

Lecture 16

Testing Hypotheses

Reading: C&O sections 8.6–8.7, K & M sections 7.1–7.3

A common question that we may have is, whether two (or more) samples of survival times may be considered to have been drawn from the same distribution: That is, whether the populations under observation are subject to the same hazard rate.

16.1 Tests in the regression setting

1) A package will produce a test of whether or not a regression coefficient is 0. It uses properties of mle's. Let the coefficient of interest be b say. Then the null hypothesis is $H_0 : b = 0$ and the alternative is $H_A : b \neq 0$. At the 5% significance level, H_0 will be accepted if the p -value $p > 0.05$, and rejected otherwise.

2) In an AL parametric model if α is the shape parameter then we can test $H_0 : \log \alpha = 0$ against the alternative $H_A : \log \alpha \neq 0$. Again mle properties are used and a p-value is produced as above. In the case of the Weibull if we accept $\log \alpha = 0$ then we have the simpler exponential distribution (with $\alpha = 1$).

3) We have already mentioned that, to test Weibull v. exponential with null hypothesis $H_0 : \text{exponential}$ is an acceptable fit, we can use

$$2 \log \hat{L}_{weib} - 2 \log \hat{L}_{exp} \sim \chi^2(1), \text{ asymptotically.}$$

16.2 Non-parametric testing of survival between groups

16.2.1 General principles

We will consider only the case where the data splits into two groups. There is a relatively easy extension to $k > 2$ groups.

We define the following notation

Event times are $0 < t_1 < t_2 < \dots < t_m$.

For $i = 1, 2, \dots, m$, and $j = 1, 2$, $d_{ij} = \#$ events at t_i in group j ,

$n_{ij} = \#$ in risk set at t_i from group j ,

$d_i = \#$ events at t_i ,

$n_i = \#$ in risk set at t_i .

Thus, when the number of groups $k = 2$, we have $d_i = d_{i1} + d_{i2}$ and $n_i = n_{i1} + n_{i2}$.

Generally we are interested in testing the null hypothesis H_0 , that there is no difference between the hazard rates of the two groups, against the two-sided alternative that there is a difference in the hazard rates. The guiding principle is quite elementary, quite similar to our approach to the proportional hazards model: We treat each event time t_i as a new and independent experiment. Under the null hypothesis, the next event is simply a random sample from the risk set. Thus, the probability of the death at time t_i being from group 1 is n_{i1}/n_i , and the probability of it being from group 2 is n_{i2}/n_i .

This describes only the setting where the events all occur at distinct times: That is, d_i are all exactly 1. More generally, the null hypothesis predicts that the group identities of the individuals whose events are at time t_i are like a sample of size d_i without replacement from a collection of n_{i1} '1's and n_{i2} '2's. The distribution of d_{i1} under such sampling is called the hypergeometric distribution. It has

$$\begin{aligned} \text{expectation} &= d_i \frac{n_{i1}}{n_i}, \text{ and} \\ \text{variance} &=: \sigma_i^2 = \frac{n_{i1}n_{i2}(n_i - d_i)d_i}{n_i^2(n_i - 1)}. \end{aligned}$$

Note that if d_i is negligible with respect to n_i , this variance formula reduces to $d_i \binom{n_{i1}}{n_i} \binom{n_{i2}}{n_i}$, which is just the variance of a binomial distribution.

Conditioned on all the events up to time t_i (hence on n_i, n_{i1}, n_{i2}) and on d_i , the random variable $d_{i1} - n_{i1} \frac{d_i}{n_i}$ has expectation 0 and variance σ_i^2 . If we multiply it by an arbitrary weight $W(t_i)$, determined by the data up to time t_i , we still have $W(t_i)(d_{i1} - n_{i1} \frac{d_i}{n_i})$ being a random variable with (conditional) expectation 0, but now (conditional) variance $W(t_i)^2 \sigma_i^2$. This means that if we define for $k = 1, \dots, m$

$$M_k := \left(\sum_{i=1}^k W(t_i) \left(d_{i1} - n_{i1} \frac{d_i}{n_i} \right) \right)_{k=1}^m,$$

these will be random variables with expectation 0 and variance $\sum_{i=1}^k W(t_i)^2 \sigma_i^2$. While the increments are not independent, we may still apply a version of the Central Limit Theorem to show that M_k is approximately normal when the sample size is large enough. (In technical terms, the sequence of random variables M_k is a *martingale*, and the appropriate theorem is the Martingale Central Limit Theorem. See [HH80] for more details.) We then base our tests on the statistic

$$Z := \frac{\sum_{i=1}^m W(t_i) \left(d_{i1} - n_{i1} \frac{d_i}{n_i} \right)}{\sqrt{\sum_{i=1}^m W(t_i)^2 \frac{n_{i1}n_{i2}(n_i - d_i)d_i}{n_i^2(n_i - 1)}}},$$

which should have a standard normal distribution under the null hypothesis.

