smfw2 24 & 26.1.2017

4. Sufficiency and Minimal Sufficiency

Recall (IS II) the idea of sufficiency as data reduction, and minimal sufficiency as data reduction carried as far as possible without loss of information. We now formalise this.

Definition (Fisher, 1922). To estimate a parameter θ from data $\mathbf{x}$, a statistic $T = T(\mathbf{x})$ is *sufficient* for θ if the conditional distribution of $\mathbf{x}$ given $T = T(\mathbf{x})$ does not depend on θ.

Interpretation. Always use what you know. We know T: is this enough? The conditional distribution of $\mathbf{x}$ given T represents the information remaining in the data $\mathbf{x}$ over and above what is in the statistic T. If this does not involve θ, the data *cannot* have anything left in it to tell us about θ beyond what is already in T.

The usual – because the easiest – way to tell when one has a sufficient statistics is the result below. The sufficiency part is due to Fisher in 1922, the necessity part to J. NEYMAN (1894-1981) in 1925.

Theorem (Factorisation Criterion; Fisher-Neyman Theorem. T is sufficient for θ if the likelihood factorises:

$$f(\mathbf{x}; \theta) = g(T(\mathbf{x}); \theta)h(\mathbf{x}),$$

where g involves the data only through T and h does not involve the parameter θ.

Proof. We give the discrete case; the density case is similar.
Necessity. If such a factorisation exists,

$$P_\theta(\mathbf{X} = \mathbf{x}) = g(T(\mathbf{x}), \theta)h(\mathbf{x}),$$

then given t_0,

$$P(T = t_0) = \sum_{\mathbf{x}:T(\mathbf{x})=t_0} P_\theta(\mathbf{X} = \mathbf{x}) = \sum_{\mathbf{x}:T(\mathbf{x})=t_0} g(T(\mathbf{x}), \theta)h(\mathbf{x}) = g(t_0, \theta) \sum_{\mathbf{x}:T(\mathbf{x})=t_0} h(\mathbf{x}).$$

So $P_\theta(\mathbf{X}=\mathbf{x}|T=t_0)=P_\theta(\mathbf{X}=\mathbf{x}\ \&\ T=T(\mathbf{X})=t_0)/P_\theta(T=t_0)$ is 0 unless $T(\mathbf{x})=t_0$, in which case it is

$$P_\theta(\mathbf{X}=\mathbf{x})/P_\theta(T=t_0)=\frac{g(t_0;\theta)h(\mathbf{x})}{g(t_0;\theta)\sum_{T(\mathbf{x})=t_0}h(\mathbf{x})}=\frac{h(\mathbf{x})}{\sum_{T(\mathbf{x})=t_0}h(\mathbf{x})}.$$

This is independent of θ, so T is sufficient.
Sufficiency. If T is sufficient, the conditional distribution of $\mathbf{X}$ given T is independent of θ:

$$P_\theta(\mathbf{X}=\mathbf{x}|T=t_0)=c(\mathbf{x},t_0), \qquad \text{say.} \tag{i}$$

The LHS is $P(\mathbf{X}=\mathbf{x}\ \&\ T(\mathbf{X})=t_0)/P(T=t_0)$. Now the numerator is 0 unless $t_0=T(\mathbf{X})$. Defining $c(\mathbf{x},t_0)$ to be 0 unless $t_0=T(\mathbf{x})$, we have (i) in all cases, and now

$$c(\mathbf{x},t_0)=P_\theta(\mathbf{X}=\mathbf{x})/P(T(\mathbf{X})=t_0),$$

as "& $T(\mathbf{X})=t_0=T(\mathbf{x})$" is redundant. So now

$$P_\theta(\mathbf{X}=\mathbf{x})=P_\theta(T(\mathbf{X})=t_0)c(\mathbf{x},t_0),$$

a factorisation of the required type. //

Cor. If $U=a(T)$ with a injective (one-to-one), T sufficient implies U sufficient.

Proof. $T=a^{-1}(U)$ as a is one-to-one, so

$$f(\mathbf{x};\theta)=g(a^{-1}(U);\theta)h(\mathbf{x})=G(U(\mathbf{x});\theta)h(\mathbf{x}),$$

say, a factorisation of Fisher-Neyman type, so U is sufficient. //

So if, e.g. T is sufficient for the population variance σ^2, $\sqrt{T}$ is sufficient for the standard deviation σ, etc.

