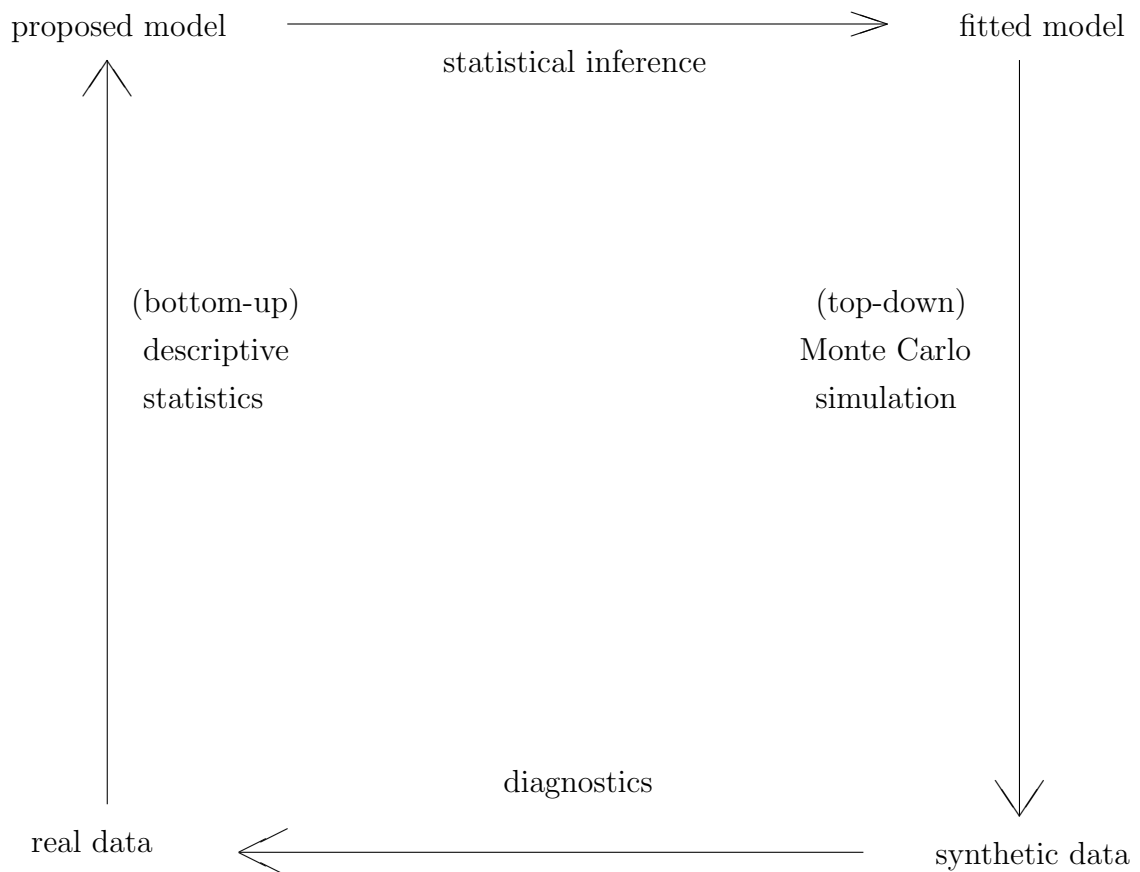


Lecture 1 Basic Elements in Decision Theory

1.1 Introduction

Put this course in a big picture: reconstruction cycle for statistical modelling



This course focuses on statistical inference. Each problem usually consists of three components:

- How to formulate it ? (formulation, setting, notation, etc.)
- How to solve it? (procedure)
- Why solve it this way? (justification, comparisons, theoretical and/or empirical, etc.)

Example 1.1 Let X_1, \dots, X_n be iid random variables having a common Bernoulli distribution with parameter θ , i.e. $P_\theta(X_1 = 1) = 1 - P_\theta(X_1 = 0) = \theta$. Estimate θ based on $X^n = (X_1, \dots, X_n)$.

Proposal 1: $d_1 = \bar{X}$

Proposal 2: $d_2 = [\inf\{1 \leq k \leq n : X_k = 1\}]^{-1}$ or $d_2 = 0$ if $X_i = 0 \forall i = 1, \dots, n$.

Proposal 3: $d_3 = 1/2$ (constant).

Pros and cons? Note that d_2 is motivated by the following fact: The random variable $T = \inf\{k \geq 1 : X_k = 1\}$ follows a geometric distribution with mean $E_\theta T = 1/\theta$. What about d_3 ? It ignores the data completely. Someone could argue that since θ is unknown anyway, he is willing to bet on $\theta = 1/2$. In case $\theta = 1/2$, then d_3 is precise. This shows the necessity of judging a proposal based on a certain criterion. d_3 turns out to be a “bad” proposal under both minimax and Bayesian criteria (later).

1.2 Basic elements in decision theory

- \mathcal{X} : sample space (data)
- Θ : parameter space (models, finite- or infinite-dimensional)
- \mathcal{A} : action space (procedures)
- $L(\cdot)$: loss function (criteria for evaluation), a mapping $L : \Theta \times \mathcal{A} \rightarrow \mathcal{R}$.

