Stat210B: Theoretical Statistics **Lecture Date: March 20, 2007**

Examples using LeCam's 3rd Lemma and Rank Statistics

Lecturer: Michael I. Jordan *Scribe: Daniel Ting*

1 Examples using LeCam's 3rd Lemma

1.1 Wilcoxon signed rank statistic

The Wilcoxon signed rank statistic is used to test if the location of a sample is equal to zero under the following assumptions.

- f is a density symmetric about θ
- $X_1, X_2, ...X_n \overset{iid}{\sim} f$
- The null hypothesis is $\theta = 0$ and the alternative is $\theta > 0$

The Wilcoxon signed rank statistic is

$$W_n = n^{-3/2} \sum_{i=1}^{n} R_{i,n}^{+} \text{sign}(X_i),$$

where $R_{i,n}^{+}$ = rank of $|X_i|$.

Under the null hypothesis, $\text{sign}(X_i) = 1$ with probability $1/2$ and -1 with probability $1/2$. It is easily verified that the sign and rank are independent, and we have

$$E(W_n) = 0 \ .$$

To get the variance, rewrite W_n as a sum over the ranks versus a sum over the data points.

$$W_n = n^{-3/2} \sum_k k J_k \ ,$$

where

$$J_k = \begin{cases} 1 & \text{if } X_i \text{ is positive, and } |X_i| \text{ is the } k^{th} \text{ largest value} \\ -1 & \text{otherwise} \end{cases} \ .$$

Again using the independence of sign and rank, we obtain

$$\begin{aligned} Var(W_n) &= n^{-3} Var(\sum_k k J_k) = n^{-3} \sum_k k^2 \\ &= n^{-3} \frac{n(n+1)(2n+1)}{6} \to 1/3 \ . \end{aligned}$$

We now have the asymptotic mean and variance of the statistic and must show it is asymptotically normal. One can show that W_n is an asymptotically linear estimator with

$$W_n = n^{-1/2} \sum_i U_i \text{sign}(X_i) + o_p(1),$$

where $U_i = G(|X_i|)$ and G is the edf of $|X_i|$. Indeed, if we replace G with the empirical edf, we recover the Wilcoxon signed rank statistic. Since W_n is asyptotically linear and the $U_i \text{sign}(X_i)$'s are bounded, it follows that W_n is asymptotically normal bt the CLT.

1.1.1 Power under shrinking alternatives

We consider a test where we reject the null, $\theta_0 = 0$, if $W_n > c$ and examine the power of the test under the alternatives $\theta_n = h/\sqrt{n}$

LeCam's 3^{rd} Lemma suggests that we look at $(W_n, \log \frac{dP_{h/\sqrt{n}}}{dP_0})$.

We first consider the case where the density f is normal. In this case,

$$(W_n, \log \frac{dP_{h/\sqrt{n}}}{dP_0}) = (n^{-1/2} \sum_i U_i \text{sign}(X_i), hn^{-1/2} \sum_i X_i - h^2/2) + o_p(1),$$

which is asymptotically bivariate normal with the covariance of the cross term

$$\tau_{12} = hCov_0 \left(G(|X_1|)\text{sign}(X_1), X_1\right) = hE_0 \left(G(|X|)|X|\right) = h/\sqrt{\pi},$$

where the covariance and expectation are taken under the null. The last equality is an exercise in integration. The α level test rejects when $W_n > z_{1-\alpha}/\sqrt{3}$. The asymptotic power of this test under the alternatives is then

$$\begin{aligned} \lim_{n\to\infty} P_{\theta_n}(W_n > z_{1-\alpha}/\sqrt{3}) &= \lim_{n\to\infty} P_{\theta_n}(W_n - h/\sqrt{\pi} > z_{1-\alpha}/\sqrt{3} - h/\sqrt{\pi}) \\ &= 1 - \Phi(z_{1-\alpha} - h\sqrt{3/\pi}). \end{aligned}$$

Dropping the simplifying normal assumption, we can obtain the asymptotic power under quadratic mean differentiability if the density is also square integrable. Recall that the q.m.d. of a location family with density $f(x-\theta)$ is $-f'(x-\theta)/f(x-\theta)$. The covariance term of interest is thus

$$\tau_{12} = Cov_0(U_1 \text{sign}(X_1), -hf'(X_1)/f(X_1)) = 2h \int f^2(x)dx,$$

where the last equality follows from integration by parts. Thus under the alternatives $\theta_n = h/\sqrt{n}$, $W_n \xrightarrow{\theta_n} N(2h \int f^2, 1/3)$.

1.2 Neyman-Pearson statistic

The log-likelihood ratio under the q.m.d. regularity conditon has

$$\log \frac{dP_{\theta + h/\sqrt{n}}}{dP_\theta} \xrightarrow{\theta} N(-\frac{1}{2} h^T I_\theta h, h^T I_\theta h).$$

In this case, the test statistic is the log-likelihood ratio, and the covariance of interest is just the asyptotic variance of the log-likelihood ratio. ie.

$$\tau_{12} = h^T I_\theta h.$$

Thus the asymptotic power of the α level test that rejects when $\log \frac{dP_{\theta+h/\sqrt{n}}}{dP_\theta} > z_{1-\alpha}\sqrt{h^T I_\theta h} - \frac{1}{2}h^T I_\theta h$ is $1 - \Phi(z_{1-\alpha} - \sqrt{h^T I_\theta h})$ under the alternatives.

2 Rank Statistics (van der Vaart, 1998, Chapter 13)

We first introduce some notation.

- Denote the order statistics of $X_1, X_2, ..., X_N$ by

  $$X_{N(1)} \leq X_{N(2)} \leq ... \leq X_{N(N)},$$

  and denote the order statistic vector by $X_{N()}$.

- Let the rank R_{Ni} be the position of X_i in the order statistic, so in the absence of ties, $X_i = X_{N(R_{Ni})}$.

Definition 1 (Linear rank statistics). A statistic is a *linear rank statistic* is it is of the form

$$T_N = \sum_{i=1}^{N} c_{Ni} a_{N,R_{Ni}}.$$

The c_{Ni}'s are called the coefficients and a_{Nk}'s are called the scores.

Here are some examples of linear rank statistics.

Example 2 (Two-Sample problems). Given two independent samples $X_1, ..., X_m$ and $Y_1, ..., Y_n$, the two-sample problem is to determine if both samples came from the same distribution, i.e., the null hypothesis is that both the X_i's and Y_j's have the same distribution. Let $N = m + n$ and R_N be the rank vector of the pooled sample $X_1, ..., X_m, Y_1, ..., Y_n$. Note that under the null, rank statistics are distribution free. They also throw away "irrelevant" properties of the distribution such as scale.

General two-sample problems often use

$$c_{Ni} = \begin{cases} 0 & \text{if } i = 1, ..., m \\ 1 & \text{if } i = m+1, ..., N \end{cases}$$

Example 3 (Wilcoxon statistic).

$$W = \sum_{i=m+1}^{N} R_{Ni}$$

Note that this is not the same as the Wilcoxon signed rank test statistic. Also, note that the Mann-Whitney statistic defined by

$$U = \sum_{i,j} 1_{X_i \leq Y_i}$$

is a U-statistic and equivalent to the Wilcoxon up to an additive constant.