introstat.tex

INTRODUCTORY LECTURES ON STATISTICS. 6-7.10.2011

N. H. BINGHAM

Syllabus.
I. Simulation (Monte Carlo, ...), Th. 6, 140, 10-12 am,
II. Estimation (Maximum likelihood, etc.), Fri 7, 140, 10-12 am.

I. SIMULATION

Our raw material here will be a sequence of (independent) random numbers x_n, each uniformly distributed on the unit interval: $x_n \sim U[0,1]$.

The point is that we want the numbers to be *random*. But we cannot strictly achieve this except by genuine sampling. This is awkward, indeed impossible to do in practice if we need very many random numbers. Instead, we use a computer to generate a (perhaps very long) sequence of *pseudo-random numbers* – numbers that are not random at all but deterministic, but which 'look random'.

That this may be possible is familiar. Think of mathematical tables, such as of logs or antilogs, trig functions, etc – say, four-figure tables. The *first* digits are informative, and systematic. The *last* digits are not: they are determined by rounding error from the first digit not displayed, and look like mere 'noise'.

Note. You should compare the numerical behaviour of first digits with that of last digits. For last digits, you would expect each of the ten possibilities 0,1,..,9 to occur with equal frequency 1/10 in the long run. They do; you can check this. (This is an instance of the 'Law of Averages', below, or the Strong Law of Large Numbers (SLLN) [Stochastic Processes II.10, L14].) By contrast, the first digits show decreasing frequency from 1 to 9! You should
(i) check this for yourselves numerically;
(ii) then check out the theory here – *Benford's Law*: Frank BENFORD (1883-1948) in 1938.

Our main theoretical tool for generating such pseudo-random sequences are *congruential generators*:

$$x_{n+1} := ax_n + c \qquad (\text{mod } m).$$

These were introduced by D. H. LEHMER (1905-1991) in 1948. We shall take it for granted that such congruential generators 'work'; for background, see e.g.
D. E. KNUTH, *The Art of Computer Programming*, Vol. 2: *Semi-numerical algorithms*, Addison-Wesley, 1969, Ch. 3.
Note. Donald E. Knuth (1938–) is also the inventor of TeX (pronounced 'tech' – ch as in 'loch'), in 1978. This is now known as plainTeX (which I prefer); more widely used nowadays, and recommended in this MSc, is LaTeX (Leslie LAMPORT (1941–) in 1986).

The uniform distribution (or more briefly, uniform law) $U[0,1]$ models *probability = length*:

$$P(a \le X \le b) = b - a \qquad (0 \le a \le b \le 1).$$

For X a random variable and F its distribution function

$$F(x) := P(X \le x)$$

(see e.g. MSF3 Stochastic Processes, Ch. II): F is non-decreasing and right-continuous ($F(x+) = F(x)$; $F(x-) \le F(x)$, with $<$ if x is a jump point of F); F increases from 0 at $-\infty$ to 1 at $+\infty$. Its inverse function

$$F^{-1}(t) := \inf\{x : F(x) \ge t\}$$

is similarly non-decreasing, but is left-continuous. So the infimum is attained, and so is a minimum:

$$F^{-1}(t) := \min\{x : F(x) \ge t\}.$$

Proposition (Probability Integral Transformation, PIT). If $U \sim U[0,1]$, then $X := F^{-1}(U) \sim F$.

Proof.

$$P(X = F^{-1}(U) \le x) = P(U \le F(x)) = F(x). \qquad //$$

So as we can generate a sequence of uniforms (above), we can hence generate a sequence sampled from any given distribution F.
Examples.
1. *Coin-tossing.* We can generate coin-tosses by taking tails (T, or 0) if $U < 1/2$, heads (H, or 1) if not.

