10. Bayesian Inference as a Decision Problem

The aim of statistical decision theory is to provide a framework for choosing between various possible *actions* after observing data.

Denote by Θ the set of all possible states of nature (values of parameter), and by $\mathcal{D}$ the set of all possible decisions (*actions*).

A *loss function* is any function

$$L : \Theta \times \mathcal{D} \to [0, \infty);$$

$L(\theta, d)$ gives the cost (penalty) associated with decision d if the true state of the world is θ.

Statistical inference as a decision problem

Suppose that we sample $x \in \mathcal{X}$ from $f(x, \theta)$, sampling distribution, and let $\pi(\theta)$ be our prior on θ, and $L(\theta, \delta)$ our loss function. Often a decision d is to evaluate or estimate a function $h(\theta)$; we would like to do this as accurately as possible.

Framework for point estimation:

Our function is $h(\theta) = \theta$; our set of decisions is $\mathcal{D} = \Theta$, and $L(\theta, d)$ is the loss in reporting d when θ is true.

Framework for hypothesis testing: Test $H_0 : \theta \in \Theta_0$; our set of decisions is just $\mathcal{D} = \{\text{accept } H_0,\ \text{reject } H_0\}$. Estimate

$$h(\theta) \ = \ \begin{cases} 1 & \text{if } \theta \in \Theta_0 \\ \\ 0 & \text{otherwise .} \end{cases}$$

We denote our loss function by

$$L(\theta,\ \text{accept } H_0) \ = \ \begin{cases} \ell_{00} & \text{if } \theta \in \Theta_0 \\ \\ \ell_{01} & \text{otherwise ;} \end{cases}$$

$$L(\theta,\ \text{reject } H_0) \ = \ \begin{cases} \ell_{10} & \text{if } \theta \in \Theta_0 \\ \\ \ell_{11} & \text{otherwise .} \end{cases}$$

Note: ℓ_{01} is the Type II-error, (accept H_0 although false), ℓ_{10} is the Type I-error (reject H_0 although true).

A *decision rule:* is a mapping $\delta : \mathcal{X} \to \mathcal{D}$. Our aim is to choose δ such that the loss is "small". In general there is no choice of δ which uniformly mimimizes $L(\theta, \delta(x))$.

Bayes estimators

In Bayesian statistics we average over the value of the parameter (not over the data) to assess the expected loss for a given decision. For a prior π and data $x \in \mathcal{X}$, the *posterior expected loss* of a decision (as a function of the data) is defined as

$$\rho(\pi, d|x) = \int_\Theta L(\theta, d)\pi(\theta|x)d\theta.$$

For a prior π the *integrated risk* of a decision rule δ is defined as

$$r(\pi, \delta) \ = \ \int_\Theta \int_\mathcal{X} L(\theta, \delta(x))f(x|\theta)dx\pi(\theta)d\theta.$$

This is a real number; we prefer the decision rule δ_1 to δ_2 if and only if $r(\pi, \delta_1) < r(\pi, \delta_2)$.

Proposition

An estimator minimizing $r(\pi, \delta)$ can be obtained by selecting, for every $x \in \mathcal{X}$, the value $\delta(x)$ that minimizes $\rho(\pi, \delta|x)$.

Proof (additional material)

$$\begin{aligned} r(\pi, \delta) &= \int_{\Theta} \int_{\mathcal{X}} L(\theta, \delta(x)) f(x|\theta) dx \pi(\theta) d\theta \\ &= \int_{\mathcal{X}} \int_{\Theta} L(\theta, \delta(x)) \pi(\theta|x) p(x) d\theta dx \\ &= \int_{\mathcal{X}} \rho(\pi, \delta|x) p(x) dx; \end{aligned}$$

recall that $p(x) = \int_{\Theta} f(x|\theta)\pi(\theta)d\theta)$ is the prior predictive distribution. The assertion follows.

A *Bayes estimator* associated with prior π, loss L, is any estimator δ^π which minimizes $r(\pi, \delta)$.

From our proposition it follows that, for $x \in \mathcal{X}$, the Bayes estimator is

$$\delta^\pi = \delta^\pi(x) = arg \min_d \rho(\pi, d|x).$$

The quantitiy $r(\pi) = r(\pi, \delta^\pi)$ is called *Bayes risk*.

The proposition is valid for proper priors, and for improper priors if $r(\pi) < \infty$. If $r(\pi) = \infty$ we define a *generalized Bayes estimator* as the minimizer, for every x, of $\rho(\pi, d|x)$.

Fact: For strictly convex loss functions, Bayes estimators are unique.