Note. From SP, you know Measure Theory, so the above proof may strike you as crude. It is. For the full story, see e.g.
P. R. HALMOS and L. J. SAVAGE, Application of the Radon-Nikodym theorem to the theory of sufficient statistics, *Annals Math. Statistics* **20** (1949),

225-241 [1].
But textbooks often proceed as above (to avoid encumbering new statistical ideas with measure-theoretic technicalities), including the classic book by Rao [R, 2d.3].

Example: Normal families $N(\mu, \sigma^2)$.
(i) The joint likelihood factorises into the product of the marginal likelihoods:

$$f(\mathbf{x}; \mu, \sigma^2) = \frac{1}{(2\pi)^{\frac{1}{2}n}\sigma^n}.\exp\{-\frac{1}{2}\sum_1^n (x_i - \mu)^2/\sigma^2\}. \tag{1}$$

Since $\bar{x} := \frac{1}{n}\sum_1^n x_i$, $\sum(x_i - \bar{x}) = 0$, so

$$\sum(x_i - \mu)^2 = \sum[(x_i - \bar{x}) + (\bar{x} - \mu)]^2 = \sum(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2 = n(S^2 + (\bar{x} - \mu)^2) :$$

the likelihood is

$$L = f(\mathbf{x}; \mu, \sigma^2) = \frac{1}{(2\pi)^{\frac{1}{2}n}\sigma^n}.\exp\{-\frac{1}{2}n(S^2 + (\bar{x} - \mu)^2)/\sigma^2\}. \tag{2}$$

By the Factorisation Criterion, $(\bar{x}, S^2)$ is (jointly) sufficient for (μ, σ^2). So for a *normal* family: only *two* numbers are needed for the two parameters μ, σ^2, namely $\bar{x}, S^2$ (equivalently, $\sum X, \sum X^2$ – note that good programmable pocket calculators have keys for $\sum X, \sum X^2$ for this purpose!)
(ii) Now suppose σ is *known* (so counts as a constant, not a parameter). Then (2) says that $\bar{x}$ *is now sufficient for* μ.
(iii) But if μ is known, (1) says that now $\sum(x_i - \mu)^2$ *is sufficient for* σ^2.

Minimal Sufficiency. Sufficiency enables *data reduction* – reducing from n numbers (n is the sample size – the bigger the better) to a much smaller number (as above). Ideally, we would like to reduce as much as possible, without loss of information. How do we know when we have done this?

Recall that when applying a function, we lose information in general (we do not lose information only when the function is injective – one-to-one, when

[1] P. R. (Paul) Halmos (1916-2006), a versatile mathematician and prolific textbook-writer;
L. J. (Jimmie) Savage (1917-1971), one of the founding fathers, and greatest champions, of Bayesian statistics.

we can go back by applying the inverse function). This leads to the following

Definition. A sufficient statistic T is *minimal (sufficient)* for θ if T is a function of any other sufficient statistic T'.

Minimal sufficient statistics are clearly desirable ('all the information with no redundancy'). The following result gives a way of constructing them.

Theorem (Lehmann & Scheffé, 1950). If T is such that the likelihood ratio $f(\mathbf{x};\theta)/f(\mathbf{y};\theta)$ is independent of θ iff $T(\mathbf{x}) = T(\mathbf{y})$, then T is a minimal sufficient statistic for θ.

We quote this. To find minimal sufficient statistics, we form the likelihood ratio, and seek to eliminate the parameters. This works very well in practice, as examples show (see Problems 2). We quote also

Basu's theorem (Debabrata Basu in 1955).

Any boundedly complete sufficient statistic is independent of any ancillary statistic.

Note. We do not define the term 'boundedly complete' here; see e.g. Rao [R, 5a].

This result is often used to show independence of two statistics: show one boundedly complete sufficient and the other ancillary. Examples (see IS II):
$N(\mu, \sigma^2)$; independence of $\bar{X}$ and S^2;
$U(\theta - \frac{1}{2}, \theta + \frac{1}{2})$: independence of *mid* and *ran*.

