smfd4.tex
Day 4. 24.10.2013.

3. Likelihood Ratio Tests

We turn now to the general case: *composite* H_0 v. *composite* H_1. We may not be able to find UMP (best) tests. Instead, we seek a general procedure for finding *good* tests.

Let θ be a parameter, H_0 be a null hypothesis – a set of parameter values $\mathbb{T}_0$, such that H_0 is true iff $\theta \in \mathbb{T}_0$, and similarly for H_1, $\mathbb{T}_1$. It is technically more convenient to take H_1 *more general than* H_0, and we can do this by replacing H_1 by "H_1 or H_0". Then $\mathbb{T}_0 \subset \mathbb{T}_1$.

With L the likelihood, we write

$$L_0 := \sup_{\theta \in \mathbb{T}_0} L(\theta), \qquad L_1 := \sup_{\theta \in \mathbb{T}_1} L(\theta).$$

As with MLE: the size of L_1 is a measure of how well the data supports H_1. So to test H_0 v. H_1, we use test statistic the *likelihood ratio (LR)* statistic,

$$\lambda := L_0/L_1,$$

and critical region: *reject* H_0 *if* λ *is too small*.

Since $\mathbb{T}_0 \subset \mathbb{T}_1$, $L_0 \leq L_1$, so

$$0 \leq \lambda \leq 1.$$

In standard examples, we may be able to find the distribution of λ. But in general this is hard to find, and we have to rely instead on large-sample asymptotics.

Theorem (S. S. WILKS, 1938). If θ is a one-dimensional parameter, and λ is the likelihood-ratio statistic for testing $H_0 : \theta = \theta_0$ v. $H_1 :$ θ unrestricted, then under the usual regularity conditions for MLEs (I.3),

$$-2 \log \lambda \to \chi^2(1) \qquad (n \to \infty).$$

Proof. $\lambda = L_0/L_1$, where $L_0 = L(\mathbf{X}; \theta_0)$, $L_1 = L(\mathbf{X}; \hat{\theta})$, with $\hat{\theta}$ the MLE (I.1). So

$$\log \lambda = \ell(\theta_0) - \ell(\hat{\theta}) = \ell_0 - \ell_1,$$

say. But

$$\ell(\theta_0) = \ell(\hat{\theta}) + (\theta_0 - \hat{\theta})\ell'(\hat{\theta}) + \frac{1}{2}(\theta_0 - \hat{\theta})\ell''(\theta^*),$$

with θ^* between θ_0 and $\hat{\theta}$, by Taylor's Theorem. As $\hat{\theta}$ is the MLE, $\ell'(\hat{\theta}) = 0$. So

$$\log\lambda = \ell_0 - \ell_1 = \frac{1}{2}(\theta_0 - \hat{\theta})^2\ell''(\theta^*), \qquad -2\log\lambda = (\theta_0 - \hat{\theta})^2[-\ell''(\theta^*)].$$

By consistency of the MLE (I.3), $\hat{\theta} \to \theta_0$ a.s. as $n \to \infty$. So also $\theta^* \to \theta_0$ (as θ^* is between θ_0 and $\hat{\theta}$). So

$$\begin{aligned} -\ell''(\theta^*) &= -\ell''(\mathbf{X};\theta^*) = n.\frac{1}{n}\sum_1^n[-\ell''(X_i;\theta^*)] \\ &\sim nE[-\ell''(X_i;\theta^*)] \qquad \text{(LLN)} \\ &= nI(\theta^*) \qquad \text{(definition of information per reading)} \\ &\sim nI(\theta_0) \qquad (\theta^* \to \theta_0). \end{aligned}$$

By I.3,

$$(\hat{\theta} - \theta_0)\sqrt{nI(\theta_0)} \to \Phi, \qquad (\hat{\theta} - \theta_0)^2.nI(\theta_0) \to \Phi^2 = \chi^2(1),$$

using Φ^2 as shorthand for 'the distribution of the square of a standard normal random variable'. So

$$-2\log\lambda \to \chi^2(1). \qquad //$$

Higher Dimensions. If $\theta = (\theta_r, \theta_s)$ is a vector parameter, with

θ_r an r-dimensional parameter of interest,

θ_s an s-dimensional nuisance parameter,

to test $H_0 : \theta_r = \theta_{r,0}$ (which is *composite* unless $s = 0$) v. $H_1 : \theta_r$ unrestricted. Similar use of the large-sample theory of MLEs for vector parameters (which involves Fisher's *information matrix*) gives

Theorem (Wilks, 1938). Under the usual regularity conditions,

$$-2\log\lambda \to \chi^2(r) \qquad (n \to \infty).$$

Note that the dimensionality s of the nuisance parameter plays no role: what counts is r, the dimension of the parameter of interest (i.e., the difference in dimension between H_1 and H_0). (We think here of a fully specified

parameter, as in H_0, as a point – of dimension 0, and of H_1 of dimension r, like θ_r. There need not be any vector-space structure here. Recall that degrees of freedom (df) correspond to effective sample size, and that for every parameter we estimate we 'use up' one df, so reducing the effective sample size by the number of parameters we estimate, so reducing also the available information. For background, see e.g. [BF], Notes 3.8, 3.9.)