In principle the loss function is part of the problem specification. The existence of a suitable loss function is a big assumption. If a decision-theoretic approach is taken, it should be shown that the decisions are reasonable under a broad class of loss functions.

Some common loss functions

The *squared error loss* is given by

$$L(\theta, d) = (\theta - d)^2.$$

This is very common; it is convex, and the following result holds.

Proposition The Bayes estimator δ^π associated with prior π under squared error loss is the posterior mean,

$$\begin{aligned} \delta^\pi(x) &= E^\pi(\theta|x) \\ &= \frac{\int_\Theta \theta f(x|\theta)\pi(\theta)d\theta}{\int_\Theta f(x|\theta)\pi(\theta)d\theta}. \end{aligned}$$

Reason: for any random variable Y, $E((Y-a)^2)$ is minimized by $a = EY$.

Under squared error loss the integrated risk is the posterior mean-square error of the decision rule δ, see *Leonard and Hsu*, p.145.

Note that the squared error loss very sensitive to large deviations, which may be a criticism.

The *absolute error loss* is given by

$$L(\theta, d) = |\theta - d|.$$

Proposition: The posterior median is a Bayes estimator under absolute error loss.

A 0-1 loss function is used Bayesian testing:

Bayesian testing

Let $f(x, \theta)$ be our sampling distribution, $x \in \mathcal{X}$, $\theta \in \Theta$. Suppose that our null hypothesis is $H_0 : \theta \in \Theta_0$, and our possible decisions are

$$\mathcal{D} = \{\text{accept } H_0, \text{ reject } H_0\} = \{1, 0\},$$

where 1 stands for acceptance.

We choose as loss function

$$L(\theta,\phi)=\begin{cases}0 & \text{if } \theta\in\Theta_0,\phi=1,\\ a_0 & \text{if } \theta\in\Theta_0,\phi=0;\\ 0 & \text{if } \theta\notin\Theta_0,\phi=0,\\ a_1 & \text{if } \theta\notin\Theta_0,\phi=1.\end{cases}$$

In general $a_0\neq a_1$; the two types of error are penalized differently.

Proposition Under this loss function, the Bayes decision rule associated with a prior distribution π is

$$\phi^{\pi}(x) = \begin{cases} 1 \text{ if } P^{\pi}(\theta \in \Theta_0 | x) > \frac{a_1}{a_0+a_1}, \\ \\ 0 \text{ otherwise .} \end{cases}$$

The Bayes decision rule depends directly on the posterior probability of Θ_0. As $\frac{a_1}{a_0+a_1} = (1 + \frac{a_0}{a_1})^{-1}$, the decision rule depends on a_0 and a_1 only through the ratio $\frac{a_0}{a_1}$. The larger this ratio, the more the type 1 error is penalized, and the smaller the posterior probability of Θ_0 needed to accept H_0.

Note a special case: If $a_0 = a_1$ then we accept H_0 if $P^{\pi}(\theta \in \Theta_0 | x) > \frac{1}{2}$.

Proof (additional material) The posterior expected loss is

$$\begin{aligned}
\rho(\pi, \phi|x) &= a_0 P^\pi(\theta \in \Theta_0|x)\mathbf{1}(\phi(x) = 0) \\
&\quad + a_1 P^\pi(\theta \notin \Theta_0|x)\mathbf{1}(\phi(x) = 1) \\
&= a_0 P^\pi(\theta \in \Theta_0|x) + \mathbf{1}(\phi(x) = 1) \\
&\quad (a_1 - (a_0 + a_1)P^\pi(\theta \in \Theta_0|x)),
\end{aligned}$$

and $a_1 - (a_0 + a_1)P^\pi(\theta \in \Theta_0|x) < 0$ if and only if $P^\pi(\theta \in \Theta_0|x) > \frac{a_1}{a_0+a_1}$.

Example: Binomial distribution. Let $X \sim Bin(n, \theta)$,

$\Theta_0 = [0, 1/2)$, and choose as prior $\pi(\theta) = 1$. Then

$$\begin{aligned} P^{\pi}\left(\theta < \frac{1}{2}|x\right) &= \frac{\int_0^{\frac{1}{2}} \theta^x (1-\theta)^{n-x} d\theta}{\int_0^1 \theta^x (1-\theta)^{n-x} d\theta} \\ &= \frac{\left(\frac{1}{2}\right)^{n+1}}{B(x+1, n-x+1)} \\ & \quad \left\{\frac{1}{x+1} + \ldots + \frac{(n-x)!x!}{(n+1)!}\right\}. \end{aligned}$$

This expression can be evaluated for particular n and x, and compared with the acceptance level $\frac{a_1}{a_0+a_1}$.

