# STAT 801 Lecture 22

**Reading for Today's Lecture:**
**Goals of Today's Lecture:**

- Summarize optimality theory of tests.
- Describe relation of tests to confidence intervals.
- Describe likelihood ratio tests.
- Introduce decision theory.

**Today's notes**

## Optimal tests

- A good test has $\pi(\theta)$ large on the alternative and small on the null.
- For one sided one parameter families with MLR a UMP test exists.
- For two sided or multiparameter families the best to be hoped for is UMP Unbiased or Invariant or Similar.
- Good tests are found as follows:

  1. Use the NP lemma to determine a good rejection region for a simple alternative.
  2. Try to express that region in terms of a statistic whose definition does not depend on the specific alternative.
  3. If this fails impose an additional criterion such as unbiasedness. Then mimic the NP lemma and again try to simplify the rejection region.

## Likelihood Ratio tests

For general composite hypotheses optimality theory is not usually successful in producing an optimal test. instead we look for heuristics to guide our choices. The simplest approach is to consider the likelihood ratio

$$\frac{f_{\theta_1}(X)}{f_{\theta_0}(X)}$$

and choose values of $\theta_1 \in \Theta_1$ and $\theta_0 \in \Theta_0$ which are reasonable estimates of $\theta$ assuming respectively the alternative or null hypothesis is true. The simplest method is to make each $\theta_i$ a maximum likelihood estimate, but maximized only over $\Theta_i$.

**Example 1**: In the $N(\mu, 1)$ problem suppose we want to test $\mu \leq 0$ against $\mu > 0$. (Remember there is a UMP test.) The log likelihood function is

$$-n(\bar{X} - \mu)^2/2$$

If $\bar{X} > 0$ then this function has its global maximum in $\Theta_1$ at $\bar{X}$. Thus $\hat{\mu}_1$ which maximizes $\ell(\mu)$ subject to $\mu > 0$ is $\bar{X}$ if $\bar{X} > 0$. When $\bar{X} \leq 0$ the maximum of $\ell(\mu)$ over $\mu > 0$ is on the boundary, at $\hat{\mu}_1 = 0$. (Technically this is in the null but in this case $\ell(0)$ is the supremum of the values $\ell(\mu)$ for $\mu > 0$. Similarly, the estimate $\hat{\mu}_0$ will be $\bar{X}$ if $\bar{X} \leq 0$ and 0 if $\bar{X} > 0$. It follows that

$$\frac{f_{\theta_1}(X)}{f_{\theta_0}(X)} = \exp\{\ell(\hat{\mu}_1) - \ell(\hat{\mu}_0)\}$$

which simplifies to

$$\exp\{n\bar{X}|\bar{X}|/2\}$$

This is a monotone increasing function of $\bar{X}$ so the rejection region will be of the form $\bar{X} > K$. To get the level right the test will have to reject if $n^{1/2}\bar{X} > z_\alpha$. Notice that the *log likelihood ratio statistic*

$$\lambda \equiv 2\log(\frac{f_{\hat{\mu}_1}(X)}{f_{\hat{\mu}_0}(X)} = n\bar{X}|\bar{X}|$$

as a simpler statistic.

**Example 2**: In the $N(\mu, 1)$ problem suppose we make the null $\mu = 0$. Then the value of $\hat{\mu}_0$ is simply 0 while the maximum of the log-likelihood over the alternative $\mu \neq 0$ occurs at $\bar{X}$. This gives

$$\lambda = n\bar{X}^2$$

which has a $\chi^2_1$ distribution. This test leads to the rejection region $\lambda > (z_{\alpha/2})^2$ which is the usual UMPU test.

**Example 3**: For the $N(\mu, \sigma^2)$ problem testing $\mu = 0$ against $\mu \neq 0$ we must find two estimates of $\mu, \sigma^2$. The maximum of the likelihood over the alternative occurs at the global mle $\bar{X}, \hat{\sigma}^2$. We find

$$\ell(\hat{\mu}, \hat{\sigma}^2) = -n/2 - n\log(\hat{\sigma})$$

We also need to maximize $\ell$ over the null hypothesis. Recall

$$\ell(\mu, \sigma) = -\frac{1}{2\sigma^2} \sum (X_i - \mu)^2 - n \log(\sigma)$$

On the null hypothesis we have $\mu = 0$ and so we must find $\hat{\sigma}_0$ by maximizing

$$\ell(0, \sigma) = -\frac{1}{2\sigma^2} \sum X_i^2 - n \log(\sigma)$$

This leads to

$$\hat{\sigma}_0^2 = \sum X_i^2 / n$$

and

$$\ell(0, \hat{\sigma}_0) = -n/2 - n \log(\hat{\sigma}_0)$$

This gives

$$\lambda = -n \log(\hat{\sigma}^2 / \hat{\sigma}_0^2)$$

Since

$$\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2} = \frac{\sum (X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2 + n\bar{X}^2}$$

we can write

$$\lambda = n \log(1 + t^2/(n-1))$$

where

$$t = \frac{n^{1/2} \bar{X}}{s}$$

is the usual $t$ statistic. The likelihood ratio test thus rejects for large values of $|t|$ which gives the usual test.

