do it for every x

- Thus, Bayes rule is

mean is the best prediction of a r.v. under MSE → reasonable? → $\hat{\theta}(x) = \begin{cases} \int \theta\, h(\theta\,|\,x)\, d\theta, & \text{in the continuous case} \\ \sum_{\theta_i} \theta_i\, h(\theta_i|x), & \text{in the discrete case} \end{cases}$

a decision function (an estimator)

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

5/22

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 ⇒ for θ, can use uniform prior: $g(\theta) = 1$, $0 \le \theta \le 1$.

a vague prior

- Let

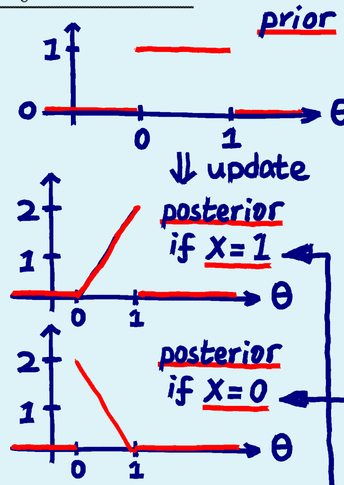
Data → $X = \begin{cases} 1, & \text{if a head appears} \\ 0, & \text{if a tail appears} \end{cases}$

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 → $g(\theta)$

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

- The Bayes estimator of θ is the posterior mean

under square error loss (L^2-norm)

$$\int_0^1 \theta \cdot h(\theta|x)\, d\theta = \begin{cases} \int_0^1 \theta \cdot (2\theta)\, d\theta = 2/3, & x = 1 \\ \int_0^1 \theta \cdot 2(1-\theta)\, d\theta = 1/3, & x = 0 \end{cases}$$

posterior median $\hat{\theta} = \begin{cases} 1/\sqrt{2}, & \text{if } x=1 \\ 1 - 1/\sqrt{2}, & \text{if } x=0 \end{cases}$

Q: Which one is more reasonable estimator?

- Note. The maximum likelihood estimator is $X = \begin{cases} 1, & x=1 \\ 0, & x=0 \end{cases}$

Frequentist approach

Note. With one observation, the data carries very little information about θ ⇒ prior becomes very helpful.

Theorem 10.3 (Bayes rule for Estimation under Absolute Error Loss)

In the case of absolute error loss, i.e.,

FYI $l(\theta, d) = |\theta - d|$ ← L^1-norm

robust estimator

median is the best predictor of a r.v. under absolute error

the Bayes estimator is the median of the posterior distribution. ← cf. Thm 10.2 (LNp.15)

50%-quantile

Definition 10.4 (dominate, strictly dominate, admissible, 2nd Ed., TBp.585)

- Let d_1, d_2 be two decision functions.
  - Say d_1 **dominates** d_2, if $R(\theta, d_1) \le R(\theta, d_2)$ for all θ.

risk function

d_1 is no worse than d_2. But, d_1, d_2 can be equally good.

  - Say d_1 **strictly dominates** d_2, if $R(\theta, d_1) \le R(\theta, d_2)$ for all θ, and the inequality is strict for some θ. ← d_1: no worse than d_2, but better for some θ

no need to consider d_2

- A decision function d is called **admissible** if d is not strictly dominated by *any* other decision function.

cf. → inadimissible (not adimissible)

- For estimation, an estimator is called **admissible** if it is not strictly dominated by *any* other estimator. (**Note**: Admissibility is a rather weak property. There could be many admissible estimators.) ← cf. → UMVUE ⇒ only one admissible est'or

Theorem 10.4 (2nd Ed., TBp.586) ← true for any decision problem, not only estimation.

Suppose that one of the following two assumptions holds:

1. ⓐ Ω is discrete, ⓑ d^* is a Bayes rule w.r.t. a prior pmf $g(\theta)$ such that ⓒ $g(\theta) > 0$ for all $\theta \in \Omega$. (countable)
2. ⓐ Ω is an interval (perhaps infinite) and ⓑ d^* is a Bayes rule w.r.t a prior pdf $g(\theta)$ such that ⓒ $g(\theta) > 0$ for all $\theta \in \Omega$ and ⓓ $R(\theta, d)$ is a continuous function of θ for all d. (uncountable) ← risk function

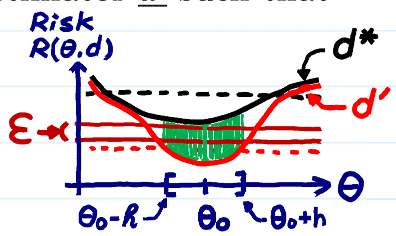
Then d^* is admissible. ← Intuition: If d' strictly dominates d^*, d' has smaller Bayes risk than d^*.