5. Location and scale; Tails

In one dimension, the mean μ gives us a natural measure of *location* for a distribution. The variance σ^2, or standard deviation (SD) σ, give us a natural measure of *scale*.

Note. The variance has much better mathematical properties (e.g., it adds over independent, or even uncorrelated, summands). But the SD has the *dimensions* of the random variable, which is better from a physical point of view. As moving between them is mathematically trivial, we do so at will, without further comment.

Example: Temperature. In the UK, before entry to the EU (or Common Market as it was then), temperature was measured in degrees Fahrenheit, F (freezing point of water 32^oF, boiling point 212^oF (these odd choices are only of historical interest – but dividing the freezing-boiling range into 180 parts rather than 100 is better attuned to homo sapiens being warm-blooded, and most people having trouble with decimals and fractions!) The natural choice for freezing is 0; 100 parts for the freezing-boiling range is also natural when using the metric system – whence the Centigrade (= Celsius) scale. Back then, one used F for ordinary life, C for science, and the conversion rules

$$C = \frac{5}{9}(F - 32), \qquad F = \frac{9}{5}C + 32$$

were part of the lives of all schoolchildren (and the mechanism by which many of them grasped the four operations of arithmetic!)

Pivotal quantities.

A *pivotal quantity*, or *pivot*, is one whose distribution is independent of parameters. Pivots are very useful in forming *confidence intervals.*

Defn. A *location family* is one where, for some reference density f, the density has the form $f(x-\mu)$; here μ is a *location parameter.* A *scale family* (usually for $x \geq 0$) is of the form $f(x/\sigma)$; here σ is a *scale parameter.* A *location-scale family* is of the form $f(\frac{x-\mu)}{\sigma})$.
Pivots here are

$$\bar{X} - \mu \quad \text{(location)}; \quad \bar{X}/\sigma \quad \text{(scale)}; \quad \frac{\bar{X} - \mu}{\sigma} \quad \text{(location-scale)}.$$

Examples. The normal family $N(\mu, \sigma^2)$ is a location-scale family.
The *Cauchy location family* is

$$f(x - \mu) = \frac{1}{\pi[1 + (x - \mu)^2]}.$$

In higher dimensions, the location parameter is the mean μ (now a *vector*); the scale parameter is now the *covariance matrix*

$$\Sigma = (\sigma_{ij}), \qquad \sigma_{ij} := cov(X_i, X_j) = E[(X_i - EX_i)(X_j - EX_j)].$$

See III: Multivariate Analysis, below.

6. CAPM.

All of this is highly relevant to Mathematical Finance. Finance was an art rather than a science before the 1952 PhD thesis of Harry MARKOWITZ (1927-; Nobel Prize 1990). Markowitz gave us two insights that have become so much part of the ambient culture that it is difficult to realise that they have not always been there. These are:
(i). *Think of risk and return together, not separately.* Now return corresponds to mean (= mean rate of return), risk corresponds to variance – hence *mean-variance* analysis (hence also the *efficient frontier*, etc. – one seeks to maximise return for a given level of risk, or minimise risk for a given return rate).
(ii). *Diversify* (don't 'put all your eggs in one basket'). Hold a *balanced portfolio* – a range of risky assets, with lots of *negative correlation* – so that when things change, one's losses on some assets will tend to be offset by gains on others.
Markowitz's work led on to the *Capital Asset Pricing Model* (CAPM – "capemm") of the 1960s (Jack TREYNOR in 1961/62, William SHARPE (1934-; Nobel Prize 1990), John LINTNER (1965), Jan MOSSIN (1966)), the first phase of the development of Mathematical Finance. The second phase was triggered by the *Black-Scholes formula* of 1973 and its follow-up by Merton (Fischer BLACK (1938-95); Myron SCHOLES (1941-; Nobel Prize 1997); Robert C. MERTON (1944-; Nobel Prize 1997)).

As a result of Markowitz's work, the vector-matrix parameter (μ, Σ) is accepted as an essential part of any model in mathematical finance. As a result of CAPM, *regression* methods (Ch. IV) are an essential part of any portfolio management programme. The x-axis is used to represent the return for the *market* (or a *portfolio*) as a whole, the y-axis for the return of the asset.

II: HYPOTHESIS TESTING

1. Formulation

The essence of the scientific method is to formulate theories, and test them experimentally. Thus a typical scientific experiment will *test* some theoretical prediction, or *hypothesis*.