2. *Rolling dice.* Similarly, we can generate die-rolls by taking the outcome as 1 if $0 \leq U < 1/6$, 1 if $1/6 \leq U < 1/3$, etc.
3. *Dealing bridge hands.* There are 52 cards in a pack (4 suits, spades ♠, hearts ♡, diamonds ♢, clubs ♣, 13 cards per suit – 2,..., 10, J, Q, K, A). By labelling each card with a number from 1 to 52, we can as above 'deal each card uniformly' (make it equally likely that each player N, S, E, W gets it). [To make each player get 13 cards is more tricky, and involves 'sampling without replacement', or we can proceed as follows. Proceed with prob. 1/4 each until one player has 13 cards, then with prob. 1/3 each till the next has 13, then with prob. 1/2 till the next, then the remaining player gets the rest.]
4. *Tennis.* In a tennis game, suppose the server wins each point with probability p. What is the probability that the server wins the game? the set? the match?

Draw a picture of a game of tennis.
In the bottom row is the starting point, 0-0.
In the first row are the two possibilities after 1 point, 15-0 and 0-15.
In the second row are the three possibilities after 2 points, 30-0, 15-15, 0-30.
In the fourth row are the four possibilities after 3 points, 40-0, 30-15, 15-30, 0-40.
In the fifth row are two of the possibilities after 4 points in which the game continues: 15-40 and 40-15.
In the top [6th] row are the five remaining possibilities after 4 or 5 points: Win (for server), 40-30, 30-30, 30-40, Lose.
If the game continues, we can combine 'Advantage in' with 40-30, 'Deuce' with 30-30, 'Advantage out' with 30-40 (you should check this!).
Note. 1. 40 is short here for 45. Formerly each player had a clock, and each point was worth 1/4 of a revolution, i.e. 15 seconds (say). To win, a player had to complete a revolution but be more than one point ahead.

With a flow-diagram to represent these 17 states, one can 'play tennis' by computer, with each point leading to the upper left and right neighbours with probabilities $p, 1-p$.
2. The general context of such flow-diagrams is that of *finite Markov chains.* See e.g. John G. KEMENY and J. Laurie SNELL, *Finite Markov chains*, Van Nostrand, 1960 (tennis, 7.2 p. 161-7),
Olle HÄGGSTRÖM, *Finite Markov chains and algorithmic applications*, CUP, 2002.
5. *Binomial tree.* In a discrete financial model, suppose that at each stage

the price of a risky asset can go up or down. One can model the price evolution over a finite time-period (say, from the start of an option to its expiry) by a *binomial tree*, with two paths leading from each node, one 'up right', one 'down right'. This gives the *Cox-Ross-Rubinstein binomial tree model*. This leads to the *discrete Black-Scholes model* for option pricing. Lurking in the background here is the binomial distribution, which in the limit gives the normal distribution (Central Limit Theorem (CLT), or 'Law of Errors' – below). This leads from the discrete Black-Scholes formula to the more familiar and famous (continuous) *Black-Scholes formula* of mathematical finance.

Densities.
1. *Exponential distribution*, $E(\lambda)$ ($\lambda > 0$ is the parameter):

$$f(x) := \lambda e^{-\lambda x} \quad (x > 0), \quad 0 \quad (x \leq 0).$$

The distribution function is

$$F(x) = \int_0^x f(u)du = 1 - e^{-\lambda x} \qquad (x > 0);$$

$$F^{-1}(u) = -\lambda^{-1} \log(1-u).$$

So by PIT, if $U \sim U[0,1]$,

$$F^{-1}(U) = -\lambda^{-1} \log(1-U) \sim E(\lambda).$$

So we use instead

$$F^{-1}(U) = -\lambda^{-1} \log U \sim E(\lambda),$$

since $1 - U \sim U[0,1]$ also. Note that the two last formulae are equivalent mathematically, but not computationally: it would be avoidably inefficient, and so count as a programming error, not to use the second.
2. *Gamma distribution* $\Gamma(n, \lambda)$.

