Part III

A Decision-Theoretic Approach and Bayesian testing

Chapter 10

Bayesian Inference as a Decision Problem

The decision-theoretic framework starts with the following situation. We would like to choose between various possible *actions* after observing data. Let Θ denote the set of all possible states of nature (values of parameter), and let $\mathcal{D}$ denote the set of all possible decisions (*actions*). A *loss function* is any function

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

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

10.1 Inference as a decision problem

Denote by $f(x, \theta)$ the sampling distribution, for a sample $x \in \mathcal{X}$. Let $L(\theta, \delta)$ be our loss function. Often the decision d is to evaluate or estimate a function $h(\theta)$ as accurately as possible.

For *point estimation*, we want to evaluate $h(\theta) = \theta$; our set of decisions is $\mathcal{D} = \Theta$, the parameter space; and $L(\theta, d)$ is the loss in reporting d when θ is true.

For *hypothesis testing*, if we want to test $H_0 : \theta \in \Theta_0$, then our possible decisions are

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

We would like to evaluate the function

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

as accurately as possible. Our loss function is

$$\begin{aligned} L(\theta,\ \text{accept } H_0) &= \begin{cases} \ell_{00} & \text{if } \theta \in \Theta_0 \\ \ell_{01} & \text{otherwise} \end{cases} \\ L(\theta,\ \text{reject } H_0) &= \begin{cases} \ell_{10} & \text{if } \theta \in \Theta_0 \\ \ell_{11} & \text{otherwise .} \end{cases} \end{aligned}$$

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

10.2 Decision rules and Bayes estimators

In general we would like to find a *decision rule:* $\delta : \mathcal{X} \to \mathcal{D}$, a function which makes a decision based on the data. We would like to choose δ such that we incur only a "small" loss. In general there is no δ that uniformly mimimizes $L(\theta, \delta(x))$.

In a *Bayesian* setting, for a prior π and data $x \in \mathcal{X}$, the *posterior expected loss* of a decision is a function of the data x, 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 the real number defined as

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

We prefer 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 $p(x) = \int_\Theta f(x|\theta)\pi(\theta)d\theta$.) This proves the assertion.

A *Bayes estimator* associated with prior π, loss L, is any estimator δ^π which minimizes $r(\pi, \delta)$: For every $x \in \mathcal{X}$ it is

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

Then $r(\pi) = r(\pi, \delta^\pi)$ is called *Bayes risk.* This is valid for proper priors, and for improper priors if $r(\pi) < \infty$. If $r(\pi) = \infty$ one can 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.

10.3 Some common loss functions

In principle the loss function is part of the problem specification.

A very popular choice is *squared error loss* $L(\theta, d) = (\theta - d)^2$. This loss function is convex, and penalizes large deviations heavily.

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

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

To see this, recall that for any random variable Y, $E((Y - a)^2)$ is minimized by $a = EY$.

Another common choice for the loss function is *absolute error loss* $L(\theta, d) = |\theta - d|$.

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

10.4 Bayesian testing

Let $f(x, \theta)$ be our sampling distribution, $x \in \mathcal{X}$, $\theta \in \Theta$, and suppose that we cant to test $H_0 : \theta \in \Theta_0$. Then the set of possible decisions is $\mathcal{D} = \{\text{accept } H_0, \text{ reject } H_0\} = \{1, 0\}$, where 1 stands for acceptance. We use the loss function

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

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

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

Note: In the special case that $a_0 = a_1$, the rule states that we accept H_0 if $P^\pi(\theta \in \Theta_0 | x) > \frac{1}{2}$.

