Survival Analysis: Time To Event Modelling

Sudipta Das

Assistant Professor,
Department of Computer Science,
Ramakrishna Mission Vivekananda Educational & Research Institute

Outline I

Introduction I

- Question: Between two banks which one is possessing riskier loans?
- Question: Does the duration of default depend on gender of the defaulter?
- Statistical Equivalent: There are censored samples from several groups/populations. Each population has its own distribution, Are the two distributions different?

Introduction II

- In the absence of censoring, there are standard answers to this question
  - Nonparametric: Wilcoxon test
- Similar tests exist for other type of censored data
- Nonparametric options available for randomly right-censored data
  - Log-rank test
  - Gehan-Wilcoxon test

Introduction III

- In this part, we shall focus on hypothesis tests that are based on
  - comparing the Nelson–Aalen estimator, obtained directly from the data,
  - to an expected estimator of the cumulative hazard rate, based on the assumed model under the null hypothesis.
- Rather than a direct comparison of these two rates, we shall examine tests that look at weighted differences between the observed and expected hazard rates.

One Sample Test I

- Null hypothesis

$$H_0 : h(t) = h_0(t) \text{ for all } 0 < t \leq \tau$$

- Alternate hypothesis

$$H_A : h(t) \neq h_0(t) \text{ for some } 0 < t \leq \tau$$

One Sample Test II

- Nelson–Aalen estimate of the cumulative hazard function $H(t)$ is

$$\tilde{H}(t) = \sum_{t_i \leq t} \frac{d_i}{Y(t_i)}$$

  - d_i is the number of events at the observed event times, $t_1, \ldots, t_D$
  - $Y(t_i)$ is the number of individuals under study just prior to the observed event time t_i.

One Sample Test III

- A crude estimate of hazard rate $h(t)$ is

$$\hat{h}(t_i) = \frac{d_i}{Y(t_i)}$$

- Under the null hypothesis, the expected hazard rate at t_i is

$$h_0(t_i)$$

- We shall compare the sum of weighted differences between the observed and expected hazard rates to test the null hypothesis.

One Sample Test IV

- Let $W(t)$ be a weight function with the property that $W(t)$ is zero whenever $Y(t)$ is zero.
- Under the null hypothesis
  - The statistic is

$$Z(\tau) = O(\tau) - E(\tau) = \sum_{i=1}^{D} W(t_i)\frac{d_i}{Y(t_i)} - \int_0^{\tau} W(s)h_0(s)ds$$

  - and the sample variance of this statistic is

$$V[Z(\tau)] = \int_0^{\tau} W^2(s)\frac{h_0(s)}{Y(s)}ds$$

One Sample Test V

- For large sample

$$\frac{Z^2(\tau)}{V[Z(\tau)]} \sim \chi^2(1),$$

equivalently

$$\frac{Z(\tau)}{\sqrt{V[Z(\tau)]}} \sim N(0,1)$$

- The null hypothesis is rejected for large values of the statistic.

One Sample Test VI

- One-sample log-rank test

$$W(t) = Y(t)$$

  - $O(\tau) = \sum_{i=1}^{D} W(t_i)\frac{d_i}{Y(t_i)} = \sum_{i=1}^{D} d_i$
    = observed number of events at or prior to time τ,
    - τ is equal to the largest time on study,
  - $E(\tau) = V[Z(\tau)] = \sum_{j=1}^{n}[H_0(T_j) - H_0(L_j)]$, where
    - $H_0(t)$ is the cumulative hazard under the null hypothesis
    - T_j be the time on study for the jth patient, $j = 1, \ldots, n$.
    - L_j be the entry time for the jth patient, $j = 1, \ldots, n$.
- See Example *7.1* at page *203*

One Sample Test VII

- One sample test based on the Harrington and Fleming (1982) family weight function

$$W_{HF}(t) = Y(t)S_0(t)^p[1 - S_0(t)]^q,\ p \geq 0, q \geq 0,$$

where $S_0(t) = e^{-H_0(t)}$ is the hypothesized survival function.

- By choice of p and q,
  - one can put more weight on early departures from the null hypothesis (p much larger than q),
  - late departures from the null hypothesis (p much smaller than q), or
  - on departures in the mid-range ($p = q > 0$).
  - The log-rank weight is a special case of this model with $p = q = 0$.

Tests for Two or More Samples I

- To compare hazard rates of $K (K \geq 2)$ populations.
- Null hypothesis

$$H_0 : h_1(t) = h_2(t) = \cdots = h_K(t), \text{ for all } t \leq \tau,$$

versus

- Alternate hypothesis

$$H_A : \text{ at least one of the } h_j(t)'s \text{ is different for some } t \leq \tau,$$

where τ is the largest time at which all of the groups have at least one subject at risk.

Tests for Two or More Samples II

- The test of H_0 is based on weighted comparisons of the estimated hazard rate of the jth ($j = 1, \ldots, K$) population under the null and alternative hypotheses, based on the Nelson–Aalen estimator.