Example: Normal means $N(\mu, \sigma^2)$, σ unknown.

Here μ is the *parameter of interest*, σ is a *nuisance parameter* – a parameter that appears in the *model*, but not in the *hypothesis* we wish to test.

$$H_0: \quad \mu = \mu_0 \qquad v. \quad H_1: \quad \mu \text{ unrestricted.}$$

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

$$L_0 = \frac{1}{\sigma^n (2\pi)^{n/2}} . \exp\{-\frac{1}{2}\sum_1^n (x_i - \mu_0)^2/\sigma^2\} = \frac{1}{\sigma^n (2\pi)^{n/2}} . \exp\{-\frac{1}{2} n S_0^2/\sigma^2\},$$

in an obvious notation. The MLEs under H_1 are $\hat{\mu} = \bar{X}$, $\hat{\sigma}^2 = S^2$, as before, and under H_0, we obtain as before $\sigma = S_0$. So

$$L_1 = \frac{e^{-\frac{1}{2}n}}{S^n (2\pi)^{\frac{1}{2}n}}; \qquad L_0 = \frac{e^{-\frac{1}{2}n}}{S_0^n (2\pi)^{\frac{1}{2}n}}.$$

So

$$\lambda := L_0/L_1 = S^n/S_0^n.$$

Now

$$n S_0^2 = \sum_1^n (X_i - \mu_0)^2 = \sum [(X_i - \bar{X}) + (\bar{X} - \mu_0)]^2$$

$$= \sum (X_i - \bar{X})^2 + n(\bar{X} - \mu_0)^2 = n S^2 + n(\bar{X} - \mu_0)^2$$

(as $\sum (X_i - \bar{X}) = 0$):

$$\frac{S_0^2}{S^2} = 1 + \frac{(\bar{X} - \mu_0)^2}{S^2}.$$

But $t := (\bar{X} - \mu_0)\sqrt{n-1}/S$ has the Student t-distribution $t(n-1)$ with n df under H_0, so

$$S_0^2/S^2 = 1 + t^2/(n-1).$$

The LR test is: reject if

$\lambda = (S/S_0)^n$ too small;

$S_0^2/S^2 = 1 + t^2/(n-1)$ too big;

t^2 too big: $|t|$ too big, which is the Student t-test:

The LR test here is the Student t-test.

2. Normal variances $N(\mu, \sigma^2)$, μ unknown (a nuisance parameter). Test

$$H_0: \quad \sigma = \sigma_0 \qquad v. \qquad H_1: \quad \sigma > \sigma_0.$$

Under H_0, $\ell = const - n \log \sigma_0 - \frac{1}{2} \sum (X_i - \mu)^2/\sigma_0^2$.
$\partial \ell / \partial \mu = 0$: $\sum_1^n (X_i - \mu) = 0$:

$$\hat{\mu} = \frac{1}{n} \sum_1^n X_i = \bar{X}.$$

So

$$L_0 = \frac{1}{\sigma_0^n (2\pi)^{n/2}} . \exp\{-\frac{1}{2} \sum_1^n (x_i - \mu_0)^2/\sigma_0^2\} = \frac{1}{\sigma_0^n (2\pi)^{n/2}} . \exp\{-\frac{1}{2} n S^2/\sigma_0^2\}.$$

Under H_1, $\ell = const - n \log \sigma - \frac{1}{2} \sum (X_i - \mu)^2/\sigma^2$. As above, the maximising value for μ is $\bar{X}$, and as $\sum_1^n (X_i - \bar{X})^2 = nS^2$,

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

$\partial / \partial \sigma = 0$: $-n/\sigma + nS^2/\sigma^3 = 0$: $\sigma^2 = S^2$.

There are two cases: I. $\sigma_0 < S$. II. $\sigma_0 \geq S$.

