STA 3000F Testing composite hypotheses **2013**

Suppose H_0 and H_1 are *composite* null hypotheses, of the form $H_0 : \theta \in \Theta_0$, $H_1 : \theta \notin \Theta_0$. Any test of level α should have level α for all values of θ in H_0: i.e. we want to maximize $\Pr_{H_1}\{Y \in \mathcal{R}\}$ subject to

$$\Pr_\theta\{Y \in \mathcal{R}\} \leq \alpha, \qquad \forall \theta \in \Theta_0.$$

The second of these requirements essentially means that we have to know the distribution of Y under H_0, but since this is not specified by H_0, $\mathcal{R}$ will need to be determined by some function of Y, say $t(Y)$, whose distribution *is* known under H_0. Assume for now that H_0 is composite of the form $\theta = (\psi, \lambda)$, with H_0 only specifying ψ,i.e. $H_0 : \psi = \psi_0$.

1. Reduction by conditioning

If we can write

$$f_Y(y;\theta) = f_1(s(y)|t(y);\psi) f_2(t(y);\psi,\lambda),$$

then we've found a function of Y whose distribution depends only on ψ, i.e. H_0 completely specifies the conditional distribution of S (given T). Consequently any test that depends on Y only through S, and considering T fixed, can be ensured to have level α for all values of λ. (In most cases, (S, T) is an original reduction from Y to a minimal sufficient statistic for the full parameter vector (ψ, λ). In that case the equation above would really be

$$f_Y(y;\theta) = f_1(s(y)|t(y);\psi) f_2(t(y);\psi,\lambda) h(y|s(y), t(y)),$$

where by sufficiency the third term does not depend on any parameters.) If our test is constructed using the conditional distribution f_1 above, then we're back to a simple null hypothesis, and we know how to find first a most powerful test under a particular alternative, and then how to check if it is uniformly most powerful for a range of ψ values. To this point there are no obvious optimality claims *for the original problem*. However, it can be proved, under fairly weak requirements on the conditioning statistic T, that all critical regions satisfying

$$\int_{\mathcal{R}} f(y;\psi_0,\lambda)dy = \alpha,$$

must be obtained by conditioning on T. Such regions are called similar, and the best among them called most powerful similar. In several exponential family examples, the resulting test is uniformly most powerful similar, uniform for $\psi > \psi_0$, but also uniform over *all* λ.

2. Reduction by marginalization

If we can write

$$f_Y(y;\theta) = f_1(s(y)|t(y);\psi,\lambda) f_2(t(y);\psi),$$

then we're in a somewhat opposite situation, in which H_0 now specifies completely the marginal distribution of T. Among all tests of H_0 that only depend on T, we can find the most powerful against a particular alternative by applying the Neyman-Pearson lemma to f_2, and then check to see if it's uniformly most powerful, etc.. With luck, we can find a UMP test, *among all tests that are functions of* $t(Y)$. Such a factorization will typically be obtainable if the distribution of T is ancillary for λ, when $\psi = \psi_0$. (In the hypothesized factorization above, T is in fact ancillary for λ for all values of ψ.)

Again, more work is needed to justify the optimality of such tests among the class of all tests. This is usually handled by an invariance argument. The details of the argument require a lot of notation, but the basic idea is that if the distribution of T only depends on ψ, and T in some measure extracts *all* the information about ψ, then the best test based on f_2 will also be best among all tests based on f. In applications, T is a maximal invariant statistic for a group of transformations G, say, that leaves H_0 unchanged. It can be proved that the distribution of T depends only on the maximal invariant parameter, which is in applications equal to or equivalent to ψ, the parameter of interest. And it can be proved that all tests that are invariant under the group G must depend only on T. So if we find the optimal test based on f_2, we have automatically found the optimal test among all invariant tests.