Proof. We prove it under 2. The proof under 1 is similar.

Suppose that d^* is inadmissible. Then, there is another estimator d' such that

d′ strictly dominates d* → $R(\theta, d^*) \geq R(\theta, d'), \quad \text{for all } \theta$

and with strict inequality for some θ, say θ_0.



Since $R(\theta, d^*) - R(\theta, d')$ is a continuous function of θ, there exist an $\epsilon > 0$ and an $h > 0$ such that

$$R(\theta, d^*) - R(\theta, d') > \epsilon > 0 \quad \text{for} \quad \theta_0 - h \leq \theta \leq \theta_0 + h$$

Then,

$$\int_{-\infty}^{\infty} \underbrace{\Big[R(\theta, d^*) - R(\theta, d')\Big]}_{\geq 0 \text{ all } \theta} \underbrace{g(\theta)\, d\theta}_{>0 \text{ all } \theta} \geq \int_{\theta_0 - h}^{\theta_0 + h} \Big[R(\theta, d^*) - R(\theta, d')\Big] g(\theta)\, d\theta$$

$B(d^*) - B(d')$ ← Bayes risk

$$\boxed{>} \; \epsilon \int_{\theta_0 - h}^{\theta_0 + h} g(\theta)\, d\theta \; \boxed{>} \; 0$$

∵ ⓒ

But this contradict the fact that d^* is Bayes rule with respect to $g(\theta)$, that is

$$B(d^*) - B(d') = \int_{-\infty}^{\infty} [R(\theta, d^*) - R(\theta, d')]\, g(\theta)\, d\theta \leq 0\,.$$

The proof is complete.

Notes.

The theorem can be regarded as both a positive and a negative result.

positive part: It identifies a certain class of estimators (i.e., Bayes estimators) as being admissible.
↑ for estimation d

(**Note**: In general, it could be difficult to examine whether a particular estimator is admissible, since by definition, one have to check that the estimator was not strictly dominated by *any* other estimators.)

negative part: There are so many admissible estimators — one for every prior distribution satisfying the hypotheses of the theorem.

⇒ admissibility is a rather weak property ⇒ can put more restriction.

e.g., unbiased → UMVUE ← unique

❖ **Reading**: textbook (2nd ed.), 15.2.4

• The Subjectivist Point of View – where the prior distributions come from?

↑ a distribution assigned on Ω
parameter: Fixed value → random variable

- Different viewpoints on probability.

  Question. how to interpret the statement:

  objective (客)
  subjective (主)

  "the probability that a coin will land heads up is 1/2"? ← $\theta \in \Omega$

Frequentists' interpretation: the long run relative frequency of heads approaches 1/2. ← long-run average ⇐ LLN ← objective → a fixed world (θ) ?→ probability that Shakespeare wrote Hamlet

Bayesian: the prior opinion (1/2) is such that they would as soon guess heads or tails if the rewards are equal. ← subjective →

Bayesian View of Probability (Personal Opinion) (2nd Ed., TBp. 587-588).

- Let A be an event (e.g., image that $A \subset S$: sample space). regarded as parameter.
- For a game in which if A occurs, the Bayesian will be rewarded \$1, $P(A)$ is the maximum amount of money the Bayesian would be willing to pay to buy into the game. → a fair game, average reward = pay, $1 \times p - R = 0 \Leftrightarrow R = p$. Frequentist approach
- Example: if the Bayesian is willing to pay at most 50 cents to buy in, $P(A) = 0.5$.
- **Note**: Bayesian probability is personal. $P(A)$ may vary from person to person. ← cf.
- Bayesians probability is a model for quantifying the strength of personal opinions or personal beliefs. Bayesian:每人心中皆可有不同的平行世界，Frequentist:每人心中的平行世界皆相同

Evolvement of Personal Opinion (2nd Ed., TBp. 588).

