Hypothesis Testing in the Maximum Likelihood Context

MLE gives us an "asymptotically" optimal estimation principle. Might anticipate that constructing tests based on MLEs will lead to tests which perform well, at least asymptotically.

Problem: $H_0: \theta \in \Omega_0$

$H_1: \theta \in \Omega_1$

(If $\Omega_0 \subset \Omega_1$, hypotheses are "nested".)

e.g.: $y = \beta_1 + \beta_2 X + \varepsilon$

$$H_0: \beta_2 = 0 \quad ; \quad H_1: \beta_2 \neq 0$$

<u>Definition</u>: A "Test" is a decision rule that leads us to reject/not reject a stated hypothesis.

Specifically: $T_n = T_n(y_1, \ldots, y_n)$ is a "test statistic" - a function of the random $\{y_i\}$ data. Our test (rule) is of the form:

Reject H_0 if $T_n \in C$,

where C is the "critical region" or "rejection region" of the sample space. Usually T_n is an estimator of θ, or a function of such an estimator.

e.g. $H_0: \beta_2 = 0$; $H_1: \beta_2 \neq 0$

$$T_n = t_{n-K} = \frac{\tilde{\beta}_2 - 0}{s.e.(\tilde{\beta}_2)}$$

We call Pr.(Type I error) the "size" of the test ("significance level").

$$\alpha = Pr[T_n \in C \mid \theta \in \Omega_0]$$

In classical hypothesis testing we set (upper limit for) α. Then try to construct a test which has small Pr.(Type II error). That is, we try to construct test so that:

$Pr.[T_n \in \bar{C} \mid \theta \in \Omega_1]$ is small; or:

$\text{Power} = Pr.[T_n \in C \mid \theta \in \Omega_1]$ is large.

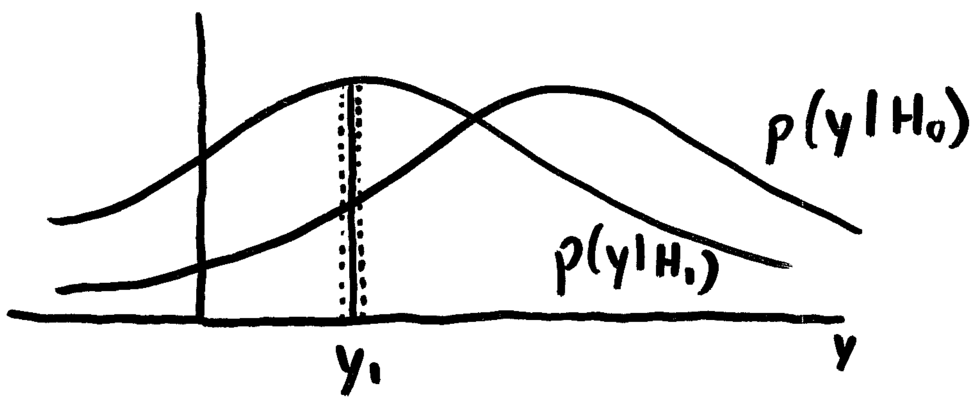
[$\text{Power} = 1 - Pr.(\text{Type II error})$].

Recall the following test properties:

* Unbiased * Consistent * Most Powerful * UMP * LMP

The big question is: For a given choice of α, how can we construct a UMP test for a particular problem? This is where the Likelihood funct. and MLE can help. Recall - Likelihood funct. provides a full description of random sample data.

Suppose we have a scalar θ, and $n=1$:



$$Pr.[y_1 - \varepsilon < y < y_1 + \varepsilon \mid H_1] > Pr.[y_1 - \varepsilon < y < y_1 + \varepsilon \mid H_0]$$

In the limit, as $\varepsilon \to 0$: $p(y_1 \mid H_1) > p(y_1 \mid H_0)$, which suggests that we should reject H_0 as being "less likely" than H_1. When $n > 1$ we generalize this using $p(y_1, \dots, y_n) \to$ Likelihood funct.

We're going to consider three general testing principles that use this intuition. They differ in terms of the point at which we evaluate the L.F. – at H_0, at H_1, or at both.