We can never *prove* that a scientific theory, or hypothesis, is *true*. To take an extreme case, look at Newton's Laws of Motion (Sir Isaac NEWTON (1642-1727); *Principia*, 1687). This was the mathematics that made possible the Scientific Revolution, and Newton's Laws were regarded as unchallengeable for more than two centuries. But in the 20th century, Quantum Mechanics showed that Newton's Laws are approximate only – useful in the macroscopic case, but inadequate at the atomic or subatomic level.

With this in mind, we should treat established theory with respect, and not replace it lightly (or textbooks would become too ephemeral!), but not regard it as sacrosanct: scientific theory is provisional, and evolving. This is part of the great strength of the scientific method.

It is customary, and convenient, to represent the existing theory by a *null hypothesis*, H_0, and to test it against a candidate new theory, an *alternative hypothesis*, H_1.

A hypothesis is *simple* if it completely specifies the parameter(s); e.g.,

$$H_0 : \qquad \theta = \theta_0,$$

composite otherwise, e.g.

$$H_0 : \qquad \theta > \theta_0.$$

As above, there is an *asymmetry* between H_0 and H_1: H_0 is the '*default option*'. We will discard H_0 in favour of H_1 only if the data gives us convincing evidence to do so.

Legal analogy.

Hypothesis test $\leftrightarrow$ Criminal trial

Null hypothesis $H_0 \leftrightarrow$ accused

H_0 accepted till shown untenable $\leftrightarrow$ accused innocent until proved guilty

Accept (= do not reject) $H_0 \leftrightarrow$ not guilty verdict

Reject H_0 (for H_1) $\leftrightarrow$ guilty verdict

Data $\leftrightarrow$ evidence

Statistician $\leftrightarrow$ jury

Significance level $\alpha \leftrightarrow$ probability of convicting an innocent person.

Significance level.

The above introduces this important term. Statistical data (like legal evidence) is random (if we re-sampled, we would get different data!) So we can never conclude *with certainty* anything from data – including that H_0 is false. But we cannot go from this to saying that we can never reject H_0 – or scientific progress would halt, being frozen at the current level. We strike a sensible balance by choosing some small probability, α, of rejecting a valid null hypothesis, and working with that. We call α the *significance level*. Common choices are $\alpha = 0.05$, or 5%, for ordinary work, and $\alpha = 0.01$, or 1%, for accurate work. But note that the choice of α is down to *you*, the statistician, so is subjective. We like to think of Science as an objective activity! So the whole framework of Hypothesis Testing is open to question – indeed, it is explicitly rejected by *Bayesian* statisticians (see Ch. VII below). (But then, the concept of a criminal trial is explicitly rejected in some forms of political thinking, such as Anarchism.)

There are two types of error in Hypothesis Testing, called *Type I error – false rejection* (rejecting H_0 wrongly, probability α – cf. convicting an innocent person), and *Type II error – false acceptance* (accepting H_0 when it is false, probability β, say – cf. acquitting a guilty person). The usual procedure is to fix α, and then try to minimise β for this α.

Usually, we decide on a suitable *test statistic*, $T = T(\mathbf{X})$, and *reject* H_0 if the data $\mathbf{X}$ falls in the *critical region* (or *rejection region*), R say, where T falls in some set S. Then abbreviating P_{θ_i} to P_i:

$$\alpha = P_0(\mathbf{X} \in R), \qquad \beta = P_1(\mathbf{X} \notin R).$$

We often look at

$$1 - \beta = P_1(\mathbf{X} \in R),$$

the probability that the test correctly picks up that H_0 is false. We can think of this as the *sensitivity* of the test; the technical term used is the *power* of the test. This depends on θ (grossly wrong hypotheses are easier to reject than marginally wrong ones!);

$$\theta \mapsto 1 - \beta(\theta)$$

is called the *power function* of the test.

Usually, we fix the significance level α and the sample size n, and then seek to choose the rejection region R so as to maximise the power $1 - \beta$ [minimise the prob. β of Type II error, false acceptance].

The *Likelihood Principle (LP)* says that all that matters is the likelihood L, which is

$L_0 := L(\mathbf{X}; \theta_0)$ if H_0 is true;

$L_1 := L(\mathbf{X}; \theta_1)$ if H_1 is true.