Proof (additional material) The posterior expected loss is

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

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: $X \sim Bin(n, \theta)$, $\Theta_0 = [0, 1/2), \pi(\theta) = 1$; then

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

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

Example: $X \sim \mathcal{N}(\theta, \sigma^2)$, σ^2 is known, $\theta \sim \mathcal{N}(\mu, \tau^2)$. We already calculated

$$\begin{aligned} \pi(\theta|x) &\sim \mathcal{N}(\mu(x), w^2) \\ \mu(x) &= \frac{\sigma^2\mu + \tau^2 x}{\sigma^2 + \tau^2} \\ w^2 &= \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}. \end{aligned}$$

To test $H_0 : \theta < 0$:

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

Let z_{a_0,a_1} be the $\frac{a_1}{a_0+a_1}$ quantile; 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 this to the *frequentist test:* Agian $\sigma^2 = 1$. Accept H_0 if $x < z_{1-\alpha} = -z_\alpha$. This corresponds to $\frac{a_0}{a_1} = \frac{1}{\alpha} - 1$. So $\frac{a_0}{a_1} = 19$ for $\alpha = 0.05$, and $\frac{a_0}{a_1} = 99$ for $\alpha = 0.01$, for example.

Note:

- If the prior probability of H_0 is 0, then so will be posterior probability.
- Testing $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$ often really means testing $H_0 : \theta \le \theta_0$ against $H_1 : \theta > \theta_0$, which is natural to test in a Bayesian setting.

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

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

It 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 that the Bayes factor still depends on the prior π.

Special case: If $H_0 : \theta = \theta_0$, $H_1 : \theta = \theta_1$, 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 this to the *frequentist likelihood ratio*

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

In Bayesian statistics we average instead of taking maxima.

Note: with ϕ^π from the Proposition, and $\rho_0 = P^\pi(\theta \in \Theta_0), \rho_1 = P^\pi(\theta \in \Theta_1)$,

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

and so

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

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: Let $X \sim Bin(n,p)$, $H_0 : p = 1/2$, $H_1 : p \neq 1/2$. Choose as prior an atom of size ρ_0 at $1/2$, otherwise uniform; then

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

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

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

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

We could have chosen an *alternatively* prior: atom of size ρ_0 at $1/2$, otherwise $Beta(1/2, 1/2)$; this prior favours 0 and 1. Then we obtain for n=10:

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

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

and

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

Example: $\rho_0 = 1/2, \tau = \sigma$, put $z = x/\sigma$

x	$P(\theta = 0\|z)$
0	0.586
0.68	0.557
1.28	0.484
1.96	0.351

For $\tau = 10\sigma$ (more diffusive prior)

x	$P(\theta = 0\|z)$
0	0.768
0.68	0.729
1.28	0.612
1.96	0.366

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

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

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

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

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

and so

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

which is not equal to 1.

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

In contrast, if σ^2, τ^2 are fixed, and $n \to \infty$ such that $\frac{\overline{x}}{(\sigma/\sqrt{n})} = k_\alpha$ fixed, which is just significant at level α in classical 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.

10.4.1 Least favourable Bayesian answers

The choice of prior affects the result. Suppose that we want to test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, and the prior probability on H_0 is $\rho_0 = 1/2$. Which prior g in H_1, after observing x, would be least favourable to H_0?

Let G be a 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 are G_S the set of distributions symmetric around θ_0, and G_{SU} the set of unimodal distributions symmetric around θ_0.

Example: Let $X \sim \mathcal{N}(\theta, 1)$, $H_0 : \theta = \theta_0$, $H_1 : \theta \neq \theta_0$, then we can calculate

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 approach

Consider the general test problem $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, and consider repetitions in which one uses the most powerful test with level $\alpha = 0.01$. In frequentist analysis, 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 the probability of 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.

10.4.2 Comparison with frequentist hypothesis testing

Frequentist hypothesis testing:

- 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: Let $X \sim \mathcal{N}(\theta, 1/2)$, $H_0 : \theta = -1$, $H_1 : \theta = 1$, and suppose that $x = 0$. Then 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: Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(\theta, \sigma^2)$, both θ, σ^2 are unknown. Then a $100(1-\alpha)$ confidence for θ is

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

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

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

Bayesian testing compares the "probability" of the actual data under the two hypotheses

Chapter 11

Hierarchical and empirical Bayesian methods

11.1 Hierarchical Bayes:

A hierarchical model consists of modelling a parameter θ through randomness at different levels; for example,

$$\theta|\beta \sim \pi_1(\theta|\beta), \text{ where } \beta \sim \pi_2(\beta);$$

so that then $\pi(\theta) = \int \pi_1(\theta|\beta)\pi_2(\beta)d\beta$.