Note that, as in the Cox regression setting, right censoring and left truncation are automatically taken care of, by appropriate choice of the risk sets.

16.2.2 Standard tests

Any choice of weights $W(t_i)$ defines a valid test. Why do we need weights? Since any choice of weights produces a *correct* test, there is no canonical choice. Changing the weights changes the power with respect to different alternatives. Which alternative you choose — hence, which

weights you choose — should depend on what deviations from equality you are most interested in detecting. As always, the test should be chosen beforehand. Multiple testing makes the interpretation of test results problematic.

Some common choices are:

1. $W(t_i) = 1, \forall i$. This is the **log rank test**, and is the test in most common use. The log rank test is aimed at detecting a consistent difference between hazards in the two groups and is best placed to consider this alternative when the proportional hazard assumption applies. It is maximally asymptotically efficient in the proportional hazards context; in fact, it is equivalent to the score test for the Cox regression parameter being 0, hence is asymptotically equivalent to the likelihood ratio test. A criticism is that it can give too much weight to the later event times when numbers in the risk sets may be relatively small.
2. R. Peto and J. Peto [PP72] proposed a test which emphasises deviations that occur early on, when there are more individuals under observation. **Petos' test** uses a weight dependent on a modified estimated survival function, estimated for the whole study. The modified estimator is

$$\tilde{S}(t) = \prod_{t_i \leq t} \frac{n_i + 1 - d_i}{n_i + 1}$$

and the suggested weight is then

$$W(t_i) = \tilde{S}(t_{i-1}) \frac{n_i}{n_i + 1}$$

This has the advantage of giving more weight to the early events and less to the later ones where the population remaining is smaller.

3. $W(t_i) = n_i$ has also been suggested (Gehan, Breslow). This again downgrades the effect of the later times.
4. D. Harrington and T. Fleming [HF82] proposed a class of tests that include Petos' test and the logrank test as special cases. The **Fleming-Harrington tests** use

$$W(t_i) = \left(\hat{S}(t_{i-1})\right)^p \left(1 - \hat{S}(t_{i-1})\right)^q$$

where \hat{S} is the Kaplan-Meier survival function, estimated for all the data. Then $p = q = 0$ gives the logrank test and $p = 1, q = 0$ gives a test very close to Peto's test and is called the Fleming-Harrington test. If we were to set $p = 0, q > 0$ this would emphasise the later event times if needed for some reason.

All of these tests may be written in the form

$$\frac{\sum(O_{i1} - E_{i1})W_i}{\sqrt{\sum \sigma_{i1}^2 W_i^2}},$$

where O_i and E_i are observed and expected numbers of events. Consequently, positive and negative fluctuations can cancel each other out. This could conceal a substantial difference between hazard rates which is not of the proportional hazards form, but where the hazard rates (for instance) cross over, with group 1 having (say) the higher hazard early, and the lower

hazard later. One way to detect such an effect is with a test statistic to which fluctuations contribute only their absolute values. For instance, we could use the standard χ^2 statistic

$$X := \sum_{i=1}^m \sum_{j=1}^k \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

Asymptotically, this should have the χ^2 distribution with $(k - 1)m$ degrees of freedom. Of course, if the number of groups $k = 2$, this is the same as

$$X := \sum_{i=1}^m \frac{(O_{i1} - E_{i1})^2}{d_i \frac{n_{i1}}{n_i} (1 - \frac{n_{i1}}{n_i})}.$$

16.3 The AML example

We can use these tests to compare the survival of the two groups in the AML experiment discussed in section 12.3. The relevant quantities are tabulated in Table 16.1.

Time	n_{i1}	n_{i2}	d_{i1}	d_{i2}	σ_i^2	Peto weight
5	11	12	0	2	0.476	0.958
8	11	10	0	2	0.474	0.875
9	11	8	1	0	0.244	0.792
12	10	8	0	1	0.247	0.750
13	10	7	1	0	0.242	0.708
18	8	6	1	0	0.245	0.661
23	7	6	1	1	0.456	0.614
27	6	5	0	1	0.248	0.519
30	5	4	0	1	0.247	0.467
31	5	3	1	0	0.234	0.416
33	4	3	0	1	0.245	0.364
34	4	2	1	0	0.222	0.312
43	3	2	0	1	0.240	0.260
45	3	1	0	1	0.188	0.208

Table 16.1: Data for testing equality of survival in AML experiment.

When the weights are all taken equal, we compute $Z = -1.84$, whereas the Peto weights — which reduce the influence of later observations — give us $Z = -1.67$. This yields one-sided p-values of 0.033 and 0.048 respectively — a marginally significant difference — or two-sided p-values of 0.065 and 0.096.

Applying the χ^2 test yields $X = 16.86$, which needs to be compared to χ^2 with 14 degrees of freedom. The resulting p-value is 0.24, which is not at all significant. This should not be seen as surprising: The differences between the two survival curves are clearly mostly in the same direction, so we lose power when applying a test that ignores the direction of the difference.