Tests for Two or More Samples III

- If the null hypothesis is true, then, an estimator of the expected hazard rate in the jth population under H_0, i.e.,

$$d_{ij}/Y_{ij}$$

will be as same as the pooled sample estimator of the hazard rate i.e.,

$$d_i/Y_i,$$

where

  - Death instants are $t_1 < t_2 < \cdots < t_D$
  - d_{ij} be the number of events in the jth sample at time t_i, $i = 1, \ldots, D$
  - Y_{ij} be the number of individuals at risk in the jth sample at time t_i
  - $d_i = \sum_{j=1}^{K} d_{ij}$ be the number of events in the pooled sample at time t_i
  - $Y_i = \sum_{j=1}^{K} Y_{ij}$ be the number of individuals at risk in the pooled sample at time t_i

Tests for Two or More Samples IV

- Test statistics

$$Z_j(\tau) = \sum_{i=1}^{D} W_j(t_i) \left[\frac{d_{ij}}{Y_{ij}} - \frac{d_i}{Y_i} \right], \; j = 1, \ldots, K.$$

  where, $W_j(t)$ be a positive weight function with the property that $W_j(t_i)$ is zero whenever Y_{ij} is zero
- Thus,
  - if all the $Z_j(\tau)$'s are close to zero,
    - then, there is little evidence to believe that the null hypothesis is false,

    whereas,
  - if one of the $Z_j(\tau)$'s is far from zero,
    - then, there is evidence that this population has a hazard rate differing from that expected under the null hypothesis.

Tests for Two or More Samples V

- In practice, we use the weight function as

$$W_j(t_i) = Y_{ij} W(t_i),$$

where $W(t_i)$ is a common weight shared by each group.

- With this choice of weight functions, the test statistics become

$$\begin{aligned} Z_j(\tau) &= \sum_{i=1}^{D} W(t_i) \left[d_{ij} - Y_{ij} \frac{d_i}{Y_i} \right] \\ &= \sum_{i=1}^{D} W(t_i) \left[\frac{d_{ij}}{d_i} - \frac{Y_{ij}}{Y_i} \right] d_i, \; j = 1, \ldots, K. \end{aligned}$$

  - Note that with this class of weights the test statistic is the sum of the weighted difference between the observed number of deaths and the expected number of deaths under H_0 in the jth sample.
  - Also, d_{ij}/d_i is multinomial with parameter 1 and probabilities $p_1 = Y_{i1}/Y_i, \ldots, p_K = Y_{iK}/Y_i$.

Tests for Two or More Samples VI

- The variance of $Z_j(\tau)$ is given by

$$\hat{\sigma}_{jj} = \sum_{i=1}^{D} W^2(t_i) \frac{Y_{ij}}{Y_i} \left(1 - \frac{Y_{ij}}{Y_i}\right) \left(\frac{Y_i - d_i}{Y_i - 1}\right) d_i, \; j = 1, \ldots, K$$

- The covariance of $Z_j(\tau)$ and $Z_g(\tau)$ is expressed by

$$\hat{\sigma}_{jg} = -\sum_{i=1}^{D} W^2(t_i) \frac{Y_{ij}}{Y_i} \frac{Y_{ig}}{Y_i} \left(\frac{Y_i - d_i}{Y_i - 1}\right) d_i, \; g \neq j$$

  - The terms $\frac{Y_{ij}}{Y_i}\left(1 - \frac{Y_{ij}}{Y_i}\right) d_i$ and $-\frac{Y_{ij}}{Y_i}\frac{Y_{ig}}{Y_i} d_i$ arise from the variance and covariance of a multinomial random variable with parameters d_i and $p_j = Y_{ij}/Y_i, j = 1, \ldots, K$.
  - The term $(Y_i - d_i)/(Y_i - 1)$, which equals one if no two individuals have a common event time, is a correction for ties.

Tests for Two or More Samples VII

- Thus, the test statistic is

$$T = [Z_1(\tau) \ \ldots \ Z_{K-1}(\tau)] \, \Sigma^{-1} \, [Z_1(\tau) \ \ldots \ Z_{K-1}(\tau)]^T$$

  - Note that, the components vector $[Z_1(\tau) \ \ldots \ Z_K(\tau)]$ are linearly dependent because $\sum_{j=1}^{K} Z_j(\tau)$ is zero.
  - Thus, the test statistic is constructed by selecting any $K-1$ of the Z_j's.
  - The estimated variance-covariance matrix of these statistics is given by the $(K-1) \times (K-1)$ matrix Σ, formed by the appropriate $\hat{\sigma}_{jg}$'s.

Tests for Two or More Samples VIII

- Under the null hypothesis H_0,

$$T \sim \chi^2(K-1),$$

  if the sample size is large enough.
- An α level test of H_0 reject is rejected in favor of $H_A : h_1(t) \neq h_2(t)$ for some $t \leq \tau$ when $|T| > \chi^2_\alpha(K-1)$.

