Part II — Principles of Statistics

Based on lectures by R. Nickl

Notes taken by Joseph Tedds using Dexter Chua's header and Gilles Castel's snippets.

Michaelmas 2019

These notes are not endorsed by the lecturers, and I have modified them (often significantly) after lectures. They are nowhere near accurate representations of what was actually lectured, and in particular, all errors are almost surely mine.

The Likelihood Principle

Basic inferential principles. Likelihood and score functions, Fisher information, Cramer-Rao lower bound, review of multivariate normal distribution. Maximum likelihood estimators and their asymptotic properties: stochastic convergence concepts, consistency, efficiency, asymptotic normality. Wald, score and likelihood ratio tests, confidence sets, Wilks theorem, profile likelihood. Examples. [8]

Bayesian Inference

Prior and posterior distributions. Conjugate families, improper priors, predictive distributions. Asymptotic theory for posterior distributions. Point estimation, credible regions, hypothesis testing and Bayes factors [3]

Decision Theory

Basic elements of a decision problem, including loss and risk functions. Decision rules, admissibility, minimax and Bayes rules. Finite decision problems, risk set. Stein estimator. [3]

Multivariate Analysis

Correlation coefficient and distribution of its sample version in a bivariate normal population. Partial correlation coefficients. Classification problems, linear discriminant analysis. Principal component analysis. [5]

Nonparametric Inference and Monte Carlo Techniques

GlivenkoCantelli theorem, KolmogorovSmirnov tests and confidence bands. Bootstrap methods: jackknife, roots (pivots), parametric and nonparametric bootstrap. Monte Carlo simulation and the Gibbs sampler. [4]

Contents

0 Introduction

Consider a random variable X defined on some probability space,

$$X : (\Omega, A, P) \mapsto \mathbb{R}.$$

We call Ω the set of outcomes, A is the set of measurable events in Ω and P is our probability measure on A. with distribution function

$$F(t) = P\left(\omega \in \Omega : X(\omega) \leq t\right), \quad t \in \mathbb{R}.$$

If X is a discrete random variable, then

$$F(t) = \sum_{x \leq t} f(x).$$

where f is the probability mass function (pmf) and if X is a continuous random variable, then

$$F(t) = \int_{-\infty}^{t} f(x) \mathrm{d}x.$$

where f is the probability density function (pdf).
We typically only write $F(t) = P\left(X \leq t\right)$, where P is the *law* of X (i.e. the image measure $P = \mathbb{P} \circ X^{-1}$).

Definition (Statistical model)**.** A *statistical model* for the law P of X is any collection

$$\{f(\theta) : \theta \in \Theta\}, \text{ or } \{P_\theta : \theta \in \Theta\}.$$

of pdf/pmf's or probability distributions. The index set Θ is the parameter space

Example. (i) $N(0,1), \theta \in \Theta = \mathbb{R}$, or $\Theta = [-1,1]$

(ii) $N(\mu, \sigma^2)$, $(\mu, \sigma^2) = \theta \in \Theta = \mathbb{R} \times (0, \infty)$

(iii) $\mathrm{Exp}(\theta), \ldots$

Definition (Correctly specified)**.** A statistical model $\{P_\theta : \theta \in \Theta\}$ is *correctly specified* (for the law P of X) if $\exists\ \theta \in \Theta$ such that $P_\theta = P$. We often write θ_0 for this specific 'true' value of θ. We say that observations $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P_\theta$ arise from the model $\{P_\theta : \theta \in \Theta\}$ in this case. We refer to n as the sample size.

The tasks of statistical inference comprise at least:

(i) Estimation - construct an estimator $\hat{\theta}_n = \hat{\theta}(x_1, \ldots, x_n) \in \Theta$ that is close with high probability to θ when $x_1, \ldots, x_n \overset{\text{iid}}{\sim} P_\theta,\ \forall\ \theta \in \Theta$.

(ii) Hypothesis testing - For $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$, we want a test (indicator) function $\psi_n = \psi(x_1, \ldots, x_n)$ such that $\psi_n = 0$ with high probability when H_0 is true, and $\psi_n = 1$ otherwise.