* Likelihood Ratio Test (Both)
  (Neyman and Pearson ; 1928–1934)

* Wald Test (H_1)
  (Wald ; 1943)

* Lagrange Multiplier Test (H_0)
  (Silvey ; 1959)
  (Variation of Rao's "Score Test" ; 1948)

Lemma (Neyman - Pearson):

Suppose we have $\{y_1, \ldots, y_n\}$ drawn from the population, $p(y_1, \ldots, y_n ; \theta)$.

Let $\lambda = \lambda(y_1, \ldots, y_n ; \theta_0, \theta_1)$

$$= \frac{p(y_1, \ldots, y_n ; \theta_0)}{p(y_1, \ldots, y_n ; \theta_1)} = \frac{L(\theta_0)}{L(\theta_1)} \quad (= T_n)$$

and let $C = \{\lambda : \lambda \leq K\}$, where 'K' is such that
$Pr[\lambda \in C \mid \theta = \theta_0] = \alpha$, and α = chosen size for test.

Then C is the best critical region for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, in the sense that the associated test is Most Powerful.

Note:

(i) No guarantee that such a test exists (it will if $p(y_1, \ldots, y_n)$ is a density)

(ii) "Simple" null and alternative - no parameters to estimate (very special case).

(iii) Idea - put points into critical region until it reaches size of α.

(iv) To max. power, points that are more likely under H_1 than under H_0 should be put into C.

(v) If a point $(y_1, \ldots, y_n)$ is more likely under H_1 than under H_0 then this increases denominator of λ and makes λ smaller. Hence, reject if $\lambda < k$.

Practical Issue – To apply test we need to know K, such that $Pr.[\lambda < K \mid \theta = \theta_0] = \alpha$. This requires knowledge of the distribution of the test statistic, λ. This may be tricky.

<u>Example</u> : Binomial Dist^n

$$L(p) = {}^nC_x\, p^x (1-p)^{n-x} \quad \text{and} \quad \tilde{p} = x/n .$$

$$H_0 : p = 1/3 \quad \text{vs.} \quad H_1 : p = 2/3$$

$$\lambda = \frac{{}^nC_x \left(\frac{1}{3}\right)^x \left(2/3\right)^{n-x}}{{}^nC_x \left(2/3\right)^x \left(1/3\right)^{n-x}} = \left(\frac{1}{2}\right)^x 2^{n-x} = 2^{n-2x}$$

So, calculate λ and reject H_0 if $\lambda < K$, where K is chosen so that $Pr[\lambda < K \mid p = 1/3] = \alpha$. How?

Now, usually in practice we don't have simple H_0 and H_1. Usually θ is a vector. For example -

$y_i \sim N(\mu, \sigma^2)$. $H_0: \mu = 0$ vs. $H_1: \mu > 0$.

Generalized Likelihood Ratio Test -

Let $\tilde{\lambda} = L(\hat{\tilde{\theta}}_0) / L(\hat{\tilde{\theta}}_1)$ and reject H_0 if $\hat{\tilde{\lambda}} < K$, where K s.t. $Pr[\hat{\tilde{\lambda}} < k \mid H_0 \text{ true}] = \alpha$.

We are now evaluating ~~the~~ L.F. twice - once under H_0 and once under H_1. Often this LRT has good power properties, but need not be (U)MP. Another point to note is that while dist^n of $\tilde{\lambda}$ may be unclear, we may be able to resolve issue after suitable transformation. For example, when $y = X\beta + \varepsilon$; $\varepsilon \sim N[0, \sigma^2 I]$; $H_0: R\beta = q$;

$H_1: R\beta \neq q$. If we take the following monotonic (decreasing) function of $\tilde{\lambda} \to \tilde{\lambda}^* = (\tilde{\lambda}^{-2/n} - 1)(n-k)/J$, then $\tilde{\lambda}^*$ is F-distributed if H_0 is true. LRT is equivalent to the F-test here, and so is a UMP test.