The idea of maximum likelihood estimation is that the data supports θ if $L(\mathbf{X}; \theta)$ is large. This suggests that a good test statistic for H_0 $v.H_1$ would be the *likelihood ratio (LR)*

$$\lambda := L_0/L_1 = L(\mathbf{X}; \theta_0)/L(\mathbf{X}; \theta_1),$$

rejecting H_0 for H_1 if λ is *too small* – that is, using the critical region

$$R := \{\mathbf{X} : \lambda(\mathbf{X}) \leq c\},$$

where c is chosen so that

$$\alpha = P_0(\mathbf{X} \in R).$$

In the density case, such a region does exist. In the discrete case, it may not: the probability may 'jump over' the level α if one more point is included. One can allow for this by randomisation (including the 'extra point' with some probability so as to get α right) but we ignore this, and deal with the density case – the important case in practice.

2. The Neyman-Pearson Lemma

The simple suggestion above is in fact best possible. This is due to J. NEYMAN (1894-1981) and E. S. PEARSON (1895-1980) in 1933.

Theorem (Neyman-Pearson Lemma). To test a simple null hypothesis $H_0 : \theta = \theta_0$ against a simple alternative hypothesis $H_1 : \theta = \theta_1$ at

significance level α, a critical region of the form

$$R := \{\mathbf{X} : \lambda \leq c\} = \{\mathbf{X} : L(\mathbf{X}; \theta_0)/L(\mathbf{X}; \theta_1) \leq c\}, \qquad \alpha = P_0(\lambda \leq c)$$

is best possible (most powerful): the $\beta = \beta(R)$ for this R is as small as possible for given α and n.

Proof. If S is any other critical region with the same significance level (or 'size') α, we need to show $\beta(S) \geq \beta(R)$, i.e.

$$\int_{S^c} f(\mathbf{x}; \theta_1) d\mathbf{x} \geq \int_{R^c} f(\mathbf{x}; \theta_1) d\mathbf{x} : \qquad \int_{S^c} f(\theta_1) \geq \int_{R^c} f(\theta_1),$$

or as densities integrate to 1,

$$\int_S f(\theta_1) \leq \int_R f(\theta_1). \qquad (*)$$

But

$$\int_R f(\theta_1) - \int_S f(\theta_1) = \int_{R \cap S} f(\theta_1) + \int_{R \setminus S} f(\theta_1) - \int_{R \cap S} f(\theta_1) - \int_{S \setminus R} f(\theta_1)$$

$$= \int_{R \setminus S} f(\theta_1) - \int_{S \setminus R} f(\theta_1).$$

Now

$$\lambda = L_0/L_1 \leq c \qquad (\mathbf{X} \in R), \qquad > c \qquad (\mathbf{X} \notin R),$$

or reverting from "L" to "f" notation,

$$f(\theta_1) \geq c^{-1} f(\theta_0) \quad \text{in} \quad R, \quad < c^{-1} f(\theta_0) \quad \text{in} \quad R^c.$$

As $R \setminus S \subset R$, this gives

$$\int_{R \setminus S} f(\theta_1) \geq c^{-1} \int_{R \setminus S} f(\theta_0).$$

Similarly,

$$\int_{S \setminus R} f(\theta_1) \leq c^{-1} \int_{S \setminus R} f(\theta_0), \qquad - \int_{S \setminus R} f(\theta_1) \geq -c^{-1} \int_{S \setminus R} f(\theta_0).$$

Add:

$$\int_R f(\theta_1) - \int_S f(\theta_1) = \int_{R\setminus S} f(\theta_1) - \int_{S\setminus R} f(\theta_1) \geq c^{-1}\Big[\int_{R\setminus S} f(\theta_0) - \int_{S\setminus R} f(\theta_0)\Big]. \tag{a}$$

But both R and S have size (θ_0-probability) α:

$$\alpha = \int_R f(\theta_0) = \int_{R\cap S} f(\theta_0) + \int_{R\setminus S} f(\theta_0),$$

$$\alpha = \int_S f(\theta_0) = \int_{R\cap S} f(\theta_0) + \int_{S\setminus R} f(\theta_0).$$

Subtract:

$$\int_{R\setminus S} f(\theta_0) = \int_{S\setminus R} f(\theta_0).$$

This says that the RHS of (a) is 0. Now (a) gives $(*)$. //

Note. The Neyman-Pearson Lemma (NP) is fine as far as it goes – simple v. simple. But most realistic hypothesis testing situations are more complicated. Fortunately, NP extends to some important cases of simple v. composite; see below. We turn to composite v. composite later, using likelihood ratio tests (LR).