$$f(x) = e^{-\lambda x} \lambda^n x^{n-1}/(n-1)! \qquad (x > 0).$$

This has moment-generating function (MGF) $(\lambda/(\lambda - t))^n$, and is the distribution of the sum of n (independent) copies of $E(\lambda)$.
3. *Chi-square* $\chi^2(n) = \Gamma(n/2, 1/2)$. If the X_i are independent copies of standard normal (below),

$$X_1^2 + \ldots + X_n^2 \sim \chi^2(n).$$

4. *Standard normal distribution*, $\Phi = N(0,1)$. Density ϕ and distribution Φ:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \qquad \Phi(x) = \int_{-\infty}^{x} \phi(u)du = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2} du.$$

There is no closed form for Φ (except that $\Phi(0) = 1/2$ by symmetry), so none for Φ^{-1}. So how do we use PIT to simulate from Φ? One method is the *Box-Muller method* of 1958. Recall that the element of area dA is $dx\ dy$ in plane cartesian coordinates and $r\ drd\theta$ in plane polar coordinates. If X, Y are independent $N(0,1)$,

$$\begin{aligned} P(X^2 + Y^2 \le R^2) &= \int\int_{x^2+y^2 \le R^2} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} . \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dxdy \\ &= \int_0^R \int_0^{2\pi} \frac{e^{-r^2/2}}{2\pi} . r dr d\theta \\ &= \int_0^R \frac{e^{-r^2/2}}{2\pi} . d(r^2/2) \\ &= 1 - e^{-R^2/2}. \end{aligned}$$

This says that $r^2 := X^2 + Y^2 \sim E(1/2)$, the exponential law with parameter $1/2$ (mean 2). We know how to simulate from this, by above!

$\theta \sim U[0, 2\pi]$ by symmetry: simulate by $\theta = 2\pi U$. So take U_1, U_2 (independent) $\sim U[0,1]$,

$$r^2 := -2 \log U_1 \quad (\sim E(1/2)),$$

$$r := \sqrt{-2 \log U_1}, \qquad \theta := 2\pi U_2.$$

Then $X := r \cos\theta$, $Y := r \sin\theta$ are independent $N(0,1)$.

Rejection Method.

This is due to John von NEUMANN (1903-1957) in 1951. Suppose we have a density f. Then the area under the curve is 1. The *subgraph* of f is $\{(x,y) : 0 \le y \le f(x)\}$. So the area of the subgraph is 1. By definition of density,

$$P(X \in [x, x+dx]) = f(x)dx = dA,$$

where A denotes area under the subgraph to the left of x. So ('probability = area') X has density f iff X is the x-coordinate of a point uniformly distributed over the subgraph of f.

Suppose we have a density g that we know how to simulate from, and a density f that we don't know how to simulate from, but

$$f(x) \leq cg(x)$$

for all x and some constant c. We proceed as follows.
1. Simulate from g, i.e. by above 1*. Sample points *uniformly* from the subgraph of g.
2. Stretch the positive y-axis by a factor c.
The points are still uniformly distributed over the subgraph of cg.
3. *Reject* all point not in the subgraph of f (contained in the subgraph of cg, as $f \leq cg$). The remaining points are still uniform, but over the subgraph of f not cg. So:
4. The x-coordinates of the points have density f.

The step that needs checking is 3 – that the non-rejected points are still uniform, but over the subgraph F of f rather than the subgraph G of cg. Before the rejection step, X is uniform over G:

$$X \sim U(G); \qquad P(X \in A) = |A|/|G|, \quad A \subset G$$

(writing $|.|$ for area). Now for $B \subset F$, the distribution of the non-rejected points (i.e. of the points conditional on their being in F) is given by

$$P(X \in B | X \in F) = P(X \in B \& X \in F)/P(X \in F) = P(X \in B \cap F)/P(X \in F)$$

$$= \frac{|B \cap F|}{|G|} / \frac{|F|}{|G|} = |B \cap F|/|F|.$$

This says that the non-rejected points are uniform over F, the subgraph of f, i.e. that they have density f, as required. //

Note. The closer the graph of f is to that of cg, the fewer points are rejected, and the greater the computational efficiency. For heavy computational use, it is worth making an effort to achieve such a 'good fit', but for details we must refer to a specialist book on simulation.