(iii) Confidence regions (inference) - Find regions (intervals) $C_n = C(x_1, \ldots, x_n, \alpha) \subseteq \Theta$ of confidence in that

$$P_\theta(\theta \in C_n) \overset{(\geq)}{=} 1 - \alpha,\ \forall\ \theta \in \Theta.$$

This quantifies the uncertainty in the inference on θ by the size (diameter) of C_n. Here $0 < \alpha < 1$ is a pre-scribed significance level.

1 Likelihood Principle

Example. Consider a sample $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\theta)$ with (unknown) $\theta > 0$. If the actual observed values are $X_1 = x_1, \ldots, X_n = x_n$, then the probability of this particular occurance of $x_1, \ldots, x_n$ as a function of θ is

$$\begin{aligned} f(x_1, \ldots, x_n, \theta) &= P_\theta(X_1 = x_1, \ldots, X_n = x_n) \\ &= \prod_{i=1}^n P_\theta(X_i = x_i) \\ &= \prod_{i=1}^n e^{-\theta} \frac{\theta^{x_i}}{x!} \\ &= e^{-n\theta} \prod_{i=1}^n \frac{\theta^{x_i}}{x_i!} \\ &\equiv L_n(\theta) \end{aligned}$$

a random function of θ.

Idea Maximise $L_n(\theta)$ over Θ, and for continuous variables, replace pmf's by pdf's. In the example above, we can equivalently maximise

$$\ell_n(\theta) = \log L_n(\theta) = -n\theta + \log\theta \sum_{i=1}^n X_i - \sum_{i=1}^n \log(x_i!) \text{ over } (0, \infty).$$

Then

$$\ell_n'(\theta) = -n + \frac{1}{\theta} \sum_{i=1}^n X_i \overset{\text{FOC}}{=} 0 \iff \hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n X_i.$$

Also,

$$\ell_n''(\theta) = -\frac{1}{\theta^2} \sum_{i=1}^n X_i < 0 \text{ if not all } X_i = 0 \text{ (in which case } \theta = 0 = \frac{1}{n} \sum_{i=1}^n X_i).$$

Definition (Likelihood function)**.** Given a statistical model $\{f(\cdot, \theta); \theta \in \Theta\}$ of pdf/pmf's for the law P of X, and given numerical observations $(x_i, i = 1, \ldots, n)$ arising as iid copies $X_i \overset{\text{iid}}{P}$, the *likelihood function of the model* is defined on

$$L_n : \Theta \mapsto \mathbb{R}, \quad L_n(\theta) = \prod_{i=1}^n f(x_i, \theta).$$

Moreover, the *log-likelihood* function is

$$\ell_n : \Theta \mapsto \mathbb{R} \cup \{-\infty\}, \ell_n(\theta) = \sum_{i=1}^n \log f(x_i, \theta),$$

and the *normalised log-likelihood function*

$$\bar{\ell}_n(\theta) = \frac{1}{n} \sum_{i=1}^n \log f(x_i, \theta).$$

We regard these functions as ('random' via the X_i's) maps of θ.

Definition (Maximum likelihood estimator)**.** A *maximum likelihood estimator* (MLE) is any $\hat{\theta} = \hat{\theta}_{\text{MLE}}(X_1, \dots, X_n) \in \Theta$ such that

$$L_n(\hat{\theta}) = \max_{\theta \in \Theta} L_n(\theta).$$

Equivalently, $\hat{\theta}$ maximises ℓ_n or $\overline{\ell}_n$ over Θ.

Example. For Poisson$(\theta), \theta \geq 0$, we have seen $\hat{\theta}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^n X_i$

Example. $N(\mu, \sigma^2)$, where $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty)$ one shows that the MLE

$$\hat{\theta}_{\text{MLE}} = \begin{pmatrix} \hat{\mu}_{\text{MLE}} \\ \hat{\sigma}^2_{\text{MLE}} \end{pmatrix} = \begin{pmatrix} \overline{X}_n \\ \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2 \end{pmatrix}, \overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$$

is obtained from simultaneously solving $\frac{\partial}{\partial \mu}\ell_n(\theta) = \frac{\partial}{\partial \sigma^2}\ell_n = 0$

Remark. Calculation of 'marginal' MLE's that optimise only one variable is not sufficient. Typically, the MLE for $\theta \in \Theta \subseteq \mathbb{R}^p$ is found by solving the *score equations*

$$S_n(\hat{\theta}) = 0, \text{ where } S_n : \Theta \mapsto \mathbb{R}^p$$

is the score function

$$S_n(\theta) = \nabla \ell_n(\theta) = \left(\frac{\partial}{\partial \theta_1}\ell_n(\theta), \dots, \frac{\partial}{\partial \theta_p}\ell_n(\theta) \right).$$

