WILKS'S THEOREM; A LIKELIHOOD RATIO TEST FOR NESTED COMPOSITE HYPOTHESES

1. Assumptions

Suppose we have a family of probability density or mass functions $f(\theta, x) > 0$ depending on a d-dimensional parameter θ which ranges over a parameter space H_1. Then θ will be written as $(\theta_1, \ldots, \theta_d)$. (In general, however, H_1 will not necessarily be a Euclidean space $\mathbb{R}^d$ or an open subset of one. It may be a curved surface or manifold in a higher-dimensional space, on which $\theta_1, ..., \theta_d$ are local coordinates.) As usual, P_θ will denote the probability distribution of x given $\theta \in H_1$, and E_θ the expectation under that distribution, possibly for $X_1, ..., X_n$ i.i.d. P_θ.

Let $L(\theta, x) := \log f(\theta, x)$ be the log likelihood function. It will be assumed that there are partial derivatives $\partial \log f(\theta, x)/\partial\theta_i$ at each $\theta \in H_1$, continuous with respect to θ. Suppose also that all elements of the matrix

$$I_{ij}(\theta) := E_\theta\left(\frac{\partial L(\theta, x)}{\partial\theta_i}\frac{\partial L(\theta, x)}{\partial\theta_j}\right),\ i, j = 1, ..., d,$$

are well-defined and finite. Then $I_{ij}(\theta)$ is a symmetric $d \times d$ matrix, called the *Fisher information matrix* at θ. It's easily seen that it's nonnegative definite, since for any $t = (t_1, \ldots, t_d)$, we have

$$\sum_{i,j=1}^{d} I_{ij}(\theta)t_i t_j = E_\theta\left(\left(\sum_{i=1}^{d} t_i \frac{\partial L(\theta, x)}{\partial\theta_i}\right)^2\right) \geq 0.$$

It will be assumed that the Fisher information matrix is strictly positive definite for all θ, in other words, in the last inequality, "≥ 0" is replaced by "> 0" if at least one $t_i \neq 0$. This will assure that the model H_1 is truly d-dimensional. For example, if for some $i = 1, \ldots, d$, $\partial L(\theta, x)/\partial\theta_i = 0$ for all θ and x, then $I_{ij}(\theta) = 0$ for all j, I_{ij} would not be positive definite, and the true dimension of the model H_1 would be $d - 1$ or less.

Although not necessary for the given definition of $I_{ij}(\theta)$, it's often convenient to assume that second partial derivatives $H_{ij}(x, \theta) := \partial^2 \log f(\theta, x)/\partial\theta_i\partial\theta_j$ exist for $i, j = 1, ..., d$,

Date: 18.650, Oct. 14, 2015.

The Fisher information matrix is important in mathematical statistics, for example in the course formerly numbered 18.466, but it will not be mentioned further in this course. It was only brought in to give an honest definition of the dimension of a model.

Let H_0 be an m-dimensional subset of H_1 for some $m < d$. It will be assumed that H_0 is "smooth" in the sense that at any point of H_0, we can select m of the parameters, say for example $\theta_1, ..., \theta_m$, for which the other $d - m$ parameters are twice differentiable functions of $\theta_1, ..., \theta_m$. Or, H_0 may be given by way of an m-dimensional parameter $\phi = (\phi_1, ..., \phi_m)$ and a mapping $\phi \mapsto \theta(\phi)$ of H_0 into H_1, such that partial derivatives $\partial/\partial\phi_j$ and $\partial^2/\partial\phi_j\partial\phi_k$ of $f(\theta(\phi), x)$ exist and have the same properties with respect to ϕ as were assumed above with respect to θ.

Example 1. Consider the family of all normal distributions $N(\mu, \sigma^2)$, with $d = 2$. Here μ can be any real number, and σ or σ^2 any number > 0. The subfamily with $\mu = 0$ has dimension $m = 1$.

Example 2. In a trinomial distribution, we have n independent trials with three possible outcomes, having respective probabilities p_1, p_2, and p_3 of occurring on each trial. Then $p_j \geq 0$ and $p_1 + p_2 + p_3 = 1$. Let's assume that $p_j > 0$ for each j. This parameter space H_1 has dimension $d = 2$. We can take for example p_1 and p_2 as coordinates, where $p_3 \equiv 1 - p_1 - p_2$. One lower-dimensional submodel H_0 is the family of Hardy–Weinberg equilibrium distributions in which for some p with $0 < p < 1$ and $q \equiv 1 - p$, $p_1 = p^2$, $p_2 = 2pq$, and $p_3 = q^2$. Thus $\phi = \phi_1 = p$ is the parameter for H_0, which has dimension $m = 1$, and $\theta(p) = \{\theta_j(p)\}_{j=1}^3 = \{p_j(p)\}_{j=1}^3 = (p^2, 2pq, q^2)$. So the mapping $\theta(\cdot)$ is nonlinear (quadratic), but derivatives with respect to p of all orders exist and are continuous.

