smfd1(13a).tex
Day 1. 11.11.2013

I. ESTIMATION OF PARAMETERS

1. PARAMETERS; LIKELIHOOD

To do Statistics – to handle the mathematics and data analysis of situations involving randomness – we need to *model* the situation. Here we confine ourselves to models that can be specified by a *parameter*, θ, which will be *finite-dimensional*. Often, θ will be one-dimensional. Usually, the dimensionality will be quite low (at most 5 or 6, say), unless the parameters are vectors or scalars (which will be the case with Multivariate Analysis, Ch. III. When infinitely many dimensions are needed, one speaks instead of a *non-parametric* model; see Ch. VI. Sometimes, one has a compound model, with a parametric part and a non-parametric part; one speaks then of a *semi-parametric model*.

Things should be kept as simple as possible (but not simpler!) So we should always work with as few parameters as possible – or, in the lowest possible number of dimensions. If we are unsure about what this is, we need to formulate a question on this, and test it on the data. This is the context of *Hypothesis testing*, Ch. II.

We deal with a probability distribution, F, describable by a parameter θ. Our data consists of a random variable X, or random variables $X_1, \ldots, X_n$, drawn from this distribution. A *statistic* is just a function of the data – something we can calculate when we have done our sampling and obtained our data; an *estimator* of θ is a statistic used to estimate a parameter θ. Often our data $X_1, \ldots, X_n$ will be *independent and identically distributed (iid)*; we call them independent *copies* drawn from F, or independent *draws* from F. We shall use the same letter F for the probability distribution or law (a measure), and the corresponding probability distribution function (a function); F will be a Lebesgue-Stieltjes measure (function) in the language of Measure Theory (Stochastic Processes, Ch. I – SP I). By the Lebesgue decomposition theorem,

$$F = F_{ac} + F_d + F_{cs} = F_{ac} + F_s,$$

where F_{ac} is the absolutely continuous component (w.r.t. Lebesgue measure; write f for its Radon-Nikodym derivative, called the (probability) *density*

(function) of F, X), F_d is the discrete component (probability mass $m_n > 0$ at a finite or countable set of points x_n), and F_{cs} is the continuous singular component. We often combine the last two, into the *singular* component, F_s. In this course, without further comment, we shall always be dealing with the absolutely continuous case, with density f, *or* with the discrete case, in which case (partly to simplify notation, partly to emphasise that here the base or reference measure is counting measure rather than Lebesgue measure) we write $f(x_n)$ for the probability mass m_n at the point x_n.

The most basic questions to ask about a random variable are 'how big is it' (on average), and this is measured by the *mean*,

$$\mu \quad \text{or} \quad \mu_X := E[X],$$

and 'how variable (or how random) is it', which is measured by the *variance*

$$\sigma^2, \quad \text{or} \quad \sigma_X^2 := E[(X - E[X])^2] = E[X^2] - [EX]^2.$$

We write

$$\mu_2 := E[X^2], \qquad \mu_n := E[X^n] \qquad (n = 1, 2, \ldots).$$

Our first task is usually to estimate the mean, and we like to be 'right on average'. We call an estimator S for θ *unbiased* if

$$ES = \theta;$$

otherwise it has *bias* $ES - \theta$. For the mean, we have an obvious estimator, the *sample mean* $\bar{X}$. This is unbiased, and by the Strong Law of Large Numbers (SLLN – Stochastic Processes), $\bar{X} \to \mu$ $(n \to \infty)$ a.s.; we say that $\bar{X}$ is *consistent* for μ (we 'get the right answer in the limit'). For the variance, matters are somewhat more complicated. The sample variance

$$S^2 := \frac{1}{n} \sum_1^n (X_k - \bar{X})^2 = \overline{X^2} - [\bar{X}]^2$$

is consistent, as by SLLN

$$S^2 \to E[X^2] - [E\ X]^2 = \sigma^2 \qquad (n \to \infty).$$

However, it is biased, and to obtain the unbiased version we have to divide by $n - 1$ instead of n (as the authors of many textbooks do for this reason –

always check!) For,

$$\begin{aligned} nS^2 &= \sum_{k=1}^{n}(X_k - \bar{X})^2 \qquad (\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i) \\ &= \sum_k X_k^2 - \frac{2}{n}\sum_{ik} X_k X_i + n.\frac{1}{n^2}\sum_{ij} X_i X_j \\ &= \sum_k X_k^2 - \frac{1}{n}\sum_{ik} X_k X_i. \end{aligned}$$