Here we use the implicit notation $S_n(\hat{\theta}) = \nabla \ell_n(\theta)\Big|_{\theta = \hat{\theta}}$

Remark. The likelihood principle 'works' as soon as a joint family $\{f(\cdot, \theta) : \theta \in \Theta\}$ pdf/pmf of $X_1, \dots, X_n$ can be specified and does not rely on the iid assumption. For instance, in the normal linear model, $N(X\beta, \sigma^2 I)$, where X is a $n \times p$ matrix $(\beta, \sigma^2 = \theta \in \mathbb{R} \times (0, \infty)$, the MLE coincides with the least squares estimator (not iid but independent).

2 Information geometry

Notation. For a random variable X of law / distribution P_θ on $\chi \subseteq \mathbb{R}^d$ and let $g : \chi \to \mathbb{R}$ be given. We will write

$$\mathbb{E}_\theta\left[g(X)\right] = \mathbb{E}_{P_\theta}\left[g(X)\right] = \int_\chi g(x) \mathrm{d}P_\theta(x)$$

which in the continuous case equals $\int_\chi g(x)f(x,\theta), \mathrm{d}x$, and in the discrete case is $\sum_{xinX} g(x)f(x_\theta)$

Observation Consider a model $\{f(\cdot,\theta) : \theta \in \Theta\}$ for X of law P on χ, and assume $\mathbb{E}_P\left[|\log f(x,\theta)|\right] < \infty$. Then $\overline{\ell}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \log f(x_i,\theta)$ as a sample approximation of

$$\ell(\theta) = \mathbb{E}_P\left[\log f(X,\theta)\right], \theta \in \Theta.$$

If the model is correctly specified, with any true value θ_0 such that $P = P_{\theta_0}$, then we can rewrite

$$\ell(\theta) = \mathbb{E}_{P_{\theta_0}}\left[\log f(X,\theta)\right] = \int_\chi (\log f(x,\theta) f(x,\theta_0) \mathrm{d}x.$$

Next we write

$$\begin{aligned}
\ell(\theta) - \ell(\theta_0) &= \mathbb{E}_{\theta_0}\left[\log \frac{f(X,\theta}{f(X,\theta_0}\right] \\
&\overset{\text{(Jensen)}}{\leq} \log \mathbb{E}_{\theta_0}\left[\frac{f(X,\theta}{f(X,\theta_0}\right] \\
&= \log \int_\chi \frac{f(X,\theta)}{f(X,\theta_0} f(X,\theta_0)\mathrm{d}x \\
&= \log \int_\chi f(x,\theta)\mathrm{d}x = 0 \; \forall\; \theta \in \Theta
\end{aligned}$$

.

Thus $\ell(\theta) \leq \ell(\theta_0) \;\forall\; \theta \in \Theta$, and approximately maximising $\ell(\theta)$ appears sensible. Note next that by the strict version of Jensen's inequality, $\ell(\theta) = \ell(\theta_0)$ can only occur when $\frac{f(X,\theta}{f(X,\theta_0}$ = constant (in X), which since $\int_\chi f(x,\theta)\mathrm{d}x = 1$ can only happen when $f(\cdot,\theta) \overset{\text{almost surely}}{=} f(\cdot,\theta_0)$ identically.

Definition (Identifiable)**.** Let us thus say that the model is *identifiable* if $f(\cdot,\theta) = f(\cdot,\theta)(\text{a.s}) \iff \theta = \theta_0$. In this case, the function $\ell(\theta)$ has a unique maximiser at the true value θ_0.

The quantity

$$0 \leq -(\ell(\theta) - \ell(\theta_0)) = \mathbb{E}_{\theta_0}\left[\log \frac{f(X,\theta_0)}{f(X,\theta)}\right] \equiv \mathrm{KL}(P_{\theta_0}, P_\theta).$$

is called the Kullback-Leibler divergence (entropy-distance), which builds the basis of statistical information theory. In particular, the differential geometry of the maps $\theta \mapsto \mathrm{KL}(P_{\theta_0}, P_\theta)$ determines what 'optimal' inference in a statistical model could be.