Monte Carlo Method

If we are to evaluate an integral $\int f(x)dx$ (typically in several or many dimensions), we may be able to interpret it as an expectation, $Ef(X)$ for

some random variable X. Then if $X_1, \ldots, X_n, \ldots$ are independent random variables with the same distribution, the SLLN gives

$$\frac{1}{n}\sum_1^n f(X_k) \to Ef(X) \qquad (n \to \infty) \qquad a.s.$$

(a.s. = almost surely, or with probability one – see MSF3 Stochastic Processes). The idea of the *Monte Carlo method* is to simulate the X_k on the left, form the average on the left numerically, and use it as an approximation to the expectation or integral on the right. The method is widely used, and very powerful.

The idea can be traced back to Buffon's needle (G. L. Leclerc, Comte de BUFFON (1707-1788) in 1777), but is due in its modern form to Stanislaw ULAM (1909-1984) in 1946. It emerged in work by physicists at the Los Alamos Laboratory (Manhattan Project, WWII, atom (fission) bomb, pre-computer, then 1950s, hydrogen (fusion) bomb, with computers).

There is a whole area of Statistics called Markov Chain Monte Carlo or MCMC, based on this idea (Professor Alastair Young of the Statistics Section here is the local expert).

Random numbers and π.

Recall that π is defined to be the ratio of the circumference of a circle to its diameter, and $\pi = 3.1415926535...$ Mnemonic:

Que j'aime à faire apprendre

Ce nombre utile aux sages.

Nothing could be less random than π! But π is so important, and interesting, that its decimal expansion has been calculated to billions of places. As the brothers David and Gregory Chudnovsky (who have been prominent in this) put it, 'Pi is a damned good fake of a random number'.

What properties should a random (real) number have? – or what properties should a 'typical' real have? One obvious one is absence of a preference for one of the digits 0,1,..,9 above another: each digit should occur with its 'correct' asymptotic frequency, 1/10. That this happens is part of Borel's Normal Number Theorem of 1909, a consequence of the Strong Law of Large Numbers mentioned earlier. This property is called (strong) *normality*; Borel's result says that *almost all* reals are strongly normal. But there is nothing special about the base 10 of decimals: we can use the base 2 of binary (as computers do), etc. Borel's result says more: almost all reals are strongly normal to *all bases simultaneously.*

Now that we know that almost all reals behave like this, it would be nice to have a specific example – and π is the obvious candidate. The decimal expansion of π has been subjected to every statistical test for randomness known to Statistics – and passed them all with flying colours (hence the Chudnovsky quotation above). This strongly *suggests* that π is indeed normal – but does not *prove* it. Indeed, there is no reason to suppose that we will *ever* be able to prove this. From the point of view of Number Theory, there is only one natural way to expand π, and that is as a *continued fraction*:

$$\pi/4 = \cfrac{1}{1+\cfrac{1^2}{2+\cfrac{3^2}{2+\cfrac{5^2}{\ddots}}}}$$

(William, Lord BROUNCKER (1620-1684), in 1655 – related to Wallis' product for π, for which see e.g. M2PM3, L32).

Of course we all know that π begins 3.14159..., so the string "14159" has us all thinking "pi". But if we start after, say, a million places, the expansion would "look random". One can even personalise this. My date of birth is 19.03.1945; the decimal expansion of π started after 19,031,945 places would be perfectly deterministic and predictable to me, but would look perfectly random to anyone who did not know this.

Such examples make one think about *what randomness is*. Consider, for example, a thousand independent tosses of a fair coin (heads = 1, tails = 0). There are $2^{1,000}$ possible outcomes, each with equal probability $2^{-1,000}$ by symmetry. Imagine two such outcomes, (i) one obtained by you, laboriously tossing a coin a thousand times – or, simulating as above, (ii) all 1s. There is one sense in which these are on the same footing (symmetry, above). There is another in which they are obviously not: the first takes a thousand bits of information to describe, the second takes two. Lurking in the background here is a whole subject – Algorithmic Information Theory.