Sufficiency and the Fishr-Neyman theorem. If T is sufficient for θ,

$$L(\mathbf{X};\theta) = g(T(\mathbf{X};\theta)h(\mathbf{X}),$$

by Fisher-Neyman. Dividing,

$$\lambda := L(\theta_0)/L(\theta_1) = g(T(\mathbf{X};\theta_0)/g(T(\mathbf{X};\theta_1)$$

is a function of T only. So if we have a sufficient statistic T, we lose nothing by restricting to test statistics which are functions of T.

Example.

1. *Normal means*, $N(\mu,\sigma^2)$, σ known.

To test $H_0: \mu = \mu_0$ v. $H_1: \mu = \mu_1$, where $\mu_1 < \mu_0$. It turns out that the NP critical region is of the form 'reject if $\bar{X}$ is too small'. (This is intuitive, as $\mu_1 < \mu_0$.) How small is too small? Because the significance level α involves probabilities under H_0, the critical region is the *same* for *all* μ_1,

provided only that $\mu_1 < \mu_0$ (if instead $\mu_1 > \mu_0$, the critical region is 'reject if $\bar{X}$ is too big'). That is, the NP test is most powerful, *uniformly* in μ_1 for all $\mu_1 < \mu_0$. We call the critical region *uniformly most powerful (UMP)* for the *simple* null hypothesis H_0: $\mu = \mu_0$ v. the *composite* alternative hypothesis H_1 : $\mu < \mu_0$. Similarly for H_1 : $\mu > \mu_0$.

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 MLEs (I.3), $\hat{\theta} \to \theta_0$ a.s. as $n \to \infty$. So also $\theta^* \to \theta_0$. So

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

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 $H_0 : \theta_r = \theta_{r,0}$ v. $H_1 :$ θ_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

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

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 $T := S/\sigma_0$:

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

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 $nS^2/\sigma_0^2 \ge c_\alpha$, i.e., *reject if* $S \ge \sigma_0^2 c_\alpha/n$.

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

4. Testing Linear Hypotheses

We follow [BF] Ch. 6. In the regression context, of estimating parameters β in a model $y = A\beta + \epsilon$ (A the design matrix, $n \times p$, known, β the p-vector of parameters, ϵ the n-vector of errors, iid $N(0,\sigma^2)$), the MLE for β is $\hat{b} = (A^TA)^{-1}Ay = C^{-1}Ay$. The total sum of squares is $SS = SSR + SSE$, the sum of the sums of squares for regression and for error. We choose β so as to minimise SS, equivalently, SSR (as SSE is a statistic – can be calculated from the data – and does not involve the unknown parameters). If we have to test a *linear hypothesis*

$$B\beta = c$$

(B is a $k \times p$ matrix, with $k \le p$, of full rank, and c is a k-vector of constants),

$$\text{minimise} \quad SSR = (\hat{\beta}-\beta)^T C(\hat{b}-\beta) \qquad \text{under} \qquad B\beta = c.$$

This is a *constrained minimisation problem*, and can be solved (as usual) by *Lagrange multipliers*. It turns out that the minimising value is

$$\beta^\dagger = \hat{\beta} - C^{-1}B^T(BC^{-1}B^T)^{-1}(B\hat{b}-c),$$

and with

$$SSH := (\hat{\beta}-\beta^\dagger)^T C(\hat{\beta}-\beta^\dagger)$$

the *sum of squares for the hypothesis*, $SSE/\sigma^2 \sim \chi^2(n-p)$ and $SSH/\sigma^2 \sim \chi^2(k)$ are independent. We test for H using the F-statistic

$$F := \frac{SSH/k}{SSE/(n-p)} \sim F(k, n-p),$$

rejecting H if F is too big (Kolodzieczyk's Theorem, 1935) [Polish l: 'Kowod*jay*chick'].