When dealing with complicated posterior distributions, rather than evaluating the integrals, we might use simulation to approximate the integrals.

For simulation in hierarchical models, we simulate first from β, then, given β, we simulate from θ. We hope that the distribution of β is easy to simulate, and also that the conditional distribution of θ given β is easy to simulate. This approach is particularly useful for MCMC (Markov chain Monte Carlo) methods, e.g.: see next term.

11.2 Empirical Bayes:

Let $x \sim f(x|\theta)$. The empirical Bayes method chooses a convenient prior family $\pi(\theta|\lambda)$ (typically conjugate), where λ is a hyperparameter, so

$$p(x|\lambda) = \int f(x|\theta)\pi(\theta|\lambda)d\theta.$$

Rather than specifying λ, we estimate λ by $\hat{\lambda}$, for example by frequentist methods, based on $p(x|\lambda)$, and we substitute $\hat{\lambda}$ for λ;

$$\pi(\theta|x, \hat{\lambda})$$

is called a *pseudo-posterior*. We plug it into Bayes' Theorem for inference.

The empirical Bayes approach
* is neither fully Bayesian nor fully frequentist;
* depends on $\hat{\lambda}$, different $\hat{\lambda}$ will lead to different procedures;
* if $\hat{\lambda}$ is consistent, then asymptotically will lead to coherent Bayesian analysis.
* often outperforms classical estimators in empirical terms.

Example: James-Stein estimators

Let $X_i \sim \mathcal{N}(\theta_i, 1)$ be independent given θ_i, $i = 1, \ldots, p$,where $p \geq 3$. In vector notation: $\mathbf{X} \sim \mathcal{N}(\theta, I_p)$. Here the vector θ is random; assume that we have realizations θ_i, $i = 1, \ldots, p$. The obvious estimate (which is the least-squares estimate) for θ_i is $\hat{\theta}_i = x_i$, leading to

$$\hat{\theta} = \mathbf{X}.$$

Abbreviate $\mathbf{X} = x$. Our decision rule would hence be $\delta(x) = x$. But this is not the best rule!

Assume that $\theta_i \sim \mathcal{N}(0, \tau^2)$, then $p(x|\tau^2) = \mathcal{N}(0, (1 + \tau^2)I_p)$, and the posterior for θ given the data is

$$\theta|x \sim \mathcal{N}\left(\frac{\tau^2}{1+\tau^2}x, \frac{1}{1+\tau^2}I_p\right).$$

Under quadratic loss, the Bayes estimator $\delta(x)$ of θ is the posterior mean

$$\frac{\tau^2}{1+\tau^2}x.$$

In the empirical Bayes approach, we would use the m.l.e. for τ^2, which is

$$\hat{\tau^2} = \left(\frac{\parallel x \parallel^2}{p} - 1\right)\mathbf{1}(\parallel x \parallel^2 > p),$$

and the empirical Bayes estimator is the estimated posterior mean,

$$\delta^{EB}(x) = \frac{\hat{\tau}^2}{1+\hat{\tau}^2} x = \left(1 - \frac{p}{\parallel x \parallel^2}\right)^+ x$$

is the *truncated James-Stein estimator.* It can can be shown to outperform the estimator $\delta(x) = x$.

Alternatively, the best unbiased estimator of $1/(1+\tau^2)$ is $\frac{p-2}{\|x\|^2}$, giving

$$\delta^{EB}(x) = \left(1 - \frac{p}{\parallel x \parallel^2}\right) x.$$

This is the *James-Stein estimator.* It can be shown that under quadratic loss function the James-Stein estimator has smaller integrated risk than $\delta(x) = x$.

Note: both estimators tend to "shrink" towards 0. It is now known to be a very general phenomenon that when comparing three or more populations, the sample mean is not the best estimator. "Shrinkage" estimators are an active area of research.