In Case I, S belongs to the region $\sigma > \sigma_0$ defining H_1, so the maximum over H_1 is attained at S, giving as before

$$L_1 = \frac{e^{-\frac{1}{2}n}}{S^n (2\pi)^{\frac{1}{2}n}}. \quad \text{So} \quad \lambda = \frac{L_0}{L_1} = \frac{S^n}{S_0^n} \exp\Big\{-\frac{1}{2} n \Big[\frac{S^2}{\sigma_0^2} - 1\Big]\Big\}. \qquad (Case\ I).$$

In Case II, the maximum of L is attained at S (L increases up to S, then decreases), so its restricted maximum in the range $\sigma \geq \sigma_0 \geq S$ is attained at σ_0, the nearest point to the overall maximum S. Then

$$L_1 = \frac{1}{\sigma_0^n (2\pi)^{n/2}} . \exp\{-\frac{1}{2} \sum_1^n (x_i - \mu_0)^2/\sigma_0^2\} = L_) : \qquad \lambda = L_0/L_1 = 1$$

$$(Case\ II).$$

Comparing, λ is a function of $T := S/\sigma_0$:

$$\lambda = 1 \quad \text{if } T \leq 1 \text{ (Case II)}, \qquad T^n \exp\{-\frac{1}{2}n[T^2 - 1]\} \quad \text{if } T \geq 1 \text{ (Case I)}.$$

Now $f(x) := x^n \exp\{-\frac{1}{2}n[x^2 - 1]\}$ takes its maximum on $(0, \infty)$ at $x = 1$, where it takes the value 1 (check by calculus). So (check by graphing λ against T!) the LR test is:

reject if λ too small, i.e. T too big, i.e. S too big – as expected.

Under H_0, nS^2/σ_0^2 is $\chi^2(n-1)$. If c_α is the upper α-point of $\chi^2(n-1)$, reject if $nS^2/\sigma_0^2 \geq c_\alpha$, i.e., *reject if* $S \geq \sigma_0^2 c_\alpha/n$.

Similarly if H_1 is $\sigma < \sigma_1$ and d_α is the lower α-point: *reject if* $S^2 \leq \sigma_0^2 d_\alpha/n$.

Testing Linear Hypotheses

We give a brief overview; for details, see Ch. IV below and e.g. [BF] Ch. 6.

In Regression (Ch. IV below) it is typically the case that one has a sample of size n – the larger the better – and seeks the best explanation of the data obtainable by projection on some suitable p-dimensional subspace. Here p is the number of parameters (in the range 2 to 6, typically), so p is much smaller than n: $p << n$. Having chosen the largest p we are prepared to consider, we might test the hypothesis that p could be reduced – by dropping the last parameter, in a set of nested models. With β the p-vector of parameters, such hypotheses can be formulated as *linear hypotheses* of the form $B\beta = c$, with B a $k \times p$ matrix and c a k-vector of constants. We compare the minimum of the relevant sum-of-squares statistic with and without the constraint $B\beta = c$. The null hypothesis H_0 is that the constraint $B\beta = c$ holds. We reject H_0 if the improvement to the fit when we drop it is too big. It turns out that the relevant test statistic has an F-distribution, and we reject H_0 if this F-statistic is *too big* (Kolodzieczyk's theorem, 1935).

One important instance of all this is Time Series (Ch. V). We have autoregressive models $AR(p)$; we formulate, and test, hypotheses on the size of p needed. Similarly for moving average models $MA(q)$, ARMA models $ARMA(p, q)$, and for their extensions (ARIMA, integrated ARMA; SARIMA, seasonal ARIMA). Similarly for stochastic volatility models, such as autoregressive conditionally heteroscedastic models $ARCH(p)$, generalised ARCH $GARCH(p, q)$, etc.; see e.g. [BF] 9.4.

III. MULTIVARIATE ANALYSIS

1. Preliminaries: Matrix Theory.

Modern Algebra splits into two main parts: Groups, Rings and Fields on the one hand, and Linear Algebra on the other. Linear Algebra deals with *linear transformations* between *vector spaces*. We confine attention here to the *finite-dimensional* case; the infinite-dimensional case needs Functional Analysis and is harder. Broadly, Parametric Statistics can be handled in finitely many dimensions, Non-Parametric Statistics (Ch. VI) needs infinitely many.

Determinants can be traced back to Leibniz (1684, unpublished in his lifetime), Cramer (below) and others; the term first appears in Gauss' thesis *Disquisitiones arithmeticae* in 1801. Although matrices logically precede determinants, they were developed after them. The term is due to J. J. SYLVESTER (1814-1897) in 1850; the theory largely stems from a paper of Arthur CAYLEY (1821-1895) in 1858 (this contains the Cayley-Hamilton Theorem, following work by Hamilton in 1853).

Given a finite-dimensional vector space V, we can always choose a *basis* (a maximal set of linearly independent vectors). All such bases contain the same number of vectors; if this is n, the vector space has *dimension* n.

Given two finite-dimensional vector spaces and a linear transformation α between the two, choice of bases $(e_1, \ldots, e_m)$ and $(f_1, \ldots, f_n)$ determines a *matrix* $A = (a_{ij})$ by

$$e_i\alpha = \sum_{j=1}^{n} a_{ij} f_j \qquad (i = 1, \ldots, m).$$

We write

$$A = \begin{pmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \ldots & a_{mn} \end{pmatrix},$$

or $A = (a_{ij})$ more briefly. The a_{ij} are called the *elements* of the matrix; we write A $(m \times n)$ for A (m rows, n columns).

