Inference

BS2 Statistical Inference, Lecture 1
Michaelmas Term 2004

Steffen Lauritzen, University of Oxford; October 11, 2004

Notation and setup

$\mathcal{X}$ denotes *sample space*, typically either finite or countable, or an open subset of $\mathcal{R}^k$.

We have observed *data* $x \in \mathcal{X}$ which are assumed to be a realisation $X = x$ of a random variable X.

The probability mass function (or density) of X is partially unknown, i.e. of the form $f(x; \theta)$ where θ is a *parameter*, varying in the *parameter space* Θ.

Statistical *inference* is concerned with saying something sensible about θ on the basis of having observed $X = x$.

Variations

The observation $X = x$ is usually more composite than just a real number.

Situations mostly considered here is when $X = (X_1, \ldots, X_n)$, corresponding to n independent repetitions under identical conditions. We then refer to x as a *sample of size* n and have

$$f(x; \theta) = f(x_1; \theta) \cdots f(x_n; \theta)$$

corresponding to $X_1, \ldots, X_n$ being independent and identically distributed.

But X can be more complex, as e.g. in regression.

Other common notations are $f_\theta(x)$, $f(x \,|\, \theta)$ and $f(x, \theta)$.

A simple example

Data may consist of 3 measurements of a height difference, in millimeters:

$$x = (119, 112, 114) = (x_1, x_2, x_3).$$

We may consider these to originate from a normal distribution, $\mathcal{N}(\mu, \sigma^2)$, so that $\theta = (\mu, \sigma^2)$ and

$$f(x; \theta) = \prod_{i=1}^{3} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}.$$

The parameter space Θ would typically be $\mathcal{R} \times \mathcal{R}_+$ corresponding to μ and σ^2 being completely unknown and $\sigma^2 > 0$.

Parameters and parameter spaces

Mostly Θ is an open subset of $\mathcal{R}^d$. Θ is sometimes implicitly specified, by *default* consisting of all values of θ where the expression $f(x;\theta)$ makes sense.

θ can be *extrinsic*, i.e. defined externally from substance matter considerations, or it can be *intrinsic*, simply labelling the *family* $\mathcal{F}$ of distributions.

$$\mathcal{F} = \{f(\cdot;\theta) \,|\, \theta \in \Theta\}.$$

The distinction between extrinsic and intrinsic is not always clear and often the true picture is mixed.

The triple $(\mathcal{X}, \mathcal{F}, \Theta)$ is referred to as the *statistical model* but the latter expression is also used in an informal sense.

Estimation

Estimation is 'guessing' the value of θ based on $X = x$. We speak about a *point estimate* if the value of the estimator is a single point:

$$\hat{\theta} = \hat{\theta}(x) = t(x) = t(x_1, \ldots, x_n),$$

whereas we speak about an *interval estimate* or *set estimate* if we just state that, based on having seen x, we conclude

$$\theta \in C = C(x) = C(x_1, \ldots, x_n)$$

with some certainty.

Estimation theory is concerned with principles and methods for finding estimates and assessing their uncertainty.

Example revisited

The parameter μ may be extrinsically defined as the true height difference, whereas σ^2 is just labelling the distributions, thus intrinsic.

On the other hand, σ^2 is also the inaccuracy of the measuring instrument, thus extrinsic.

Potential point estimates for μ and σ are

$$\hat{\mu} = \bar{x} = \frac{x_1 + x_2 + x_3}{3} = 115, \quad \tilde{\mu} = \texttt{median}(x) = x_{(2)} = 114.$$

$$\hat{\sigma} = s = 3.61, \quad \tilde{\sigma} = \texttt{range}(x)/2 = (x_{(3)} - x_{(1)})/2 = 3.5.$$

A potential set estimate for μ is $[x_{(1)}, x_{(3)}] = [112, 119]$.

Unbiased estimators

An estimator $\hat{\theta} = t(X)$ is said to be *unbiased* for a function θ if it equals θ in expectation:

$$\mathbf{E}_\theta\{t(X)\} = \mathbf{E}\{\hat{\theta}\} = \theta.$$

More generally $\hat{g}(\theta) = t(x)$ is *unbiased for a function* $g(\theta)$ if

$$\mathbf{E}_\theta\{t(X)\} = g(\theta).$$

In the example, $\hat{\mu}$ and $\tilde{\mu}$ are unbiased for μ whereas neither of $\hat{\sigma}$ and $\tilde{\sigma}$ are unbiased for σ.

Show this as an *exercise* (problem sheet 1).

Frequentist and Bayesian probability

Frequency interpretation of probability identifies $P(A)$ with the long-run relative frequency of an event A in repeated trials:

$$P(A) = \lim_{n \to \infty} \frac{\#\text{ of occurrences of } A \text{ in } n \text{ trials}}{n}.$$

Bayesian probability is *subjective* and $P(A)$ is Your personal belief that A will occur or has occurred, so that $P(A)/\{1 - P(A)\}$ are *fair odds* to You for a bet on the event A.

Probability seems most meaningful when aspects of both interpretations pertain to the problem in question.

Frequentist paradigm for inference

Uses frequency interpretation of probability.

Reports uncertainty of an estimator in terms of the *sampling distribution* of a point estimator $\hat{\theta} = t(x)$ or set estimator $C(x)$.

For example the uncertainty of $\hat{\mu} = \bar{x}$ when $X_i \sim \mathcal{N}(\mu, \sigma^2)$ can be reported as

$$\hat{\mu} \sim \mathcal{N}(\mu, \sigma^2/n)$$

which makes fully sense when σ^2 is known. Otherwise the reporting is more complex and uses the t-distribution.

Note that *unbiasedness is a frequentist concept*.

Bayesian inference

1. Express Your belief about θ in terms of a *prior distribution* π, based on knowledge you have before observing x.

2. Form the *joint distribution* of data and parameter by interpreting $f(x;\theta)$ as the conditional distribution of data given θ, therefore often written $f(x \mid \theta)$:

$$f(x,\theta) = f(x \mid \theta)\pi(\theta).$$

3. For *inference*, calculate the *posterior distribution* π^* of θ by Bayes' formula (conditional probability)

$$\pi^*(\theta) = f(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{\int f(x \mid \eta)\pi(\eta)\, d\eta} \propto L_x(\theta)\pi(\theta).$$

Likelihood

The function $L_x(\theta) = f(x \mid \theta) = f(x; \theta)$ is known as the *likelihood function*, obtained from $f(x; \theta)$ by considering x fixed and θ varying.

Its logarithm is traditionally denoted $l_x(\theta) = \log L_x(\theta)$.

The likelihood function is *equally important for frequentist and Bayesian* inference and *likelihood* may well be the *most fundamental concept* within the entire subject of statistics.

The likelihood function ranges the parameter values according to how much probability they give to the data.

It is only well-defined up to a multiplicative constant: *Proportional likelihood functions are equivalent*.