Now if $i = k$ $E[X_i X_k] = E[X_k^2] = \mu_2$, and if $i \neq k$ $E[X_i X_k = E[X_i].E[X_k] = \mu^2$, by independence. So

$$nE[S^2] = n\mu_2 - \frac{1}{n}[n\mu_2 + n(n-1)\mu^2] = (n-1)[\mu_2 - \mu^2] = (n-1)\sigma^2.$$

So

$$E[S^2] = \frac{n-1}{n}\sigma^2 : \qquad E\Big[\frac{n}{n-1}S^2\Big] = \sigma^2,$$

or

$$E[S_u^2] = \sigma^2, \qquad S_u^2 := \frac{1}{n-1}\sum_1^n (X_k - \bar{X})^2.$$

Here S_u^2 is called the *unbiased* (version of the) sample variance.

We recapitulate from Introductory Statistics Ch. II (IS II).

Likelihood.

We write θ for a parameter (scalar or vector), and write such examples as $f(x|\theta)$, which we will call the *density* (w.r.t. Lebesgue measure in the first three examples, counting measure in the fourth – see SP I). Here x is the *argument* of a function, the density function.

If we have n independent copies sampled from this density, the joint density is the product of the marginal densities:

$$f(x_1, \ldots, x_n|\theta) = f(x_1|\theta).\ldots.f(x_n|\theta) : \quad f(.,\ldots,.|\theta) = f(.|\theta).\ldots.f(.|\theta). \qquad (*)$$

DATA.

Now suppose that the numerical values of the random variables in our data set are $x_1, \ldots, x_n$. Fisher's great idea of 1912 was to put the data x_i where the arguments x_i were in $(*)$. He called this (later, 1921 on) the *likelihood*, L – a function of the parameter θ:

$$L(\theta) := f(x_1, \ldots, x_n|\theta) = f(x_1|\theta).\ldots.f(x_n|\theta). \qquad (L)$$

The data point will tend to be concentrated where the probability is concentrated. Fisher advocated choosing as our estimate of the (unknown, but non-random) parameter θ, the value(s) $\hat{\theta}$ (or $\hat{\theta}_n$) for which the likelihood $L(\lambda)$ is maximised. This gives the *maximum likelihood estimator* (MLE); the method is the *Method of Maximum Likelihood*. It is intuitive, simple to use and very powerful – 'everyone's favourite method of estimating parameters'.

It is often more convenient to use the *log-likelihood*,

$$\ell := \log L,$$

and maximise that instead (the same, as log is increasing).

Examples.

1. *Normal*, $N(\mu, \sigma)$ (or $N(\mu, \sigma^2)$).

As in IS II, the MLEs are

$$\hat{\mu} = \bar{X}, \qquad \hat{\sigma^2} = S^2 (= \frac{1}{n} \sum_1^n (X_k - \bar{X})^2).$$

But by above, this is biased: to obtain an unbiased estimator for σ^2, we have to use S_u^2 and divide by $n-1$ instead of n. So desirable properties of estimators (e.g. being MLE and unbiased, as here) may be incompatible.

Note. If we use X here (in $X_1, \ldots, X_n$, $\bar{X}$ etc.), we are thinking of the Xs as random variables ("before sampling"). If we use the corresponding lowercase letters, we are thinking of them as data – the numerical values obtained ("after sampling"). We shall feel free to use either, depending on convenience – but the second is customary in Statistics, it is our default option here.

We quote (see e.g. [BF] Th. 2.4) that for $N(\mu, \sigma^2)$

(i) $\bar{X}$ and S^2 are independent;

(ii) $\bar{X} \sim N(\mu, \sigma^2/n)$;

(iii) $nS^2/\sigma^2 \sim \chi^2(n-1)$.

So (by definition of the Student t-distribution)

$$\sqrt{n-1}.\frac{\sqrt{n}(\bar{X} - \mu)/\sigma}{\sqrt{n}S/\sigma} = \sqrt{n-1}(\bar{X} - \mu)/S \sim t(n-1).$$

Note that σ (a nuisance parameter if we are interested in the mean) cancels.

As in IS II:

2. *Poisson* $P(\lambda)$: $\bar{\lambda} = \bar{x}$.
3. *Exponential* $E(\lambda)$: $\bar{\lambda} = 1/\bar{x}$.