3. Remarks

In each of the two cases above, reduction by sufficiency and reduction by ancillarity, there are two somewhat distinct arguments being advanced. The first is pragmatic: if we can factorize the joint density so that we get rid of λ in one factor, we'll just use that factor. The second is optimality: to argue that all 'good' tests must essentially be based on the factor that we want to use. Classical studies of hypothesis testing don't emphasize the pragmatic aspect, which in any case is hard to generalize. On the other hand the restriction to certain classes of tests (similar or invariant) can also be hard to justify, especially when one looks at seemingly innocuous problems for which 'reasonable' tests exist, but similar or invariant tests may not exist.

The factorizations illustrated above may be an oversimplification, in the following sense. H_0 is a composite hypothesis, say $\theta \in \Theta_0$, which I have assumed above takes the form $\psi = \psi_0, \lambda$ unspecified, where ψ and λ are in some way the 'obvious' parametrization of the problem. In fact we might have the conditional density f_1 in 1., or the marginal density f_2 in 2., depending on some *function* of θ that has the same dimension as the argument of the (conditional or marginal) density. That is, the reduction referred to above is a reduction in dimension, which is the important part. So ψ and λ might not be the original parameter of interest and nuisance parameter, but might be some function of those, for which H_0 is unaffected. This seems to happen more often in problems of type 2., where reduction of dimension uses invariance arguments. The t-test provides an example of this phenomenon.

The similar tests constructed in 1. have the property that the size is α throughout the full null hypothesis $\Theta_0 = (\psi_0 \times \Lambda)$. Tests with this property are called *unbiased*. (The precise definition of unbiased is that the power is nowhere less than the size, which in particular means the size must be α throughout H_0.) All similar tests are unbiased, but there exist unbiased tests that are not similar, i.e. not constructed from a conditional distribution that is free of the nuisance parameter. These tests are not in fact useful for problems with nuisance parameters, but the requirement of unbiasedness is also sometimes imposed in the problem of testing a *simple* H_0 against a composite H_1 when no UMP test exists. One example for real θ is $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$.

4. Examples

(a): Reduction by conditioning

1. Ratio of Poissons

   If Y_1 and Y_2 are independent Poissons with mean λ and $\psi\lambda$, then $S = Y_1 + Y_2$ has a Poisson $\lambda + \lambda\psi$ and Y_1 given S has a distribution depending only on ψ, so we are in setting 1.

2. Comparing two Binomials

   If Y_1 and Y_2 are independent Binomials with probability of success p_1 and p_2 respectively, as done in class, then $S = Y_1 + Y_2$ has a distribution which depends on ψ and λ, and the conditional distribution of Y_1 given S depends only on ψ. Again we're in situation 1.. The distribution of Y_1 given S is

   $$f_{Y_1|S}(y_1|s;\psi) = \exp(\psi y_1)\binom{n_1}{y_1}\binom{n_2}{y_2}c(\psi, s),$$

   where $c(\psi, s)$ is a normalizing constant for the conditional density. The most powerful test of $H_0 : \psi = \psi_0$ vs. $H_1 : \psi = \psi_1$ has critical region determined by $f_1(y_1|s;\psi_1) > kf_1(y_1|s;\psi_0)$. Since $f_1(y_1|s;\psi)$ has monotone likelihood ratio in y_1 (for fixed s), this is equivalent to $y_1 > a$, where a is determined from $\Pr\{Y_1 > a|S = s;\psi_0\} = \alpha$.

3. Testing a gamma shape parameter

   Suppose $Y_1, \ldots, Y_n$ are i.i.d. from the gamma density:

   $$f_Y(y;\beta,\mu) = \frac{1}{\Gamma(\beta)}\frac{\beta}{\mu}(\frac{\beta y}{\mu})^{\beta-1}\exp(-y\beta/\mu)$$

   and it is desired to test $H_0 : \beta = \beta_0$, with μ a nuisance parameter. The minimal sufficient statistic for (μ, β) is $S = (\sum \log Y_i, \sum Y_i)$, as can be seen from the form of the joint density for the sample:

   $$f_{Y_1,\ldots,Y_n}(y_1,\ldots,y_n) = \exp\{(\beta-1)\sum \log y_i - \frac{\beta}{\mu}\sum y_i - n\log\{\Gamma(\beta)\} + n\log\left(\frac{\beta}{\mu}\right)\}.$$