Note: In general, $x \in \mathcal{X}$ is a sample (data) generated from a model $\{f_\theta(\cdot) = f(\cdot|\theta), \theta \in \Theta\}$ (frequentist or Bayesian notation), where x may be a vector, i.e. $x = x^n = (x_1, \dots, x_n)$, and $f_\theta(x)$ is the density (Radon-Nikodym derivative) with respect to a reference measure (Lebesgue or counting). In what follows, we assume $\Theta \subset \mathbb{R}^k$ is finite-dimensional, i.e. we consider parametric models unless specified otherwise. A decision rule d is a mapping $d : \mathcal{X} \rightarrow \mathcal{A}$. Choices of \mathcal{A} and $L(\cdot)$ are problem-specific. See several inference problems for illustration.

Example 1.2 (point estimation) Estimate θ by $d(x)$, and measure the error by a weighted squared distance $L(\theta, d(x)) = w(\theta) \|d(x) - \theta\|^2$. $w(\theta) = 1$ is a special (unweighted) case.

Example 1.3 (hypothesis testing) Consider the partition $\Theta = \Theta_0 \cup \Theta_1$ where $\Theta_0 \cap \Theta_1 = \emptyset$. The action a_i claims $\theta \in \Theta_i$, $i = 0, 1$. It is a two-decision problem with $\mathcal{A} = \{a_0, a_1\}$. Usually, assume the 0-1 loss $L(\theta, a_i) = I_{\{\theta \notin \Theta_i\}}$ which corresponds to the type I error $L(\theta, a_1)$ and type II error $L(\theta, a_0)$ respectively.

Example 1.4 (multiple decision problems, e.g. classification) $\Theta = \Theta_1 \cup \dots \cup \Theta_m$, $m \geq 2$ (disjoint union). For $i = 1, \dots, m$, the decision a_i claims $\theta \in \Theta_i$, subject to the 0-1 loss or

an extension $L(\theta, a_i) = c_i I_{\{\theta \notin \Theta_i\}}$ with a cost $c_i > 0$.

Example 1.5 (confidence sets) Estimate θ by a set $C \subset \Theta$. In particular, $C = [c_1, c_2]$ is called a confidence interval (CI), with the case $c_1 = -\infty$ (resp. $c_2 = \infty$) referred to an upper (resp. lower) confidence bound. Usually, the 0-1 loss $L(\theta, C) = I_{\{C \not\ni \theta\}}$ is imposed. A more interesting loss function is $L(\theta, C) = I_{\{C \not\ni \theta\}} + \lambda m(C)$, where the constant $\lambda > 0$ balances the first term (reliability) and the second term (accuracy), and $m(C)$ represents the measure of set C . Obviously, $m(C) = c_2 - c_1$ with $C = [c_1, c_2]$.

To average out the randomness in data X , we define the risk function (a function of θ) for the decision d as the expected loss

$$R(\theta, d) = E_\theta L(\theta, d(X)).$$

Example 1.6 Let X_1, \dots, X_n be iid $N(\theta, 1)$ random variables. Consider the following inference problems:

- (i) Estimate θ with the restriction $\theta \geq 0$. A proposal: $d(X^n) = (\bar{X})^+$. (Is it better than \bar{X} ? Why?) Using a squared error loss, $R(\theta, d) = E_\theta [d(X^n) - \theta]^2 = \int_0^\infty (x - \theta)^2 \sqrt{\frac{n}{2\pi}} \exp[-\frac{n}{2}(x - \theta)^2] dx$ (no closed form).
- (ii) Test $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$ with a prescribed significance level $\alpha > 0$. Note the usual rejection region (critical region) is characterized by $d(X^n) = a_1$, or more specifically, by $|\sqrt{n} (\bar{X} - \theta_0)| > z_{\alpha/2}$. The risk function in this case is just the power function $R(\theta, d) = P_\theta (|\sqrt{n} (\bar{X} - \theta_0)| > z_{\alpha/2})$ which equals α (the type I error probability) when $\theta = \theta_0$ and the power of the test when $\theta \neq \theta_0$.
- (iii) A CI with confidence level $1 - \alpha$ (with the 0-1 loss) is given by $d(X^n) = [\bar{X} - z_{\theta/2}/\sqrt{n}, \bar{X} + z_{\theta/2}/\sqrt{n}]$. Now consider a CI of the type $d(X^n) = [\bar{X} - c, \bar{X} + c]$ for some $c > 0$ subject to the loss function $L(\theta, C) = I_{\{C \not\ni \theta\}} + \lambda m(C)$. We are to determine what value of c minimizes the risk. Let $g(c) = R(\theta, d)/2$. Then

$$g(c) = P_\theta (|\bar{X} - \theta| > c)/2 + \lambda c = \lambda c + \int_{c\sqrt{n}}^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

Note that $g(0) = 1/2$, $g(\infty) = \infty$, and $\frac{dg}{dc} = \lambda - \sqrt{\frac{n}{2\pi}} \exp(-nc^2/2)$. Assume $n > 2\pi\lambda^2$, then $\frac{dg}{dc}|_{c=0} < 0$. Setting $\frac{dg}{dc} = 0$ yields $c = \sqrt{\frac{1}{n} (\log \frac{n}{2\pi} - 2 \log \lambda)}$, which gets small as n gets large.