The first example is a two-parameter problem, the next two are one-parameter problems. But the first example contains two one-parameter sub-problems:
1a. *Normal* $N(\mu, \sigma^2)$, σ *known*. The calculation above gives $\hat{\mu} = \bar{x}$ again. Note that $\hat{\mu} \sim N(\mu, \sigma^2/n)$ (whether or not σ is known).
1b. *Normal* $N(\mu, \sigma^2)$, μ *known*. The calculation above gives

$$\hat{\sigma^2} = \frac{1}{n} \sum_1^n (x_i - \mu)^2.$$

This is now a statistics, as μ is known – call it S_μ^2. Then (recall that $\chi^2(r)$ is the distribution of the sum of the squares of r copies of standard normals)

$$nS_\mu^2/\sigma^2 \sim \chi^2(n).$$

By contrast, in Ex. 1,

$$nS^2/\sigma^2 \sim \chi^2(n-1)$$

(see e.g. [BF], Th. 2.4). We shall see other differences in Ch. II on Hypothesis Testing: the tests used vary depending on what is known.

2. THE CRAMÉR-RAO INEQUALITY

As above: we like parameter estimates to be unbiased ("get it right on average"). We also like estimates to be precise ("have values close together" – as little randomness as possible). We can think of precision as the reciprocal of the variance, so we like maximum precision, or minimum variance. Thus an ideal estimator is minimum-variance unbiased (MVU), and we shall study such estimators below.

But before we do this, it is important to consider the trade-off between precision and bias. Consider, by analogy, setting the sights for a rifle. Bias concerns whether the weapon fires, say, too high or to the right, etc. Precision concerns the grouping of a number of shots. One would prefer a precision weapon firing a bit high to a blunderbuss, with its shots all over the place but 'right on average'. One can formalise this, using the language of Decision Theory, but we shall not do this.

We now focus on MVU estimators. The remarkable thing is that there are theoretical limits to the accuracy they can attain.

As above, we have a joint density $f = f(x_1, \ldots, x_n; \theta)$, which we write as $f = f(x; \theta)$. This integrates to 1: $\int f(x; \theta) dx = 1$ (where dx is n-dimensional

Lebesgue measure), which we abbreviate to

$$\int f = 1.$$

We assume throughout that $f(x;\theta)$ is smooth enough for use to differentiate under the integral sign (w.r.t. dx, understood) w.r.t. θ, twice. Then

$$\int \frac{\partial f}{\partial \theta} = \frac{\partial}{\partial \theta}\int f = \frac{\partial}{\partial \theta}1 = 0 : \quad \int \big(\frac{1}{f}\frac{\partial f}{\partial \theta}\big).f = 0 : \quad \int \big(\frac{\partial}{\partial \theta}\log f\big).f = 0.$$

Now $E[g(X)] = \int g(x)f(x;\theta)dx = \int gf$, so in probabilistic language this says

$$E\big[\frac{\partial \log L}{\partial \theta}\big] = 0 : \qquad E\big[\frac{\partial \ell}{\partial \theta}\big] = 0 : \qquad E[\ell'(\theta)] = 0.$$

We now introduce the (Fisher) *score function*

$$s(\theta) := \ell'(\theta) : \qquad E[s(\theta)] = 0. \tag{a}$$

Differentiate under the integral sign wrt θ again:

$$\frac{\partial}{\partial \theta}\int \big(\frac{1}{f}\frac{\partial f}{\partial \theta}\big).f = 0, \qquad \int \frac{\partial}{\partial \theta}\big[\big(\frac{1}{f}\frac{\partial f}{\partial \theta}\big).f\big] = 0 :$$

$$\int \big[\big(\frac{1}{f}\frac{\partial f}{\partial \theta}\big)\frac{\partial f}{\partial \theta} + f\frac{\partial}{\partial \theta}\big(\frac{1}{f}\frac{\partial f}{\partial \theta}\big)\big] = 0.$$

As the bracket in the second term is $\partial \log f/\partial\theta$, this says

$$\int \big[\big(\frac{1}{f}\frac{\partial f}{\partial \theta}\big)^2 + \frac{\partial}{\partial \theta}\big(\frac{\partial \log f}{\partial \theta}\big)\big]f = 0, \quad \int \big[\big(\frac{\partial \log f}{\partial \theta}\big)^2 + \frac{\partial^2}{\partial \theta^2}(\log f)\big]f = 0,$$

or as above

$$E\big[\big(\frac{\partial}{\partial \theta}\log L\big)^2 + \frac{\partial^2}{\partial \theta^2}\log L\big] = 0 : \qquad E[\{\ell'(\theta)\}^2 + \ell''(\theta)] = 0 :$$

$$E[s(\theta)^2 + s'(\theta)] = 0. \tag{b}$$

We write

$$I(\theta) := E[\{\ell'(\theta)\}^2] = -E[\ell''(\theta)] : \qquad I(\theta) = E[s^2(\theta)] = -E[s'(\theta)], \tag{c}$$

and call $I(\theta)$ the (Fisher) *information* on θ. By (a) and (c):

Proposition. The score function $s(\theta) := \ell'(\theta)$ has mean 0 and variance $I(\theta)$.