Notice that if $n$ is large we have

$$\lambda \approx n[1 + t^2/(n-1) + O(n^{-2})] \approx t^2 \, .$$

Since the $t$ statistic is approximately standard normal if $n$ is large we see that

$$\lambda = 2[\ell(\hat{\theta}_1) - \ell(\hat{\theta}_0)]$$

has nearly a $\chi_1^2$ distribution.

This is a general phenomenon when the null hypothesis being tested is of the form $\phi = 0$. Here is the general theory. Suppose that the vector of $p + q$ parameters $\theta$ can be partitioned into $\theta = (\phi, \gamma)$ with $\phi$ a vector of $p$ parameters and $\gamma$ a vector of $q$ parameters. To test $\phi = \phi_0$ we find two

mles of $\theta$. First the global mle $\hat{\theta} = (\hat{\phi}, \hat{\gamma})$ maximizes the likelihood over $\Theta_1 = \{\theta : \phi \neq \phi_0\}$ (because typically the probability that $\hat{\phi}$ is exactly $\phi_0$ is 0).

Now we maximize the likelihood over the null hypothesis, that is we find $\hat{\theta}_0 = (\phi_0, \hat{\gamma}_0)$ to maximize

$$\ell(\phi_0, \gamma)$$

The log-likelihood ratio statistic is

$$2[\ell(\hat{\theta}) - \ell(\hat{\theta}_0)]$$

Now suppose that the true value of $\theta$ is $\phi_0, \gamma_0$ (so that the null hypothesis is true). The score function is a vector of length $p+q$ and can be partitioned as $U = (U_\phi, U_\gamma)$. The Fisher information matrix can be partitioned as

$$\begin{bmatrix} I_{\phi\phi} & I_{\phi\gamma} \\ I_{\phi\gamma} & I_{\gamma\gamma} \end{bmatrix} .$$

According to our large sample theory for the mle we have

$$\hat{\theta} \approx \theta + I^{-1} U$$

and

$$\hat{\gamma}_0 \approx \gamma_0 + I_{\gamma\gamma}^{-1} U_\gamma$$

If you carry out a two term Taylor expansion of both $\ell(\hat{\theta})$ and $\ell(\hat{\theta}_0)$ around $\theta_0$ you get

$$\ell(\hat{\theta}) \approx \ell(\theta_0) + U^t I^{-1} U + \frac{1}{2} U^t I^{-1} V(\theta) I^{-1} U$$

where $V$ is the second derivative matrix of $\ell$. Remember that $V \approx -I$ and you get

$$2[\ell(\hat{\theta}) - \ell(\theta_0)] \approx U^t I^{-1} U .$$

A similar expansion for $\hat{\theta}_0$ gives

$$2[\ell(\hat{\theta}_0) - \ell(\theta_0)] \approx U_\gamma^t I_{\gamma\gamma}^{-1} U_\gamma .$$

If you subtract these you find that

$$2[\ell(\hat{\theta}) - \ell(\hat{\theta}_0)]$$

can be written in the approximate form

$$U^t M U$$

for a suitable matrix $M$. It is now possible to use the general theory of the distribution of $X^t M X$ where $X$ is $MVN(0, \Sigma)$ to demonstrate that
**Theorem**: The log-likelihood ratio statistic

$$\lambda = 2[\ell(\hat{\theta}) - \ell(\hat{\theta}_0)]$$

has, under the null hypothesis, approximately a $\chi^2_p$ distribution.
**Aside:**
**Theorem**: Suppose that $X \sim MVN(0, \Sigma)$ with $\Sigma$ non-singular and $M$ is a symmetric matrix. If $\Sigma M \Sigma M \Sigma = \Sigma M \Sigma$ then $X^t M X$ has a $\chi^2$ distribution with degrees of freedom $\nu = trace(M\Sigma)$.
**Proof**: We have $X = AZ$ where $AA^t = \Sigma$ and $Z$ is standard multivariate normal. So $X^t M X = Z^t A^t M A Z$. Let $Q = A^t M A$. Since $AA^t = \Sigma$ the condition in the theorem is actually

$$AQQA^t = AQA^t$$

Since $\Sigma$ is non-singular so is $A$. Multiply by $A^{-1}$ on the left and $(A^t)^{-1}$ on the right to discover $QQ = Q$.

