University of Illinois | Department of Economics
Fall 2001 | Roger Koenker

Economics 478
Lecture 4
Two Examples

1. The Log Rank test

Consider the problem of deciding whether there is a significant difference between two survival distributions. One might consider Kolmogorov-Smirnov type tests based on the Kaplan-Meier estimates of the two survival distributions. But a more conventional approach involves the so-called log-rank statistic.

Let N_{it} and Y_{it} $t = t_1, \ldots, t_T$ and $i = 1, 2$ denote the number of observed events and the number at risk in the groups 1 and 2 at the merged ordered event times $t_1, \ldots, t_T$. Let N_t and Y_t denote the corresponding counts in the combined sample. At each observed time we have a two way table

Failure	Group 1	Group 2	Total
Yes	N_{1t}	N_{2t}	N_t
No	$Y_{1t} - N_{1t}$	$Y_{2t} - N_{2t}$	$Y_t - N_t$
Total	Y_{1t}	Y_{2t}	Y_t

Given Y_{it}, recall that this is the number at risk at t and thus predictable wrt $\mathcal{F}_t$, the N_{it} are binomial with sample size Y_{it} and under the null hypothesis of identical survival curves a common failure rate $\lambda(t)$, so approximate event probability $\lambda(t)\Delta t$.

A standard way of evaluating whether the two samples have the same probability is Fisher's "exact" test which is based on conditioning on the marginal total N_t, then

$$\begin{aligned} E_{it} &= EN_{it} = \frac{N_t Y_{it}}{Y_t} \\ V_{it} &= VN_{it} = N_t \, \frac{Y_{1t} Y_{2t}}{Y_t^2} \cdot \frac{Y_t - N_t}{Y_t - 1} \end{aligned}$$

and the log rank statistic

$$T = \sum_{t=1}^{T} (N_{1t} - E_{1t}) / (\sum_{t=1}^{T} V_{1t})^{1/2}$$

This would be all very reasonable if the terms $N_{1t} - E_{1t}$ were independent since then standard CLT results (Lindeberg-Feller) would yield approximate normality. However, this argument isn't really justified here, how should we proceed?

2. Digression on linear rank statistics (for two sample tests of scale)

We have $\underbrace{X_1, \dots, X_m,}_{\text{Sample 1}} \ \underbrace{X_{m+1}, \dots, X_{m+n}}_{\text{Sample 2}}$

We believe that X's come from common distribution, but would like a test to focus on the H_0 that they may differ in scale. Many tests are based on ranking full sample then considering ranks of the first sample $R_1, \dots, R_m$ and forming a linear rank statistic

$$S = \sum_{i=1}^{m} a(R_i)$$

Ideally, we should choose $a(\cdot)$ so that

$$a(i) = E(V_{(i)})$$

where $V_{(i)}$ is i^{th} order statistic from the distribution underlying the hypothesis.

Examples

1. Klotz test take $F = \Phi$ and use

$$S = \sum_{i=1}^{m} \left(\Phi^{-1} \left(\frac{R_i}{m+n+1} \right) \right)^2$$

Note this has an inherent robustness, to deviations from normality

$$ES = \frac{m}{m+n} \sum_{i=1}^{m+n} \Phi \left(\frac{i}{m+n+1} \right)$$

$$V(X) = \frac{mn}{(m+n)(m+n-1)} \sum \left(\Phi^{-1} \left(\frac{i}{m+n+1} \right) \right)^4 - \frac{n}{m(m+n-1)} (ES)^2$$

so

$$T = \frac{S - E(S)}{\sqrt{V(X)}} \sim \mathcal{N}(0,1)$$

Savage Test

$$S = \sum_{i=1}^{m} \left(\sum_{j=m+n+1-R_i}^{m+n} 1/j \right))$$

Note that

$$\begin{aligned} 1 - \sum 1/j &\approx 1 + \sum \log(1 - 1/j) \\ &= 1 + \log \left(\frac{m+n+1-R_i}{m+n+1} \right) \end{aligned}$$

so S is (almost) a sum of log (ranks). Here $ES = m$ and

$$V(S) = \frac{mn}{m+n+1}\left(1 - \frac{1}{m+n}\sum_{j=1}^{m+n}\frac{1}{j}\right)$$

again

$$\frac{S - E(S)}{\sqrt{V(S)}} \rightsquigarrow \mathcal{N}(0,1)$$

and optimality holds when F is exponential.