II. ESTIMATION (of parameters); LIKELIHOOD

We start with some examples.
Normal, $N(\mu, \sigma)$ (μ real, $\sigma > 0$), density

$$f(x|\mu, \sigma) := \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{1}{2}(x-\mu)^2/\sigma^2\}.$$

Here μ is the *mean*, σ^2 the *variance* ($\sigma > 0$ the *standard deviation*, or SD); μ, σ are the *parameters* (the term is due to R. A. (Sir Ronald) FISHER (1890-1962) in 1922).
Exponential $E(\lambda)$ ($\lambda > 0$): $f(x|\lambda) := \lambda e^{-\lambda x}$, $x > 0$.
Uniform, $U[a, b]$ ($a < b$): $f(x|a, b) := 1/(b-a)$ on $[a, b]$, 0 elewhere.
Poisson, $P(\lambda)$ ($\lambda > 0$): $f(k|\lambda) := e^{-\lambda}\lambda^k/k!$ ($k = 0, 1, \ldots$).

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 MSF3 Ch. 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:

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

which we may abbreviate to

$$f(., \ldots, .|\theta) = f(.|\theta).\ldots.f(.|\theta),$$

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 (as log is increasing, maximising L and ℓ are the same). *Examples.*

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

$$L = \frac{1}{\sigma^n 2\pi^{n/2}} . \exp\{-\frac{1}{2}\sum_1^n (x_i - \mu)^2/\sigma^2\},$$

$$\ell = const - n \log \sigma - \frac{1}{2}\sum_1^n (x_i - \mu)^2/\sigma^2.$$

$$\partial\ell/\partial\mu = 0 : \quad \sum_1^n (x_i - \mu) = 0, \quad \mu = \frac{1}{n}\sum_1^n x_i.$$

So the MLE of the (population) mean μ is the *sample mean* (average of the data points):

$$\hat{\mu} = \bar{x}, \qquad \overline{x} := \frac{1}{n}\sum_1^n x_i.$$

This makes sense: one would hope and expect that the sample mean is informative about the population mean. Indeed, by SLLN,

$$\hat{\mu} = \overline{X} \to EX \qquad (n \to \infty) \qquad a.s.$$

(we revert to capitals for random variables; we use lower case for data values, = observed values of random variables).

$$\partial\ell/\partial\sigma = 0 : \quad -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_1^n (x_i - \mu)^2 = 0, \quad \sigma^2 = \frac{1}{n}\sum_1^n (x_i - \mu)^2.$$

At the maximum, $\mu = \overline{x}$ (above), giving the MLE of σ^2 as

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

The RHS is called the *sample variance*, S^2. So the MLEs of the population mean and variance μ, σ are the sample mean and variance $\overline{x}, S^2$.

Note. 1. Many authors use $1/(n-1)$ in place of $1/n$ in the definition of the

sample variance (this is needed to get the estimate *unbiased*). But for large n, there is little difference.
2. We can extend the bar notation:

$$\hat{\sigma}^2 = \overline{(X - \overline{X})^2} = \overline{X^2 - 2X\overline{X} + \overline{X}^2} = \overline{X^2} - 2\overline{X}^2 + \overline{X}^2 = \overline{X^2} - \overline{X}^2.$$

Then by SLLN,

$$\hat{\sigma}^2 = \overline{X^2} - \bar{X}^2 \to E[X^2] - [EX]^2 = E[(X - EX)^2] = var(X) = \sigma^2.$$

Thus the bar notation is ideally suited to use of SLLN. We can show similarly that the (suitably defined) sample covariance and correlation tend to the corresponding population covariance and correlation, etc.
3. The above shows clearly that desirable properties of estimators (e.g. being MLEs and being unbiased) may be incompatible.
2. *Exponential.* For $E(\lambda)$, the mean $EX = 1/\lambda$.

$$L = \lambda^n \exp\{-\lambda \sum_1^n x_i\} = \lambda^n \exp\{-n\lambda\bar{x}\}, \quad \ell = n \log \lambda - n\lambda\bar{x}.$$

$$\partial\ell/\partial\lambda = 0 : \quad n/\lambda = n\bar{x} : \hat{\lambda} = 1/\bar{x}.$$