The matrix $Q$ is symmetric and so can be written in the form $P\Lambda P^t$ where $\Lambda$ is a diagonal matrix containing the eigenvalues of $Q$ and $P$ is an orthogonal matrix whose columns are the corresponding orthonormal eigenvectors. It follows that we can rewrite

$$Z^t Q Z = (P^t Z)^t \Lambda (PZ)$$

The variable $W = P^t Z$ is multivariate normal with mean 0 and variance covariance matrix $P^t P = I$; that is, $W$ is standard multivariate normal. Now

$$W^t \Lambda W = \sum \lambda_i W_i^2$$

We have established that the general distribution of any quadratic form $X^t M X$ is a linear combination of $\chi^2$ variables. Now go back to the condition $QQ = Q$. If $\lambda$ is an eigenvalue of $Q$ and $v \neq 0$ is a corresponding eigenvector then $QQv = Q(\lambda v) = \lambda Q v = \lambda^2 v$ but also $QQv = Qv = \lambda v$. Thus $\lambda(1 - \lambda)v = 0$. It follows that either $\lambda = 0$ or $\lambda = 1$. This means that the weights

in the linear combination are all 1 or 0 and that $X^tMX$ has a $\chi^2$ distribution with degrees of freedom, $\nu$, equal to the number of $\lambda_i$ which are equal to 1. This is the same as the sum of the $\lambda_i$ so

$$\nu = trace(\Lambda)$$

But

$$\begin{aligned} trace(M\Sigma) &= trace(MAA^t) \\ &= trace(A^tMA) \\ &= trace(Q) \\ &= trace(P\Lambda P^t) \\ &= trace(\Lambda P^tP) \\ &= trace(\Lambda) \end{aligned}$$

In the application $\Sigma$ is $\mathcal{I}$ the Fisher information and $M = \mathcal{I}^{-1} - J$ where

$$J = \begin{bmatrix} 0 & 0 \\ 0 & I_{\gamma\gamma}^{-1} \end{bmatrix}$$

It is easy to check that $M\Sigma$ becomes

$$\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}$$

where $I$ is a $p \times p$ identity matrix. It follows that $M\Sigma M\Sigma = M\Sigma$ and that $trace(M\Sigma) = p$).

## Confidence Sets

A level $\beta$ confidence set for a parameter $\phi(\theta)$ is a random subset $C$, of the set of possible values of $\phi$ such that for each $\theta$ we have

$$P_\theta(\phi(\theta) \in C) \geq \beta$$

Confidence sets are very closely connected with hypothesis tests:

Suppose $C$ is a level $\beta = 1 - \alpha$ confidence set for $\phi$. To test $\phi = \phi_0$ we consider the test which rejects if $\phi \notin C$. This test has level $\alpha$. Conversely, suppose that for each $\phi_0$ we have available a level $\alpha$ test of $\phi = \phi_0$ who rejection region is say $R_{\phi_0}$. Then if we define $C = \{\phi_0 : \phi = \phi_0 \text{ is not rejected}\}$

we get a level $1-\alpha$ confidence for $\phi$. The usual $t$ test gives rise in this way to the usual $t$ confidence intervals

$$\bar{X} \pm t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$$

which you know well.

## Confidence sets from Pivots

**Definition**: A **pivot** (or pivotal quantity) is a function $g(\theta, X)$ whose distribution is the same for all $\theta$. (As usual the $\theta$ in the pivot is the same $\theta$ as the one being used to calculate the distribution of $g(\theta, X)$.

Pivots can be used to generate confidence sets as follows. Pick a set $A$ in the space of possible values for $g$. Let $\beta = P_\theta(g(\theta, X) \in A)$; since $g$ is pivotal $\beta$ is the same for all $\theta$. Now given a data set $X$ solve the relation

$$g(\theta, X) \in A$$

to get

$$\theta \in C(X, A)\,.$$

**Example**: The quantity

$$(n-1)s^2/\sigma^2$$

is a pivot in the $N(\mu, \sigma^2)$ model. It has a $\chi^2_{n-1}$ distribution. Given $\beta = 1-\alpha$ consider the two points $\chi^2_{n-1,1-\alpha/2}$ and $\chi^2_{n-1,\alpha/2}$. Then

$$P(\chi^2_{n-1,1-\alpha/2} \leq (n-1)s^2/\sigma^2 \leq \chi^2_{n-1,\alpha/2}) = \beta$$

for all $\mu, \sigma$. We can solve this relation to get

$$P(\frac{(n-1)^{1/2}s}{\chi_{n-1,\alpha/2}} \leq \sigma \leq \frac{(n-1)^{1/2}s}{\chi_{n-1,1-\alpha/2}}) = \beta$$

so that the interval from $(n-1)^{1/2}s/\chi_{n-1,\alpha/2}$ to $(n-1)^{1/2}s/\chi_{n-1,1-\alpha/2}$ is a level $1-\alpha$ confidence interval.

In the same model we also have

$$P(\chi^2_{n-1,1-\alpha} \leq (n-1)s^2/\sigma^2) = \beta$$

which can be solved to get

$$P(\sigma \leq \frac{(n-1)^{1/2}s}{\chi_{n-1,1-\alpha/2}}) = \beta$$

This gives a level $1-\alpha$ interval $(0, (n-1)^{1/2}s/\chi_{n-1,1-\alpha})$. The right hand end of this interval is usually called a confidence upper bound.

In general the interval from $(n-1)^{1/2}s/\chi_{n-1,\alpha_1}$ to $(n-1)^{1/2}s/\chi_{n-1,1-\alpha_2}$ has level $\beta = 1-\alpha_1-\alpha_2$. For a fixed value of $\beta$ we can minimize the length of the resulting interval numerically. This sort of optimization is rarely used. See your homework for an example of the method.