As usual, write (T_{ij}, δ_{ij}) as the event times and censoring indicators for the two samples $j = 1, 2$, and set

$$N_{ij}(t) = I_{\{T_{ij}<t, \delta_{ij}=1\}}$$
$$Y_{ij}(t) = I_{\{T_{ij}\geq t\}}$$
$$N_i(t) = \sum_{j=1}^{n_i} N_{ij}(t) \qquad N(t) = \sum_{i=1}^{2} N_i(t)$$
$$Y_i(t) = \sum_{j=1}^{n_i} N_{ij}(t) \qquad Y(t) = \sum_{i=1}^{2} Y_i(t)$$

Under the hypothesis of a common failure rate,

$$\begin{aligned}
S_T = \sum_{t=1}^{T}(N_{1t} - E_{1t}) &= \sum_{t=1}^{T} N_{1t} - \sum_{t=1}^{T}\frac{Y_{1t}}{Y_t}N_t \\
&= \sum_{j=1}^{n_1}\int_0^\infty dN_{1j}(s) - \sum_{i=1}^{2}\sum_{j=1}^{n_1}\int_0^\infty \frac{Y_1(s)}{Y(s)}dN_{ij}(s) \\
&= \sum_{j=1}^{n_1}\int_0^\infty \frac{Y_2(s)}{Y(s)}dN_{1j}(s) - \sum_{j=1}^{n_2}\int_0^\infty \frac{Y_1(s)}{Y(s)}dN_{2j}(s)
\end{aligned}$$

Note $\qquad 1 - \frac{Y_1(s)}{Y(s)} = \frac{Y(s)-Y_1(s)}{Y(s)} = \frac{Y_2(s)}{Y(s)}.$

$$\begin{aligned}
&= \sum_{j=1}^{n_1}\int_0^\infty \frac{Y_1(s)Y_2(s)}{Y(s)}(Y_1(s))^{-1}(dN_{1j}(s) - Y_{1j}(s)\lambda(s)ds) \\
&\quad - \sum_{i=1}^{n_2}\int_0^\infty \frac{Y_1(s)Y_2(s)}{Y(s)}(Y_2(s))^{-1}(dN_2j(s) - Y_{2j}(s)\lambda(s)ds) \\
&= \sum_{i=1}^{2}\sum_{j=1}^{n_1}\int_0^\infty (-1)^{i-1}\frac{Y_1(s)Y_2(s)}{Y(s)}(Y_i(s))^{-1}dM_{ij}(s) \\
&= \sum_i\sum_j\int_0^\infty H_i(s)dM_{ij}(s).
\end{aligned}$$

And *this* provides a rationale for the asymptotic normality of the log rank statistic by the arguments of the last lecture.

Under the alternative hypothesis the two samples have different hazards $\lambda_1 \neq \lambda_2$. In this case from above we have

$$\begin{aligned} S_T &= \sum_{j=1}^{n_1} \int \frac{Y_2(s)}{Y(s)} dM_{1j} - \sum_{j=1}^{n_2} \int \frac{Y_1(s)}{Y(s)} dM_{2j} \\ &\qquad + \int \frac{Y_1(s) Y_2(s)}{Y(s)} (\lambda_1(s) - \lambda_2(s)) ds \end{aligned}$$

This provides some insight into the power of tests based on S_T to distinguish λ_1 and λ_2. Local alternatives that have non-trivial power would require that

$$\lim_{n_1 \to \infty; n_2 \to \infty} \left(\frac{n_1 n_2}{n_1 + n_2} \right)^2 (\lambda_1(s) - \lambda_s(s)) = k(s)$$

for some function $k(s)$ that is bounded.

3. The Cox Model

Suppose we have the Cox model

$$\lambda(t|z) = \lambda_0(t) e^{z'\beta}$$

and we have our usual (Y_i, δ_i, z_i) where z_i is a predictable covariate process, i.e. $z(t)$ is left continuous with right limits (cag|ad). The Cox partial likelihood score process is

$$U(\beta) = \sum_{i=1}^{n} \int_0^\infty \left(z_i(s) - \frac{\sum_j Y_i(s) z_i(s) e^{z_j'\beta}}{\sum_j Y_j(s) e^{z_j'\beta}} \right) dN_i(s)$$

and under the null hypothesis that $\beta = \beta_0$ we can write this as,

$$U(\beta_0) = \sum_{i=1}^{n} \int_0^\infty \left(z_i(s) - \frac{\sum Y_j(s) z_i(s) e^{z_j \beta_0}}{\sum Y_j(s) e^{z_j'\beta_0}} \right) dM_i(s)$$

where $M_i(s) = N_i(s) - \int_0^s Y_i(u) \lambda_0(u) \exp\{z_i(s)'\beta_0\} du$.

And this again leads to the conclusion that the score vector is asymptotically normal.

Ref. Fleming and Harrington (1991).

A simple special case of the Cox Model that relates the developments back to the log rank statistic involves the case in which z_i is just a treatment control indicator variable. In this case we have the score

$$U = \sum_{i=1}^{n} \int_0^\infty \left(z_i - \frac{\sum Y_j(s) z_i}{\sum Y_j(s)} \right) dN_i(s)$$

and under H_0 $N_i(s)$ can be replaced by

$$M_i(s) = N_i(s) - \int_0^s I_{(X_i \geq u)} \lambda(u) du$$

Why? Note that

$$dM_i(s) = dN_i(s) - \lambda(s) ds$$

so the claim amounts to saying that

$$\sum_{i=1}^{n} \int_0^\infty \left(z_i - \frac{\sum Y_j(s) z_i}{\sum Y_j(s)} \right) \lambda(s) ds = 0$$

but (obviously)

$$\sum_i \sum_j z_i Y_j(s) = \sum_i \sum_j Y_j(s) z_j \qquad \square$$