However, in other problems, may be impossible to get dist^n for $\tilde{\lambda}$ in finite samples, even after transformation. There is an extremely important class of testing problems where we can apply LRT with <u>asymptotic</u> validity - "<u>Nested</u>" hypotheses. Suppose H_0 is nested within H_1. That is, H_0 involves imposing restrictions on H_1. Let $m = \#$ independent restrictions to get H_0 from H_1. Then - if the null and alternative hypotheses are nested, and the usual regularity conditions are satisfied:

$$LRT = -2\log\tilde{\lambda} \xrightarrow{d} \chi^2_{(m)} \quad \text{if } H_0 \text{ true.}$$

Note:

* This last result is only for "nested" case.
* Only an asymptotic result.
* LRT is a "consistent" test.
* Power may be low in finite samples.
* Computational point - need to evaluate $L(\theta)$ twice - $L(\tilde{\theta}_0)$ and $L(\tilde{\Theta}_1)$. This is usually simple in practice.

The Wald Test:

This test is also designed for testing "nested" hypotheses, but requires only $\hat{\tilde{\theta}}_1$ to be calculated. Useful if calculation of $\tilde{\theta}_0$ is computationally difficult.

Consider linear restrictions: $H_0: R\theta = r$ vs. $H_1: R\theta \neq r$.

If H_0 is true, expect $R\hat{\tilde{\theta}}_1 \approx r$, at least as $n \to \infty$.

We know that $\sqrt{n}(\hat{\tilde{\theta}}_1 - \theta) \xrightarrow{d} N[0, IA(\theta)^{-1}]$. So:

$\sqrt{n}R(\tilde{\theta}_1 - \theta) \xrightarrow{d} N[0, RIA^{-1}R']$ and if H_0 is true $(R\theta = r)$ -

$\sqrt{n}(R\tilde{\theta}_1 - r) \xrightarrow{d} N[0, RIA^{-1}R']$.

So: $\sqrt{n}(R\tilde{\theta}_1 - r)'[RIA^{-1}R']^{-1}\sqrt{n}(R\hat{\tilde{\theta}}_1 - r)$

$$= n(R\hat{\tilde{\theta}}_1 - r)'[RIA^{-1}R']^{-1}(R\hat{\tilde{\theta}}_1 - r) \xrightarrow{d} \chi^2_{(J)} \text{ if } H_0 \text{ true.}$$

Problem: $IA = IA(\theta)$ and is unobservable. Just replace it with any consistent estimator, and asymptotic distⁿ is unchanged. Estimate $IA(\theta)$ by $\frac{1}{n} I^*(\tilde{\theta}_1)$, where

$$\text{plim}\left[\tfrac{1}{n} I^*(\tilde{\theta}_1)\right] = IA(\theta).$$ So, replace $\frac{1}{n} IA^{-1}$ by $I^*(\tilde{\theta}_1)^{-1}$. Wald test statistic is:

$$W = (R\tilde{\theta}_1 - r)'[R I^*(\tilde{\theta}_1)^{-1} R']^{-1}(R\tilde{\theta}_1 - r) \xrightarrow{d} \chi^2_J \quad \text{if } H_0 \text{ true.}$$

If H_0 is true, expect that $R\tilde{\theta}_1 \simeq r$, and so W will be small.

* Although Wald test statistic has same asymptotic distⁿ as LRT, usually differs in finite samples.
* Also, distⁿˢ of Wald and LRT may differ if H_0 false ↳ different powers.

The Lagrange Multiplier Test:

Again, nested hypotheses - base test only on the <u>restricted</u> MLE for parameters. (Especially useful if imposing the restrictions makes MLE easier).

Starting point - Ignoring any restrictions:

$$[\partial \log L_1(\theta) \,/\, \partial\theta] \big|_{\tilde{\theta}_1} = 0$$

i.e. $D \log L_1(\tilde{\theta}_1) = 0$, for simplicity. If H_0 is true, expect also that $D \log L_1(\tilde{\theta}_0) \approx 0$. This suggests a test statistic of the form:

$$LM = D \log L_1(\tilde{\theta}_0)' \, I^*(\tilde{\theta}_0)^{-1} \, D \log L_1(\tilde{\theta}_0) \xrightarrow{d} \chi^2_J \text{, if } H_0 \text{ true.}$$