Again, this is natural and what we would expect.
3. *Poisson*, $P(\lambda)$. Recall that this has mean λ (and also variance λ). Writing the data again as x_i (these are non-negative integers k_i),

$$L = e^{-n\lambda} \prod_1^n \lambda^{x_i} / \prod_1^n x_i!, \quad \ell = -n\lambda + \sum_1^n x_i \log \lambda - \sum_1^n \log x_i!$$

$$\partial\ell/\partial\lambda = 0 : \quad -n + \sum_1^n x_i/\lambda = 0, \quad \hat{\lambda} = \frac{1}{n}\sum_1^n x_i = \bar{x}.$$

Again, this is natural, and what we would expect.
4. *Uniform* $U[a, b]$. Here

$$L = 1/(b-a)^n \qquad (a \le x_1, \ldots, x_n \le b), \qquad 0 \text{ otherwise},$$

or (with $\min := \min(x_1, \ldots, x_n)$, and similary for max)

$$L = (b-a)^{-n} I(a \le \min, \max \le b).$$

We are to maximise this wrt a, b. Don't use calculus as above (we can't – the RHS is discontinuous, so not differentiable). Instead, we can do the maximising of L on sight: the MLEs are

$$\hat{a} = \min, \qquad \hat{b} = \max .$$

Note. We shall see later that this example is less well behaved, and different from, the ones above.

SUFFICIENCY (Data Reduction).

If in the expression for the normal likelihood we substitute $x_i - \mu = (x_i - \bar{x}) + (\bar{x} - \mu)$, square and expand, we get (as $\sum(x_i - \bar{x}) = 0$, so the cross-terms vanish)

$$L = \frac{1}{\sigma^n 2\pi^{n/2}} . \exp\{-\frac{n}{2}[S^2 + (\bar{x} - \mu)^2]\}.$$

This involves the data $x_1, \ldots, x_n$ only through $\bar{x}$ and S^2 (equivalently, only through $\bar{x}$ and $\overline{x^2}$). Suppose $n = 1,000$, and we record these two statistics, but lose the data. Does it matter? There are two plausible views:
(i) Yes. A thousand numbers are more informative than two.
(ii) No. As above, we expect $\bar{x}$, S^2 to be informative about μ, σ^2, and the fact that the likelihood only involves these confirms this.

In fact the optimistic view (ii) is the correct one: we can reduce the data set of 1,000 values down to just *two*, without loss of information. This is the idea of *sufficiency*, due to Fisher in 1920. It can be formulated in various equivalent ways, one of which is that above: we say that a statistic T is *sufficient* for a parameter θ if the likelihood factorises into factors, one which involves the data only through T (and may involve the parameters), the other free of the parameters.

Sufficiency (or data reduction) is such a good idea that it should always be used, to reduce the data. We would naturally like to be able to reduce the data as much as possible (without loss of information). This is the idea of *minimal sufficiency* (Lehmann and Scheffé – omitted).
Exampe: Uniform distribution. The form of the likelihood above shows that $(\min, \max)$ is sufficient for (a, b). [It is also minimal sufficient, but this is clear: we couldn't expect a further reduction, to a one-dimensional statistic, to suffice for a two-dimensional parameter.]

The Likelihood Principle (LP).

The likelihood is of central importance. The *Likelihood Principle* (LP) says in effect that the likelihood is all that matters. The LP is due implicitly to Fisher in the 1920s, and to G. A. BARNARD (1915-2002) in the 1940s, explicitly to Alan BIRNBAUM in 1962. We shall use it informally (a full discussion would involve foundational questions in Statistics).

Limit Theorems

Laws of Large Numbers.

The 'Law of Averages' is known to the man/woman in the street. It says that in the long run things happen about as often as they should (e.g., when tossing a fair coin one gets heads about half the time). The precise form of this is the Law of Large Numbers (LLN) in Probability Theory. There are actually many forms of LLN; see MSF3 Stochastic Processes for two, the weak LLN [II.7, L12] and strong LLN [II.10, L14].