→ prior distribution

lot to lot in Ex10.1 (LNp.1)

- Bayes Thm describes how personal opinions evolve with experience.
- Suppose that the prior probability of Θ is $P(\theta)$. On observation of an event A, ← every person observe same A the opinion about Θ changes to: posterior dist.

$$\underbrace{P(\theta \mid A) = \frac{\boxed{P(A \mid \theta)}\, \overset{\text{update}}{P(\theta)}}{\boxed{P(A)}}}_{\text{Bayes' Thm (LN, CH1\sim6, p.23)}}.$$

★ can result in consistent probability with the long-run average interpretation

Difference btw Frequentist and Bayesian Approaches – Point Estimation (TBp. 588)

- Data X (random) has a joint pdf/pmf $f(x \mid \theta)$. ← parameter; Freq.: dist. of X; Baye.: cond. dist. of X
- **Frequentist**: θ has some fixed value that is unknown. Since θ is not random, it makes no sense to assign θ a pdf/pmf. cf.
- **Bayesian**: The prior opinion is $g(\theta)$ for Θ (random), and ★$f(x|\theta) \to f_{X|\Theta}(x|\theta)$. ← conditional dist. Having observed the data $X = x$, the new opinion about Θ is cf.

$$\underset{\text{posterior}}{\overset{\text{update}}{h(\theta \mid x)}} = \frac{\boxed{f(x \mid \theta)}\, g(\theta)}{\boxed{\int f(x \mid \theta) g(\theta)\, d\theta}}.$$

Bayes est'or

$f_X(x)$ or $p_X(x)$: marginal

prior can reflect personal experience. However, because it's personal, it can be ★biased

- **Note 1:** As a function of θ, the posterior density $h(\theta \mid x)$ is proportional to $f(x \mid \theta) g(\theta)$, i.e.,

$$h(\theta \mid x) \propto f(x \mid \theta)\ g(\theta).$$

if $g(\theta) \approx c$, a constant. no particular preference on some specific θ's

Check Ex10.7 (LNp.15): usually centering → posterior (Bayesian) ← cf.; usually maximizing → likelihood (Frequentist)

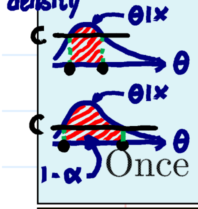
- **Note 2:** If $g(\theta)$ is nearly uniform in the region of θ where $f(x|\theta)$, regarded as a function of θ, has almost all its mass, then $h(\theta \mid x)$ is nearly proportional to the likelihood function $\mathcal{L}(\theta) = f(x \mid \theta)$. ← In likelihood: x: fixed, θ: variable

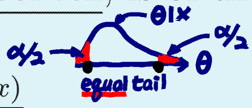
Difference btw Frequentist and Bayesian Approaches – Interval Estimation (TBp. 588)

- **Frequentist**: The confidence interval is a random interval $(\theta_L(X), \theta_U(X))$. A $100(1-\alpha)\%$ C.I. covers the true, fixed, unknown value of θ with probability $1-\alpha$. Once $X = x$ is observed, $P(\theta \in (\hat\theta_L(x), \hat\theta_U(x)))$ is 0 or 1. cf. Bayesian approach

- **Bayesian**: A Bayesian $100(1-\alpha)\%$ confidence interval, also called a **credibility interval**, is of the form $(\theta_L(x), \theta_U(x))$ where $\theta_L(x)$ and $\theta_U(x)$ satisfies

highest posterior density





$X|\theta$ — $P(\theta \in (\theta_L(X), \theta_U(X)))$ ← Frequentist approach (θ fixed, $(\theta_L(X), \theta_U(X))$ random)

$$\int_{\theta_L(x)}^{\theta_U(x)} h(\theta\,|\,x)\,d\theta = 1-\alpha := P(\theta \in (\theta_L(x), \theta_U(x)) \,|\, X=x)$$

$\theta|X$, random, fixed — cf.

Once $X=x$ is observed, the interval is fixed and θ is random.

Difference btw Frequentist and Bayesian Approaches - Testing (2nd Ed., TBp. 589)

- **Frequentist**: probabilities of Type I and Type II errors (α and β)
- **Bayesian**: deciding between two hypotheses (i.e., two subsets of Ω) in light of data reduces to comparing their posterior probabilities. — Ω_0 & Ω_A

$h(\theta|x),\ \theta\in\Omega_0$
$h(\theta|x),\ \theta\in\Omega_A$

check LNp.32~35

❖ **Reading**: textbook (2nd ed.), 15.3

- Bayesian Inference for the Normal distribution — $X|\theta$

dist. of data in Frequentist
cond'nal dist. of data in Bayesian

Theorem 10.5 (2nd Ed., TBp. 590)

- Suppose that a single observation X is taken, and
  - $X\,|\,\mu \sim N(\mu, \sigma^2)$, where μ is parameter and σ^2 known. (random, fixed)
  - $\mu \sim N(\mu_0, \sigma_0^2)$ ← Bayesian (random) — assumed known for simplicity

- Let $\xi = 1/\sigma^2$ and $\xi_0 = 1/\sigma_0^2$. (precision; $X|\theta$; θ)
- The posterior distribution of μ is normal with mean

best predictor of μ under prior ⇒ best estimate of μ before observing data

Frequentist estimator of μ

update $N(\mu_0, \sigma_0^2)$

Bayes est'or of μ under square loss →
$$\mu_1 = \frac{\xi_0\mu_0 + \xi x}{\xi + \xi_0} = \frac{\xi_0}{\xi+\xi_0}\mu_0 + \frac{\xi}{\xi+\xi_0}x$$

sum=1

if $\xi \gg \xi_0$
if $\xi \ll \xi_0$

and variance $\sigma_1^2 = 1/\xi_1$, where $\xi_1 = \xi + \xi_0$.

