\underline{X}}(\cdot|\theta)\right)^2\right]}}_{\text{information number or Fisher information}} \Big\} = 1 \quad (\text{with } \frac{d}{d\theta} E_\theta(W) = \theta)$$

If $\mathrm{Var}_\theta(W)$ saturates (satisfies equality in) this inequality,

then W is the best unbiased estimater for any value of θ.

⚠ Otherwise for two W the $\mathrm{Var}_\theta(W)$ may cross.

Remark: The following relation holds:

If $\frac{d}{d\theta} E\left(\frac{\partial}{\partial\theta} \ln f(\cdot|\theta)\right) = \int \frac{\partial}{\partial\theta}\left[\frac{\partial}{\partial\theta} \ln f(\cdot|\theta)\right] f(x|\theta)\, dx$ then

$$E_\theta\left[\left(\frac{\partial}{\partial\theta} \ln f(\cdot|\theta)\right)^2\right] = -E_\theta\left(\frac{\partial^2}{\partial\theta^2} \ln f(\cdot|\theta)\right)$$

2.2 Loss function optimality

So far, $MSE = E((W-\theta)^2) = E(\mathcal{L}(W,\theta))$ with $\mathcal{L}(s,\theta) = (s-\theta)^2$

but other function $\mathcal{L}$ = loss function could be chosen. For example:

- $\mathcal{L}(s,\theta) = |s-\theta|$
- $\mathcal{L}(s,\theta) = \begin{cases} (s-\theta)^2 & \text{for } s<\theta \\ 10(s-\theta)^2 & \text{for } s \geq \theta \end{cases}$
- $\mathcal{L}(s,\theta) = \frac{(s-\theta)^2}{|\theta|+1}$
- $\mathcal{L}(s,\theta) = \frac{s}{\theta} - 1 - \ln\left(\frac{s}{\theta}\right)$ …

Then the risk function is defined by $\mathcal{R}(\theta, W) := E(\mathcal{L}(W,\theta))$

and we want to have a value of $\mathcal{R}$ close to 0.

Remark: the nice feature $E((W-\theta)^2) = \mathrm{Var}(W) + \mathrm{Bias}(W)^2$

will not be possible in general.

⇒ minimizing $\mathcal{R}(\theta, W)$ can be complicated for other function $\mathcal{L}$.

⇒ It is a complicated question.