Matrices may be subjected to various operations:

1. *Matrix addition.* If $A = (a_{ij})$, $B = (b_{ij})$ have the same size, then

$$A \pm B := (a_{ij} \pm b_{ij})$$

(this represents $\alpha \pm \beta$ if α, β are the underlying linear transformations).

2. *Scalar multiplication.* If $A = (a_{ij})$ and c is a scalar (real, unless we specify

complex), then the matrix

$$cA := (ca_{ij})$$

represents $c\alpha$.

3. *Matrix multiplication.* If A is $m \times n$, B is $n \times p$, then $C := AB$ is $m \times p$, where $C = (c_{ij})$ and

$$c_{ij} := \sum_{k=1}^{n} a_{ik} b_{kj}$$

(this represents the product, or composition, $\alpha\beta$ or $x \mapsto x\alpha\beta$).

Note. Matrix multiplication is non-commutative! – $AB \neq BA$ in general, even when both are defined (which can only happen for A, B square of the same size).

Partitioning.

We may *partition* a matrix A in various ways. for instance, A as above partitions as

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},$$

where A_{11} is $r \times s$, A_{12} is $r \times (n-s)$, A_{21} is $(m-r) \times s$, A_{22} is $(m-r) \times (n-s)$, etc. In the same way, A may be partitioned as

(i) a column of its rows; (ii) a row of its columns.

Rank.

The maximal number of linearly independent *rows* of A is always the same as the maximal number of independent *columns*. This number, r, is called the *rank* of A. When $r = \min(m, n)$ is as big as it could be, the matrix A has *full rank*.

Inverses.

When a square matrix A $(n \times n)$ has full rank n, the linear transformation $\alpha : V \to V$ that it represents is *invertible*, and so has an inverse map $\alpha^{-1} : V \to V$ such that $\alpha\alpha^{-1} = \alpha^{-1}\alpha = i$, the identity map, and α^{-1} is also a linear transformation. The matrix representing α^{-1} is called A^{-1}, the *inverse matrix* of A:

$$AA^{-1} = A^{-1}A = I,$$

the *identity matrix* of size n: $I = (\delta_{ij})$ ($\delta_{ij} = 1$ if $i = j$, 0 otherwise – the *Kronecker delta*).

Transpose.

If $A = (a_{ij})$, the *transpose* is A', or $A^T := (a_{ji})$.

Note that, when all the matrices are defined,

$$(AB)^{-1} = B^{-1}A^{-1}$$

(as this gives $(AB)(AB)^{-1} = ABB^{-1}A^{-1} = AA^{-1} = I$, and similarly $(AB)^{-1}(AB) = I$, as required), and

$$(AB)^T = B^T A^T$$

(the (i, j) element is $\sum_k (B^T)_{ik}(A^T)_{kj} = \sum_k b_{ki}a_{jk} = \sum_k a_{jk}b_{ki} = (AB)_{ji}$).
Determinants.

There are $n!$ *permutations* σ of the set

$$\mathbb{N}_n := \{1, 2, \dots, n\}$$

– bijections $\sigma : \mathbb{N}_n \to \mathbb{N}_n$. Each permutation may be decomposed into a product of *transpositions* (interchanges of two elements), and the *parity* of the number of transpositions in any such decomposition is always the same. Call σ *odd* or *even* according as this number is odd or even. Write

$$\text{sgn } \sigma := 1 \quad \text{if } \sigma \text{ is even}, \quad -1 \quad \text{if } \sigma \text{ is odd}$$

for the *sign* or *signum* of σ. For A a square matrix of size n, the function

$$\det A, \text{ or } |A|, := \sum_\sigma (-1)^{\text{ sgn } \sigma} a_{1,\sigma(1)}a_{2,\sigma(2)} \dots a_{n,\sigma(n)},$$

where the summation extends over all $n!$ permutations, is called the *determinant* of A, $\det A$ or $|A|$.
Properties.
1. $|A^T| = |A|$.
Proof. If σ^{-1} is the inverse permutation to σ, σ and σ^{-1} have the same parity, so the sums for their determinants have the same terms, in a different order.
2. If two rows (or columns) of A coincide, $|A| = 0$.
Proof. Interchanging two rows changes the sign of $|A|$ (extra transposition, which changes the parity), but leaves A and so $|A|$ unaltered (as the two rows coincide). So $|A| = -|A|$, giving $|A| = 0$.
3. $|A|$ depends linearly on each row (or column) (det is a *multilinear* function, and this area is called Multilinear Algebra).
4. If A is $n \times n$, $|A| = 0$ iff A has rank $r < n$. For then, some row is a linear combination of others. Expanding by this row gives sum of determinants with two rows identical, giving 0.