When $x_1, \ldots, x_n$ are independent, the joint density is the product of the marginal densities; so the log-likelihoods ℓ add; so the informations $-E[\ell''] = -E[s']$ (from (c)) add: *the information in a sample of size n is n times the information per reading.* Also from (c), $s^2 \geq 0$, so $E[s^2] \geq 0$: *information is non-negative.* These two properties suggest that the term information is indeed well chosen.

Theorem (Cramér-Rao Inequality, or Information Inequality, H. Cramér (1946), C. R. Rao (1945). Let $Y = u(\mathbf{X})$ be any unbiased estimator of θ. Then the minimum variance bound for $var\ Y$ is

$$var\ Y \geq 1/I(\theta, \mathbf{X}) = 1/(nI(\theta)),$$

where $I(\theta)$ is the information per reading.

Proof. As $Y = u(\mathbf{X})$ is unbiased,

$$\theta = E[u(\mathbf{X})] = \int u(\mathbf{x}) f(\mathbf{x}; \theta) d\mathbf{x} = \int uf.$$

$\partial/\partial\theta$:

$$1 = \frac{\partial}{\partial\theta} \int uf = \int u\Big(\frac{1}{f}\frac{\partial f}{\partial\theta}\Big) f = \int u(\partial \log f/\partial\theta) f :$$

$$1 = E[u\partial \log L/\partial\theta] = E[u\ell'] = E[us].$$

By (a), (b) and (c),

$$var\ s = var\ \ell' = E[(\ell')^2] = I(\theta; \mathbf{X}), = I(\theta),$$

say. The correlation coefficient is

$$\rho := \rho(u, s) = \frac{cov(u, s)}{\sqrt{var\ u}\sqrt{vars}} = \frac{E[us] - E[u]E[s]}{\sqrt{var\ u}\sqrt{I}} = \frac{1}{\sqrt{var\ u}\sqrt{I}},$$

as $E[s] = 0$, $E[us] = 1$. But $\rho^2 \leq 1$ (correlation bound: Cauchy-Schwarz Inequality). So

$$var\ u \geq 1/I. \qquad //$$

Defn. We call an estimator *efficient* if it is unbiased and its variance achieves the CR lower bound, *asymptotically efficient* if its bias tends to 0 and its variance achieves the CR bound asymptotically.

An efficient (= minimum-variance unbiased, MVU) estimator is also called a *best* estimator.

When dealing with regression (Ch. IV), we shall often meet *linear* estimators; the above then become *BLUEs* (best linear unbiased estimators).

Iterative solution of the Likelihood Equation

It may not be possible to solve the Likelihood Equation $\ell' = 0$ (LE) in closed form. In such cases, we have to proceed as elsewhere in Mathematics – in particular, in Numerical Analysis – and proceed iteratively.

To assess the problem, begin by drawing a rough graph of ℓ. By looking for sign changes, and using trial values, it is usually possible (without excessive effort) to find a rough approximation to the desired root (there may – will in general – be multiple roots, but usually the root we need will be clear enough from context). Call this trial value t. Then (with $s = \ell'$)

$$0 = s(\hat{\theta}) = s(t) + ((\hat{\theta} - t)s'(\theta^*),$$

with θ^* between t and $\hat{\theta}$. Solving,

$$\hat{\theta} = t - s(t)/s'(\theta^*). \qquad (*)$$

We now have a choice about how to proceed. We know that $\hat{\theta}$ is (strongly) consistent, $\hat{\theta} \to \theta_0$, so $\hat{\theta} \sim \theta_0$, so with a good enough starting value t, also $t \sim \hat{\theta} (\sim \theta_0)$ and $\theta^* \sim \hat{\theta} (\sim \theta_0)$.

Newton-Raphson iteration.

This is also known as the *tangent approximation*. It relies on replacing a function by its tangent near a point. If x_n is near a root of

$$f(x) = 0,$$

then a better approximation is

$$x_{n+1} := x_n - f(x_n)/f'(x_n).$$

So starting from the approximation t, replacing $s'(\theta^*)$ in $(*)$ by $s'(t)$ gives a better approximation; this is the Newton-Raphson method.