The three most popular survival analysis methods are KM, log-rank test, and Cox model.

Name	Notation	Properties
Survival Function	$S(t)$	$0 \le S(t) \le 1$, Non-Increasing
Hazard Function (or Rate)	$h(t)$	$h(t) \ge 0$, Any Shape
Cumulative Hazard Function	$H(t)$	$H(t) \ge 0$, Non-Decreasing
PDF of Failure Time	$f(t)$	Non-negative, Integrates to 1

$$H(t) = -\ln S(t)$$
$$S(t) = h(t)f(t)$$
$$f(t) = -S'(t) = S(t)h(t)$$
$$S(t) = \exp[-H(t)]$$
$$h(t) = \frac{d}{dt}H(t) = -\frac{d}{dt}\ln S(t) = -\frac{S'(t)}{S(t)} = \frac{f(t)}{S(t)}$$

Distribution	Hazard Rate $h(t)$	Survival Function $S(t)$	PDF $f(t)$	Mean $E(T)$
Exponential $\lambda > 0, t \ge 0$	λ	$\exp[-\lambda t]$	$\lambda\exp(-\lambda t)$	$1/\lambda$
Weibull $\alpha, \lambda > 0,\ t \ge 0$	$\alpha\lambda t^{\alpha-1}$	$\exp[-\lambda t^{\alpha}]$	$\alpha\lambda t^{\alpha-1}\exp[-\lambda t^{\alpha}]$	$\Gamma(1+1/\alpha)/\lambda^{1/\alpha}$

Mean Survival Time = Area Under Survival Curve. To see this, integrate by parts:

$$\int_0^\infty t f(t)\,dt = -\int_0^\infty tS'(t)\,dt = [t\cdot S(t)]_0^\infty + \int_0^\infty S(t)\,dt$$

If mean survival is estimated from KM curve, and if last is censored, then the mean survival time will be underestimated, since the curve will not go down to zero.

Competing Risks. Let $T = \min(X_1, \ldots, X_k)$. Then the Cause Specific Hazard Rate for risk i is given by

$$h_i(t) = \lim_{\Delta t\to 0}\frac{P(t \le T < t+\Delta t, \delta = 1 \mid T \ge t)}{\Delta t} = \lim_{\Delta t\to 0}\frac{P(t \le X_i < t+\Delta t, \delta = 1 \mid X_j \ge t, j = 1,\ldots,K)}{\Delta t}$$

Then the overall hazard rate is given by $h_T(t) = \sum_{i=1}^{K} h_i(t)$

What makes survival analysis different? Censoring, truncation, non-normality

Censoring Types:

- Type I Censoring is when the censoring time is pre-specified.
- Type II is when you stop observation when a predefined number of events have occurred.
- With **Random Censoring**, Censoring time is not predetermined.
- **Left Censoring** is when you know the event occurred, but you don't know when it happened.
- With **Interval Censoring**, you know that an event occurred between two times, but you don't know when.

Censoring Reasons: LTFU. Withdrawal from Study. Study is Determined. Censoring may be caused by Competing Events.

Key Assumption of Censoring: Independence.

While **Censoring** is about leaving the study, **Truncation** is about not entering the study.

Likelihood Construction. Let C denote the censoring time, $G(t)$ denote $P(C > t)$, and $g(t)$ denote the density at time t for the censoring time distribution. Under independent censoring,

$$\begin{aligned}\text{Likelihood} &= \prod_{i=1}^{n}[f(t_i)G(t_i)]^{\delta_i}[S(t_i)g(t_i)]^{1-\delta_i}\\ &\propto \prod_{i=1}^{n}[h(t_i)]^{\delta_i}\exp[-H(t_i)]\end{aligned}$$

Example. $T \sim \exp(\lambda)$. Then $f(t) = \lambda\exp(-\lambda t)$. Then $S(t) = \exp(-\lambda t)$ and $h(t) = \lambda$, and $H(t) = \lambda t$. Then the likelihood function is given by

$$L = \prod_{i=1}^{n}\lambda^{\delta_i}\exp(-\lambda t_i)$$

where δ_i is the event indicator for subject i. So when $\delta_i = 1$, that means subject i expereienced the event. Otherwise, $\delta_i = 0$.