The density of S is therefore

$$f_{S_1,S_2}(s_1, s_2; \mu, \beta) = \exp\{(\beta-1)s_1-(\beta/\mu)s_2-n\log\{\Gamma(\beta)\}+n\log\left(\frac{\beta}{\mu}\right)-d(s_1, s_2)\},$$

where d is a fairly complicated function. We don't have to know that function, though, to verify that the conditional distribution of S_1, given S_2, depends only on β, and further that the marginal distribution of S_2 depends on μ and β. (In fact, we know that the marginal distribution of S_2 is Gamma$(\mu, n\beta)$.) Further, the conditional distribution of S_1 has monotone likelihood ratio in s_1, for fixed s_2, so we can get a UMP similar test for onesided alternatives.

4. Normal variance

   The same type of structure obtains for testing $H_0 : \sigma^2 = \sigma_0^2$, with an i.i.d. sample from a $N(\mu, \sigma^2)$ distribution.

   All these examples are testing a canonical parameter or a difference of canonical parameters in a full exponential model. I don't think I know of any examples outside the exponential family, although there undoubtedly are some. In all of the examples the statistic that is sufficient for λ, when ψ is fixed, does not depend on the particular fixed value of ψ. In problems of testing a *ratio* of canonical parameters in a full exponential model, the conditioning statistic does depend on the particular fixed value of ψ. The argument still goes through to give UMP similar tests, however. An example is testing the ratio of two exponential means, with two i.i.d. samples.

(b): Reduction by marginalization

1. Normal mean

   Let $Y_1, \ldots, Y_n$ be i.i.d $N(\mu, \sigma^2)$ random variables, and assume that we want to test the null hypothesis $H_0 : \mu = 0$ against the alternative $H_1 : \mu > 0$. Since we know $(\bar{Y}, S^2)$ is minimal sufficient, and the density of $\bar{Y}$ depends on μ and σ^2, whereas that of S^2 depends only on σ^2, it looks like we won't be able to reduce following pattern 1.. However, note that $T = \bar{Y}/S$ is *invariant* with respect to scale changes ($Y_i \to cY_i$), and so are H_0 and H_1. That is, if we replaced each observation by cY_i, the new mean would be $c\mu$, the new variance would be $c^2\sigma^2$, and the new minimal sufficient statistic would be $(c\bar{Y}, c^2S^2)$. The statistic T is a maximal invariant for the group of scale transformations. Under the group of scale transformations, H_0 and H_1 are unchanged (provided $c > 0$). If we replace T by the equivalent maximal invariant $T' = \sqrt{n}\sqrt{(n-1)}T$ we can easily derive its distribution. T' follows a *noncentral t distribution*, with $n-1$ degrees of freedom and noncentrality parameter μ/σ. So letting $\psi = \mu/\sigma$, and $\lambda = \sigma$, we are in situation 2. above: T' is ancillary for σ, when $\psi = \mu/\sigma$ is fixed, and tests constructed from the marginal distribution of T' are invariant.

Since the noncentral t- distribution has monotone likelihood ratio for all $\psi > 0$, the test that has rejection region $\{y; t'(y) > c\}$ will be UMP invariant. The constant c is determined by $Pr(T' > c; \mu = 0) = \alpha$, and since $T' \sim t_{n-1}$ when $\mu = 0$, this is the usual t-test.

If H_1 is $\mu \neq 0$, then the appropriate group of transformations is as above, with $c \neq 0$. The maximal invariant turns out to be T^2, or equivalently $|T|$, and the UMP invariant test is just the usual two-sided t-test.

Many normal theory examples can be handled by invariance arguments, as can many multivariate normal theory examples. Tests about the mean of the gamma can also be derived by invariance arguments. In fact, tests about the mean of a normal or a gamma can also be derived by conditioning arguments, but the details are tricky and the invariance argument is much simpler. The normal and gamma distributions are the only ones for which methods 1. and 2. can both be applied. Invariance arguments are not available for discrete problems, except in very special cases.