Example: Normal distribution, known variance. Let $X \sim \mathcal{N}(\theta, \sigma^2)$, where σ^2 is known, and $\theta \sim \mathcal{N}(\mu, \tau^2)$. Then we calculated

$$\pi(\theta|x) \sim \mathcal{N}(\mu(x), w^2)$$

with

$$\mu(x) = \frac{\sigma^2\mu + \tau^2 x}{\sigma^2 + \tau^2} \text{ and } w^2 = \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}.$$

For $H_0 : \theta < 0$ we obtain

$$\begin{aligned} P^\pi(\theta < 0|x) &= P^\pi\left(\frac{\theta - \mu(x)}{w} < -\frac{\mu(x)}{w}\right) \\ &= \Phi\left(-\frac{\mu(x)}{w}\right). \end{aligned}$$

Let z_{a_0,a_1} be the $\frac{a_1}{a_0+a_1}$ quantile of the standard normal, then we accept H_0 if $-\mu(x) > z_{a_0,a_1} w$, or, equivalently, if

$$x < -\frac{\sigma^2}{\tau^2}\mu - \left(1 + \frac{\sigma^2}{\tau^2}\right) z_{a_0,a_1} w.$$

For $\sigma^2 = 1, \mu = 0, \tau^2 \to \infty$, we accept H_0 if $x < -z_{a_0,a_1}$.

Compare to the *frequentist approach:* We accept $H_0 : \theta = 0$ against the one-sided alternative $H_1 : \theta > 0$ if $x < z_{1-\alpha} = -z_\alpha$.

This corresponds to

$$\frac{a_0}{a_1} = \frac{1}{\alpha} - 1.$$

So, for example, $\frac{a_0}{a_1} = 19$ for $\alpha = 0.05$, and $\frac{a_0}{a_1} = 99$ for $\alpha = 0.01$.

Note:

1) If the prior probability of H_0 is 0, then so will be posterior probability.

2) Testing $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$ often really means testing $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$, which is natural to test in a Bayesian setting.

Strictly speaking, Bayesians do not test hypothesis; testing is a frequentist concept. Bayesians compare posterior probabilities.

Definition: The *Bayes factor* for testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ is

$$B^{\pi}(x) = \frac{P^{\pi}(\theta \in \Theta_0|x)/P^{\pi}(\theta \in \Theta_1|x)}{P^{\pi}(\theta \in \Theta_0)/P^{\pi}(\theta \in \Theta_1)}.$$

The Bayes factor measures the extent to which the data x will change the odds of Θ_0 relative to Θ_1 .

If $B^\pi(x) > 1$, the data adds support to H_0; if $B^\pi(x) < 1$, the data adds support to H_1; if $B^\pi(x) = 1$, the data does not help to distinguish between H_0 and H_1.

Note: the Bayes factor still depends on the prior π.

Special case: If both null hypothesis and alternative hypothesis are simple, $H_0 : \theta = \theta_0$, $H_1 : \theta = \theta_1$, then

$$B^{\pi}(x) = \frac{f(x|\theta_0)}{f(x|\theta_1)}$$

is the likelihood ratio.

More generally,

$$\begin{aligned} B^{\pi}(x) &= \frac{\int_{\Theta_0} \pi(\theta) f(x|\theta) d\theta}{\int_{\Theta_1} \pi(\theta) f(x|\theta) d\theta} \Bigg/ \frac{\int_{\Theta_0} \pi(\theta) d\theta}{\int_{\Theta_1} \pi(\theta) d\theta} \\ &= \frac{\int_{\Theta_0} \pi(\theta) f(x|\theta) / P^{\pi}(\theta \in \Theta_0) d\theta}{\int_{\Theta_1} \pi(\theta) f(x|\theta)) / P^{\pi}(\theta \in \Theta_1) d\theta} \\ &= \frac{p(x|\theta \in \Theta_0)}{p(x|\theta \in \Theta_1)} \end{aligned}$$

is the ratio of how likely the data is under H_0 and how likely the data is under H_1.

Compare: the frequentist generalized likelihood ratio is

$$\Lambda(x) = \frac{\sup_{\theta \in \Theta_0} f(x|\theta)}{\sup_{\theta \in \Theta_1} f(x|\theta)}.$$

In Bayesian statistics we average instead of taking suprema.