2. Defining the test statistic

Assume that observations $X_1, \ldots, X_n$ are i.i.d. with likelihood function $f(\theta, x)$ for some $\theta \in H_1$. We want to test the hypothesis that $\theta \in H_0$. Let $L(\theta) = \Pi_{j=1}^n f(\theta, X_j)$ be the likelihood function. Let ML_d be the maximum of the likelihood for θ in H_1, in other words $ML_d = L(\widehat{\theta}_d)$ where $\widehat{\theta}_d$ is the MLE of θ in H_1, provided it exists. Let ML_m be, likewise, the maximum of the likelihood for θ in H_0. Then $ML_m \leq ML_d$ because $H_0 \subset H_1$. Let Λ be the *likelihood ratio*, $\Lambda = ML_m/ML_d$, so that $0 < \Lambda \leq 1$. We would want to reject H_0 if Λ is small, or sufficiently less than 1, depending on n, but not reject it if Λ is close to 1. What is a quantitative criterion, i.e. a test of H_0?

S. S. Wilks in 1938 proposed the following test: let $W = -2 \log \Lambda$, so that $0 \leq W < \infty$. Wilks found that if the hypothesis H_0 is true, then the distribution of W converges as $n \to \infty$ to a χ^2 distribution with $d - m$ degrees of freedom, not depending on the true $\theta = \theta_0 \in H_0$. Thus, H_0 would be rejected if W is too large in terms of the tabulated χ^2_{d-m} distribution.

In a multinomial, e.g. a trinomial distribution for which the jth outcome has probability p_j, if the jth outcome is observed X_j times in n trials, then since X_j has a binomial (n, p_j) distribution, the maximum likelihood estimate of p_j is X_j/n for each j. For the Hardy–Weinberg submodel of the trinomial, the maximum likelihood estimate of p is also easy to find: the likelihood function is

$$(p^2)^{X_1}(2pq)^{X_2}(q^2)^{X_3}$$

times factors not depending on p, or $p^{2X_1+X_2}q^{X_2+2X_3}$ times such factors. This has the form of a binomial likelihood function for success probability p and $2n$ trials. So it would be possible to find the likelihood ratio and Wilks statistic Λ and W respectively.

Recall Example 1 where H_1 is the set of all normal distributions with $d = 2$, and H_0 is that $\mu = 0$, having dimension $m = 1$. In this case another test of H_0 would be whether 0 is outside the $1 - \alpha$ confidence interval for μ. Question: is this test, or the Wilks test, preferable in this case? The Wilks test is only applicable when n is large and with an approximation, whereas the test based on a confidence interval uses the t distribution which is exact for any $n \geq 2$.

The likelihood ratio test of a multinomial hypothesis for k categories (Rice, Section 9.6) is a special case of Wilks's test with $d = k - 1$.

The fact that Wilks's statistic has the given asymptotic distribution if H_0 is true is called Wilks's theorem. It holds under some hypotheses, not all of which have been stated, but which are given in references as indicated in the Notes.

3. Notes

Wilks first published his theorem in a paper, Wilks (1938), then gave an exposition of it in his book, Wilks (1962, §13.8). Chernoff (1954) gave another proof. Van der Vaart (1998, Chapter 16) gives a more recent exposition. The Notes by van der Vaart (1998, p. 240) suggest that Wilks's original proof was not rigorous. The proof in the 1962 book seems rather long. A proof is given in Dudley (2003), Section 3.9.

REFERENCES

Chernoff, Herman (1954). On the distribution of the likelihood ratio statistic. *Ann. Math. Statist.* **25**, 573-578.

Dudley, R. M. (2003). *Mathematical Statistics*, 18.466 lecture notes, Spring 2003. On MIT OCW (OpenCourseWare) website, 2004.

van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.* **9**, 60-62.

Wilks, S. S. (1962). *Mathematical Statistics.* Wiley, New York; 2d printing, corrected, 1963.