Central Limit Theorem.

The 'Law of Errors' is known to the physicist in the street, as 'errors are normally distributed about the mean'. The precise form is the Central Limit Theorem (CLT); see MSF3 [II.7, L12].

Large-sample theory of MLEs.

The idea of *information per reading* about a parameter θ, $I(\theta)$, (which goes back to Fisher in 1934) can be made precise. It is related to the *Cramér-Rao* inequality, or *information inequality*, which gives a minimum-variance bound for unbiased estimators (we must omit this through lack of time; see the index of any good book on Statistics).

Under suitable regularity conditions, one can show that
(i) the MLE is *consistent*:

$$\hat{\theta} \to \theta \qquad (n \to \infty) \qquad a.s.;$$

(ii) the MLE is *asymptotically normal*, with convergence *rate* $\sqrt{n}$ (or $1/\sqrt{n}$, depending on usage):

$$(\overline{\theta} - \theta)\sqrt{n}\sqrt{I(\theta)} \to \Phi = N(0,1) \qquad (n \to \infty) \qquad \text{in distribution.}$$

This qualifying phrase 'under suitable regularity conditions' is ubiquitous in large-sample theory in Statistics. What is needed is that the model be smooth enough to justify differentiating under the integral sign wrt θ twice (this is what is needed to get the Cramér-Rao lower bound). Some such condition is needed, as the following example shows.

Example: Uniform distribution, $U[a,b]$. Here $\hat{a} = \min$, $\hat{b} = \max$. So

$$P(n(\hat{a}-a)/(b-a) \geq x) = P(n(X_i - a)/(b-a) \geq x, i = 1, \dots n)$$

$$= [P(n(X_i - a)/(b-a) \geq x]^n = (1 - x/n)^n \to e^{-x} \qquad (n \to \infty).$$

That is,

$$n(\hat{a}-a)/(b-a) \to E(1) \qquad (n \to \infty) \qquad \text{in distribution,}$$

(see MSF3 Ch. II), and similarly

$$n(b-\hat{b})/(b-a) \to E(1) \qquad (n \to \infty) \qquad \text{in distribution.}$$

Note the contrast to the above! There, the limit is normal, and the rate of convergence is $\sqrt{n}$. Here, the limit is exponential $E(1)$, and the rate of convergence is n, much faster. It looks as if this faster rate of convergence is an advantage. So it is – but we pay a heavy price for it. We have no protection against contamination of the data. If *one* of the data points is way too small (say), it will permanently affect the sample minimum, and so the estimate of a. By contrast, in the examples above, the influence of one bad data point will be damped out as the sample size increases.

The branch of Statistics concerned with such protection against data contamination is called Robust Statistics. The phenomenon noted above ('too rapid convergence') is called *super-efficiency* – and indicates extreme non-robustness. When the 'suitable regularity conditions' hold, we have a *regular* maximum-likelihood estimation problem. The example above of the uniform distribution $U[a,b]$ is *non-regular*. The non-regularity results from the *support* of the distribution (the region where it is positive) depending on the parameters (the support is $[a,b]$).

Uniform distribution (continued).

Suppose now that the length $b-a$ of the interval in $U[a,b]$ is known. Then we can w.l.o.g. take it as 1. This is now a *one*-parameter problem rather

than a two-parameter one; the natural choice of parameter is the mid-point, θ, giving $U[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$. The likelihood now is

$$L(\theta) = 1 \quad (\theta - \frac{1}{2} \leq x_1, \ldots, x_n \leq \theta + \frac{1}{2}) : \quad L(\theta) = 1 \quad (\theta - \frac{1}{2} \min, \max \leq \theta + \frac{1}{2}).$$