With ϕ^π from the Proposition, and denoting the prior probabilites

by $\rho_0 = P^\pi(\theta \in \Theta_0), \rho_1 = P^\pi(\theta \in \Theta_1)$, we obtain that

$$B^\pi(x) = \frac{P^\pi(\theta \in \Theta_0|x)/(1 - P^\pi(\theta \in \Theta_0|x))}{\rho_0/\rho_1},$$

and so

$$\phi^\pi(x) = 1 \iff B^\pi(x) > \frac{a_1}{a_0} \Big/ \frac{\rho_0}{\rho_1}.$$

Also, by inverting the equality it follows that

$$P^\pi(\theta \in \Theta_0|x) = \left(1 + \frac{\rho_1}{\rho_0}(B^\pi(x))^{-1}\right)^{-1}.$$

Example: Binomial distribution. Let $X \sim Bin(n,p)$, and $H_0 : p = 1/2$, $H_1 : p \neq 1/2$. We need to avoid that the prior puts zero probability on H_0, hence we choose as prior an atom of size ρ_0 at $1/2$, otherwise uniform. This gives for the Bayes factor

$$\begin{aligned} B^\pi(x) &= \frac{p(x|p=1/2)}{p(x|p \in \Theta_1)} \\ &= \frac{\binom{n}{x} 2^{-n}}{\binom{n}{x} B(x+1, n-x+1)}. \end{aligned}$$

So

$$P\left(p = \frac{1}{2}|x\right) = \left(1 + \frac{(1-\rho_0)}{\rho_0} \frac{x!(n-x)!}{(n-1)!} 2^n\right)^{-1}.$$

If $\rho_0 = 1/2, n = 5, x = 3$: $B^\pi(x) = \frac{15}{8} > 1$,

$$P\left(p = \frac{1}{2}|x\right) = \left(1 + \frac{2}{120} 2^5\right)^{-1} = \frac{15}{23}.$$

The data adds support to H_0, the posterior probability of H_0 is $15/23 > 1/2$.

Alternatively we could choose as prior an atom of size ρ_0 at $1/2$, otherwise $Beta(1/2, 1/2)$. This prior favours 0 and 1. For n=10 we obtain

x	$P(p = \frac{1}{2}\|x)$
0	0.005
1	0.095
2	0.374
3	0.642
4	0.769
5	0.803

Example: Normal distribution, known variance. Let $X \sim \mathcal{N}(\theta, \sigma^2)$, where σ^2 is known, and $H_0 : \theta = 0$. We choose as prior mass ρ_0 at $\theta = 0$, otherwise $\sim \mathcal{N}(0, \tau^2)$. Then

$$\begin{aligned}(B^\pi)^{-1} &= \frac{p(x|\theta \neq 0)}{p(x|\theta = 0)} \\ &= \frac{(\sigma^2+\tau^2)^{-1/2}\exp\{-x^2/(2(\sigma^2+\tau^2))\}}{\sigma^{-1}\exp\{-x^2/(2\sigma^2)\}}\end{aligned}$$

and

$$\begin{aligned}&P(\theta = 0|x) \\ &= \left(1 + \frac{1-\rho_0}{\rho_0}\sqrt{\frac{\sigma^2}{\sigma^2+\tau^2}}\exp\left(\frac{\tau^2 x^2}{2\sigma^2(\sigma^2+\tau^2)}\right)\right)^{-1}.\end{aligned}$$

Numerical example: $\rho_0 = 1/2, \tau = \sigma$, put $z = x/\sigma$. Alternatively:

$\tau = 10\sigma$ (which is a more diffusive prior).

x	$P(\theta = 0\|z)$	$P(\theta = 0\|z)$ under $\tau = 10\sigma$
0	0.586	0.768
0.68	0.557	0.729
1.28	0.484	0.612
1.96	0.351	0.366

so x gives stronger support for H_0 than under the tighter prior.

Note: For x fixed and $\tau^2 \to \infty$, $\rho_0 > 0$, we have

$$P(\theta = 0|x) \to 1.$$

For the noninformative prior $\pi(\theta) \propto 1$ we have

$$\begin{aligned} p(x|\pi(\theta)) &= \int (2\pi\sigma^2)^{-1/2} e^{-\frac{(x-\theta)^2}{2\sigma^2}} d\theta \\ &= (2\pi\sigma^2)^{-1/2} \int e^{-\frac{(\theta-x)^2}{2\sigma^2}} d\theta \\ &= 1, \end{aligned}$$

and so, if $\sigma^2 = 1$,

$$P(\theta = 0|x) = \left(1 + \frac{1-\rho_0}{\rho_0}\sqrt{2\pi}\exp(x^2/2)\right)^{-1}$$

which does not tend to 1.

Lindley's paradox

Let $\overline{X} \sim \mathcal{N}(\theta, \sigma^2/n)$, $H_0 : \theta = 0$, and n is fixed. If $\frac{\overline{x}}{(\sigma/\sqrt{n})}$ is large enough to reject H_0 in the frequentist test, then for large enough τ^2 the Bayes factor will be larger than 1, indicating support for H_0.