Proof. The posterior distribution of μ is

similar form: 2nd-order polynomial of μ

$$h(\mu|x) \;\propto\; f(x|\mu)\,g(\mu)$$

$N(\mu,\sigma^2)$ — $N(\mu_0,\sigma_0^2)$

This is a pdf of μ

That is, it is known $\int h(\mu|x)d\mu = 1$

$$\propto \exp\left[-\frac{1}{2\,\sigma^2}(x-\mu)^2\right] \times \exp\left[-\frac{1}{2\,\sigma_0^2}(\mu-\mu_0)^2\right]$$

$$\sigma_1^2 = \frac{\sigma^2\sigma_0^2}{\sigma^2+\sigma_0^2}$$

$$= \exp\left\{-\frac{1}{2}\left[\mu^2\underbrace{\left(\frac{1}{\sigma^2}+\frac{1}{\sigma_0^2}\right)}_{a\,\leftarrow\,\xi_1} - 2\mu\underbrace{\left(\frac{x}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)}_{b} + \frac{x^2}{\sigma^2}+\frac{\mu_0^2}{\sigma_0^2}\right]\right\}$$

When will σ_1^2 small?

$$\propto \exp\left[-\frac{1}{2\,a^{-1}}\left(\mu-\frac{b}{a}\right)^2\right]$$ → It has the form of normal pdf.

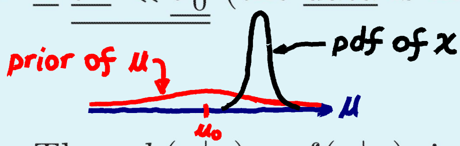
where $a = \sigma^{-2}+\sigma_0^{-2}$ and $b = x\sigma^{-2}+\mu_0\sigma_0^{-2}$. Thus the posterior distribution of μ is normal with mean

$$\mu_1 = \frac{b}{a} = \frac{x(1/\sigma^2)+\mu_0(1/\sigma_0^2)}{1/\sigma^2+1/\sigma_0^2} = \frac{x\,\xi + \mu_0\,\xi_0}{\xi+\xi_0}$$

and variance $a^{-1} = \left(\sigma^{-2}+\sigma_0^{-2}\right)^{-1} = (\xi+\xi_0)^{-1}$.

Notes (2nd Ed., TBp. 590).

- For a normal distribution $N(\mu, \sigma^2)$, $1/\sigma^2$ is called its **precision**. ↑ when σ^2 ↓
- data, prior, posterior
  ξ $(= 1/\sigma^2)$, ξ_0 $(= 1/\sigma_0^2)$, and ξ_1 $(= 1/\sigma_1^2)$ are precisions of $X|\mu$, prior, and posterior distributions, respectively. Notice that $\xi_1 = \xi + \xi_0$. ∵ combine information
  biased estimator under $X|\mu$ $\sigma_1^2 < \sigma_0^2$ ← posterior, data, prior
- The posterior mean is weighted average of prior mean μ_0 and data x, with weights proportional to the respective precisions.
- If $\sigma^2 \ll \sigma_0^2$ (the data is much more informative than the prior), then



Bayes est'or
$$\xi \gg \xi_0 \quad \text{and} \quad \xi_1 \approx \xi, \quad \Rightarrow \quad \mu_1 \approx x$$

Thus, $h(\mu|x) \approx f(x|\mu)$, i.e., $\mu|x$ is nearly distributed as $N(x, 1/\xi)$.
posterior, likelihood, update $N(\mu_0, 1/\xi_0)$

Exercise: What if $\sigma^2 \gg \sigma_0^2$? → prior dominates

- If prior distribution is quite flat relative to $f(x|\mu)$, as a function of μ, likelihood
  1. the prior distribution has little influence on the posterior,
  2. the posterior distribution is approximately proportional to the likelihood function. → Bayesian inference ←consistent (may not identical)→ Frequentist inference

  Such a prior is often called a **vague**, or **noninformative**, prior.