Bayesian computation of posterior probabilities can be very computer-intensive; see the MCMC and Applied Bayesian Statistics course.

Chapter 12

Principles of Inference

12.1 The Likelihood Principle

The Likelihood Principle: The information brought by an observation x about θ is entirely contained in the likelihood function $L(\theta|x)$.

From this follows: if x_1 and x_2 are two observations with likelihoods $L_1(\theta|x)$ and $L_2(\theta|x)$, and if

$$L_1(\theta|x) = c(x_1, x_2)L_2(\theta|x)$$

then x_1 and x_2 must lead to identical inferences.

Example. We know that $\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$. If $f_1(x|\theta) \propto f_2(x|\theta)$ as a function of θ, then they have the same posterior, so they lead to the same Bayesian inference.

Example: Binomial versus negative binomial. (a) Let $X \sim Bin(n, \theta)$ be the number of successes in n independent trials, with p.m.f.

$$f_1(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

then

$$\begin{aligned}\pi(\theta|x) &\propto \binom{n}{x}\theta^x(1-\theta)^{n-x}\pi(\theta)\\ &\propto \theta^x(1-\theta)^{n-x}\pi(\theta).\end{aligned}$$

(b) Let $N \sim NegBin(x, \theta)$ be the number of independent trials until x successes, with p.m.f.

$$f_2(n|\theta) = \binom{n-1}{x-1} \theta^x (1-\theta)^{n-x}$$

and

$$\pi(\theta|x) \;\; \propto \;\; \theta^x (1-\theta)^{n-x} \pi(\theta).$$

Bayesian inference about θ does not depend on whether a binomial or a negative binomial sampling scheme was used.

M.l.e.'s satisfy the likelihood principle, but many frequentist procedures do not!

Example: $Bin(n, \theta)$-sampling. We observe $(x_1, \ldots, x_n) = (0, \ldots, 0, 1)$. An unbiased estimate for θ is $\hat{\theta} = 1/n$. If instead we view n as geometric(θ), then the only unbiased estimator for θ is $\hat{\theta} = \mathbf{1}(n = 1)$.

Unbiasedness typically violates the likelihood principle: it involves integrals over the sample space, so it depends on the value of $f(x|\theta)$ for values of x other than the observed value.

Example: (a) We observe a Binomial randome variable, n=12; we observe 9 heads, 3 tails. Suppose that we want to test $H_0 : \theta = 1/2$ against $H_1 : \theta > 1/2$. We can calculate that the UMP test has $P(X \geq 9) = 0.075$.
(b) If instead we continue tossing until 3 tails recorded, and observe that $N = 12$ tosses are needed, then the underlying distribution is negative binomial, and $P(N \geq 12) = 0.0325$.

12.2 The conditionality perspective

Example (*Cox 1958*) A scientist wants to measure a physical quantity θ. Machine 1 gives measurements $X_1 \sim \mathcal{N}(\theta, 1)$, but is often busy. Machine 2 gives measurements $X_1 \sim \mathcal{N}(\theta, 100)$. The availability of machine 1 is beyond the scientist's control, independent of object to be measured. Assume that on any given occasion machine 1 is available with probability 1/2; if available, the scientist chooses machine 1.

A standard 95% confidence interval is about $(x - 16.4, x + 16.4)$ because of the possibility that machine 2 was used.

Conditionality Principle: If two experiments on the parameter θ are available, and if one of these two experiments is selected with probability $1/2$, then the resulting inference on θ should only depend on the selected experiment.

The conditionality principle is satisfied in Bayesian analysis. In the frequentist approach, we could condition on an ancillary statistic, but such statistic is not always available.

A related principle is the *Stopping rule principle (SRP):* A *it stopping rule* is a random variable that tells when to stop the experiment; this random variable depends only on the outcome of the first n experiments (does not look into the future). The *stopping rule principle* states that if a sequence of experiments is directed by a stopping rule, then, given the resulting sample, the inference about θ should not depend on the nature of the stopping rule.

The likelihood principle implies the SRP. The SRP is satisfied in Bayesian inference, but it is not always satisfied in frequentist analysis.