The maximum value is 1, but this is now attained on an entire *interval*, $[\min, \max]$. So we have *infinitely many* MLEs! Of course, we prefer the symmetrical choice, of the *sample mid-range*. This is the average of max and min, while the *sample range* is their difference:

$$\text{mid} := \frac{1}{2}(\max + \min), \qquad \text{ran} := \max - \min .$$

The argument above shows that the MLE mid is super-efficient for θ, with convergence rate n as for $U[a, b]$, but now the limit law is the symmetric exponential, $SE(1)$, with density

$$f(x) = \frac{1}{2} \exp\{-|x|\} :$$

$$n(\text{mid} - \theta) \to SE(1) \qquad (n \to \infty) \qquad \text{in distribution.}$$

The argument above also shows that $\{\min, \max\}$ is (minimal) sufficient for θ, but mid is not! So now we have (i) a non-unique MLE; (ii) a two-dimensional minimal sufficient statistic $\{\min, \max\}$ – equivalently, $\{\text{mid}, \text{ran}\}$ – for a one-dimensional parameter θ.

Note. This shows that ran is relevant. But this is strange: the parameter θ is a *location* parameter, while $\text{ran} = \max - \min$ is translation-invariant, and so tells us nothing about the location of the interval, i.e. about θ. True – but what ran *does* tell us about is the accuracy of mid as an estimator for θ. If we are lucky and get a 'good' sample, max and min will be nearly as far apart as they can be (1), and their average mid will be close to θ. But if we are unlucky and have a 'bad' sample, they may be close, and then their average mid may be nearly as far away from θ as it can be (1/2). So ran is uninformative about θ on its own, but $\{\text{mid}, \text{ran}\}$ is more informative than mid alone. We say that ran is *ancillary* for θ (Fisher, 1934).

Higher dimensions

The above goes through with minimal changes in higher dimensions.

Normal distribution, $N(\mu, \Sigma)$. In d dimensions, the mean μ is a d-vector and

the covariance matrix $\Sigma = (\sigma_{ij})$ is a positive definite symmetric $d \times d$ matrix. The sample mean $\bar{x}$ (a d-vector) and sample covariance matrix S (a $d \times d$ matrix) can be defined. They are the MLEs for μ, Σ, and by the LLN

$$\bar{x} \to \mu, \qquad S \to \Sigma.$$

Also $(\bar{x}, S)$ are (minimal) sufficient for (μ, Σ). See e.g.
N. H. BINGHAM and John M. FRY, *Regression: Linear Models in Statistics*, SUMS (Springer Undergraduate Mathematics Series), 2010.

The information (per reading) $I(\theta)$ becomes the *information matrix*. With these changes, the above theory goes through much as before.

Method of Moments

This is due to Karl PEARSON (1857-1936) in the 1880s. It is generally inferior to the Method of Maximum Likelihood, but can be useful. It consists of estimating parameters by matching sample moments to population moments.

Example: The Binomial distribution, $B(k, p)$. Here $k = 1, 2, \ldots$, $p \in (0, 1)$,

$$P(X = k) = \binom{k}{i} p^i (1-p)^{k-i}.$$

This counts the number of successes in k Bernoulli trials (or heads in k tosses of a biased coin).

Now assume that both p and k are unknown. Sample $x_1, \ldots, x_n$ from $B(k, p)$. The first two sample moments are $\overline{x}$, $\overline{x^2} = S^2 + [\overline{x}]^2$. The mean and variance of the Bernoulli are $\mu = kp$, $\sigma^2 = kp(1-p) = kp - kp^2$. Matching moments gives two equations for the two unknown parameters:

$$\overline{x} = kp, \qquad \overline{x^2} = kp(1-p) + k^2 p^2.$$

So

$$S^2 = kp - kp^2, \qquad \overline{x} = kp,$$

or $\overline{x} - S^2 = kp^2$, $(\overline{x})^2 = k^2 p^2$, giving the estimates

$$\tilde{k} = (\overline{x})^2 / (\overline{x} - S^2), \qquad \tilde{p} = \overline{x}/\tilde{k}.$$

Application: under-reported crime (example: rape); k is the number of crimes, p is the probability that a crime is reported.

NHB