Tests for Two or More Samples IX

- A special case, when $K = 2$ the test statistic can be written as

$$T = \frac{\sum_{i=1}^{D} W(t_i)\left[d_{i1} - Y_{i1}\frac{d_i}{Y_i}\right]}{\sqrt{\sum_{i=1}^{D} W^2(t_i)\frac{Y_{i1}}{Y_i}\left(1 - \frac{Y_{i1}}{Y_i}\right)\left[\frac{Y_i - d_i}{Y_i - 1}\right]d_i}}$$

- Under the null hypothesis H_0,

$$T \sim N(0, 1),$$

  if the sample size is large enough.
- An α level test of H_0 is rejected
  - in favor of $H_A : h_1(t) > h_2(t)$ for some $t \leq \tau$ when $T > Z_\alpha$.
  - in favor of $H_A : h_1(t) \neq h_2(t)$ for some $t \leq \tau$ when $|T| > Z_{\frac{\alpha}{2}}$.

Tests for Two or More Samples X

- Different weight function yields different tests
  - Log-rank test: $W(t_i) = 1$
    - It gives uniform weight to differences between the observed and expected number of deaths in sample j at time points.
  - Gehan: $W(t_i) = Y_i$
  - Tarone-Ware: $W(t_i) = {Y_i}^{1/2}$
    - These two give more weight to differences between the observed and expected number of deaths in sample j at time points where there is the most data.
    - However, they depend heavily on the event times and censoring distributions.
    - Hence, these weights can have misleading results when the censoring patterns are different in the individual samples

Tests for Two or More Samples XI

- Peto-Peto: $W(t_i) = \tilde{S}(t_i)$
- Modified Peto-Peto: $W(t_i) = \frac{Y_i}{Y_i + 1}\tilde{S}(t_i)$, where

$$\tilde{S}(t_i) = \prod_{t_l \leq t} \left(1 - \frac{d_l}{Y_i + 1}\right)$$

is an estimator based on the combined/pooled sample.

- These two weight functions depend on the combined survival experience in the pooled sample.
- Hence, they overcome the limitations of Gehan or Taone-Ware weights.

Tests for Two or More Samples XII

- Fleming-Harrington: $W_{p,q}(t_i) = \hat{S}(t_{i-1})^p[1 - \hat{S}(t_{i-1})]^q,\ p \geq 0, q \geq 0,$ where $\hat{S}(t)$ be the Product-Limit estimator based on the combined/pooled sample
  - When $p = q = 0$ for this class, we have the log-rank test.
  - When $p = 1, q = 0$, we have a version of the Mann-Whitney-Wilcoxon test.
  - When $q = 0$ and $p > 0$, these weights give the most weight to early departures between the hazard rates in the K populations
  - When $p = 0$ and $q > 0$, these tests give most weight to departures which occur late in time.
  - By an appropriate choice of p and q, one can construct tests which have the most power against alternatives which have the K hazard rates differing over any desired region.

Tests for Two or More Samples XIII

- Note
  - In most applications, we compute the statistics using the log-rank weight $W(t_i) = 1$ and the Gehan weight with $W(t_i) = Y_i$.
  - Because, tests using these weights are available in most statistical packages which makes their application routine in most problems.
  - However, in some applications, one of the other weight functions may be more appropriate, based on the investigator's desire to emphasize either late or early departures between the hazard rates.
- See Example *7.2* at page *209*

Example I

- Example based on bank credit data
- Common survival function and cumulative hazard rate
- **FIGURE 6A**

Example II

- Grouping
  - Split data into two halves, according to sex of customer
- **FIGURE 6B**
- Comparing survival functions grouped by *sex*
  - Null Hypothesis: Data come from same group
    - i.e. survival function does not depend on sex of customer
  - p-value of logrank test is order of 10^{-5}
  - p-value of Gehan-Wilcoxon test is order of 10^{-4}

Example III

- Example contd..
  - Divide data into three parts, according to *skill-set* of customer
- **FIGURE 6C**
- Comparing survival functions grouped by *skill-set*, considering **all groups at a time**
  - Null Hypothesis: Data come from same group
    - i.e. survival function does not depend on skill-set of customer
  - p-value of logrank test is order of 0.001
  - p-value of Gehan-Wilcoxon test is order of 0.001

Example IV

- Comparing survival functions grouped by *skill-set*, considering **two groups at a time**
  - Null Hypothesis: No difference in types of default by *Skilled* and *Highly Skilled* Customer
    - P-value of logrank test is 0.3
    - P-value of Gehan-Wilcoxon test is 0.4
  - Null Hypothesis: No difference in types of default by *Unskilled* and *Skilled* Customer
    - P-value of logrank test is 0.002
    - P-value of Gehan-Wilcoxon test is 0.0009
  - Null Hypothesis: No difference in types of default by *Unskilled* and *Highly* Skilled Customer
    - P-value of logrank test is 0.002
    - P-value of Gehan-Wilcoxon test is 0.001