If σ^2, τ^2 are fixed, $n \to \infty$ such that $\frac{\overline{x}}{(\sigma/\sqrt{n})} = k_\alpha$ is fixed to be just significant at level α in the frequentist test, then $B^\pi(\overline{x}) \to \infty$.

Results which are just significant at some fixed level in the classical test will, for large n, actually be much more likely under H_0 than under H_1.

A very diffusive prior proclaims great scepticism, which may overwhelm the contrary evidence of the observations.

Least favourable Bayesian answers

For a thorough statistical analysis we should discuss the results for different choices of priors.

Assume that $H_0 : \theta = \theta_0$, $H_1 : \theta \neq \theta_0$, and our prior probability on H_0 is $\rho_0 = 1/2$.

Which is the prior g in H_1, which is, after observing x, least favourable to H_0?

Let G family of priors on H_1; put

$$\underline{B}(x, G) \;=\; \inf_{g \in G} \frac{f(x|\theta_0)}{\int_\Theta f(x|\theta) g(\theta) d\theta}$$

and

$$\underline{P}(x, G) = \frac{f(x|\theta_0)}{f(x|\theta_0) + \sup_{g \in G} \int_\Theta f(x|\theta) g(\theta) d\theta}$$

$$= \left(1 + \frac{1}{\underline{B}(x, G)}\right)^{-1}.$$

A Bayesian prior $g \in G$ on H_0 will then have posterior probability

at least $\underline{P}(x, G)$ on H_0 (for $\rho_0 = 1/2$).

If $\hat{\theta}$ is the m.l.e. of θ, and if G_A is the set of all prior distributions,

then

$$\underline{B}(x, G_A) = \frac{f(x|\theta_0)}{f(x|\hat{\theta}(x))}$$

and

$$\underline{P}(x, G_A) = \left(1 + \frac{f(x|\hat{\theta}(x))}{f(x|\theta_0)}\right)^{-1}.$$

Other natural families include:

G_S, the set of distributions symmetric around θ_0;

G_{SU}, the set of unimodal distributions symmetric around θ_0.

Example: Normal, unit variance.

Let $X \sim \mathcal{N}(\theta, 1)$, and $H_0 : \theta = \theta_0$, $H_1 : \theta \neq \theta_0$.

p-value	$\underline{P}(x, G_A)$	$\underline{P}(x, G_{SU})$
0.1	0.205	0.392
0.01	0.035	0.109

The Bayesian approach will typically reject H_0 less frequently than the frequentist apporach.

Comparison with classical hypothesis testing

Classical:

* asymmetry between H_0, H_1: fix type I error, minimize type II error;

* UMP tests do not always exist (general 2-sided tests, e.g.);

* p-values:

- have no intrinsic optimality, space of p-values lacks a decision-theoretic foundation,

- are routinely misinterpreted,

- do not take type II error into account.

* confidence regions:

- are a pre-data measure, can often have very different post data coverage probabilities.

Example: General simple null and simple alternative hypotheses.

Let $H_0 : \theta = \theta_0$, $H_1 : \theta = \theta_1$ be simple. Consider repetitions in which one uses the most powerful test with level $\alpha = 0.01$. In frequentist tests, only 1% of the true H_0 will be rejected. But this does not say anything about the proportion of errors made when rejecting.

Example: Suppose that the probability of the type II error is 0.99, and θ_0 and θ_1 occur equally often, then about half of the rejections of H_0 will be in error.

Example: Normal, known variance. Let $X \sim \mathcal{N}(\theta, 1/2)$,

$H_0 : \theta = -1$, $H_1 : \theta = 1$. If $x = 0$, the UMP p-value is 0.072.

But the p-value for the test of H_1 against H_0 takes exactly the

same value!

Example: Normal, unknown variance. Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(\theta, \sigma^2)$, and both θ, σ^2 are unknown. Let

$$C = \left(\overline{x} - t_{\alpha/2} \frac{s}{\sqrt{n}}, \overline{x} + t_{\alpha/2} \frac{s}{\sqrt{n}} \right)$$

be the frequentist $100(1-\alpha)\%$ confidence interval.

If $n = 2, \alpha = 0.5$, the pre- data coverage probability is 0.5. However, *Brown (Ann.Math.Stat. 38, 1967, 1068-1071)* showed that

$$P(\theta \in C | |\overline{x}|/s < 1 + \sqrt{2}) > 2/3;$$

the post-data coverage is very different.

The *Bayesian* approach compares the "probability" of the actual data under the two hypotheses.