Kaplan-Meier estimator for estimating the survival function $S(t)$

$$\hat{S}(t) = \prod_{j:t_j\le t}\left(1 - \frac{d_j}{Y_j}\right) \qquad \hat{V}(t) = \hat{S}(t)^2\sum_{t_i\le t}\frac{d_i}{Y_i(Y_i - d_i)}$$

Linear CI for KM: $\hat{S}(t) \pm Z_{\alpha/2}\sqrt{\text{Var}[\hat{S}(t)]}$. May be inappropriate. Maybe use log, double log, or arcsin transformation.

Nelson-Aalen estimator for estimating cumulative hazard function $H(t)$.

$$\tilde{H}(t) = \begin{cases}0, & t \le t_1\\ \sum_{t_i\le t}(d_i/Y_i), & t_1 \le t\end{cases} \qquad \hat{\sigma}^2_{\tilde{H}}(t) = \sum_{t_i\le t}\frac{d_i}{Y_i^2}$$

We prefer to use KM to estimate $S(t)$ and NA to estimate $H(t)$.

Both KM and NA are non-parametric, but we do need to make assumptions about independent censoring.

Mantel-Haenszel Test and its relationship to the **Log-Rank Test**. The MH statistic is

$$MH = \frac{\sum_{k=1}^{D}[d_{1k} - E(d_{1k})]}{\sqrt{\sum_{k=1}^{D}\text{Var}(d_{1k})}} \sim \text{N}(0,1), \quad H_0 : S_1(t) = S_2(t) \leftrightarrow H_0 : h_1(t) = h_2(t), \quad T \sim N(0,1)$$

This is the same as the Log-Rank Test!

Log-rank vs. Wilcoxon Tests. With the Log-rank test, we have equal weights across time. With the Wilcoxon test, the weight is the number of subjects at risk, so the Wilcoxon test assigns more weight to early-on events.

Question. Under what situation will the log-rank test be more powerful than the Wilcoxon test? When the hazard rate constant. In this case, all events are weighted the same.

Test for Trend with K Populations. Assume there are K samples. Null and alternative hypotheses:

$$H_0 : h_1(t) = h_2(t) = \cdots = h_k(t), \quad \forall t < \tau, \qquad H_A : h_1(t) \le h_2(t) \le \cdots \le h_k(t), \quad \forall t < \tau, \text{ at least one strict inequality}$$

where τ is the smallest final time point, among all samples. The test statistics are $Z_1, Z_2, \ldots, Z_K$. Then the overall test statistic is

$$Z = \frac{\sum_{j=1}^{K} a_j Z_j(\tau)}{\sqrt{\text{variance}}} \sim N(0,1), \quad a_j = j(\text{score})$$

Stratified Tests. Adjust for covariates. Regression: put the covariates into the model. Alternatively, you can use the covariates to stratify the data. Assume K populations and M strata. Null hypothesis:

$$H_0 : h_{1s}(t) = h_{2s}(t) = \cdots = h_{ks}(t), \quad \forall t < \tau, \quad s = 1, \ldots, M$$

Here we have $K \times M$ test statistics: Z_{js}, $j = 1, \ldots, K$, $s = 1, \ldots, M$. Then we sum over the M strata: $Z_j = \sum_{s=1}^{M} Z_{js}$. The test statistic will follow a χ^2_{K-1} distribution.

Stratified Tests. Mathched Pairs. Within ear pair (strata), subjects were dependent. Assume two samples and M pairs. Same null hypothesis as the one for stratified tests. Here, the test statistic is D_1 = number of subjects in sample 1 who had the event first, while D_2 = number of subjects in sample 2 who had the event first. Then the test statistic is $\frac{D_1 - D_2}{\sqrt{D_1 + D_2}} \sim N(0,1)$

Power and sample size calculation for logrank test. The alternative hypothesis for the Log-Rank Test: $H_0 : h_1(t) \neq h_2(t)$. The test statistic T is ditributed as $N(\phi, 1)$ when $h_1(t) \approx h_2(t)$. Under the proportional hazard assumptions, we have $\frac{h_2(t)}{h_1(t)} = e^{\theta}$. Define $\phi = \theta\sqrt{\pi(1-\pi)D}$, where D is the expected total number of deaths under the alternative and π is the proportion in, say, group 2. The required total numberof deaths to have power of $1 - \beta$ is

$$D = \frac{[z_{1-\alpha/2} + z_{1-\beta}]^2}{\theta^2 \pi(1-\pi)}$$

Competing Risks: Logrank vs. Gray's tests. How do the log-rank and Gray's test differ? Cause-specific hazard vs. sub-distrubtion hazard. Difference is the risk set. For the cause-specific hazard, the risk set decreases with time. For the sub-distribution hazard, individuals who fail from a competing cause remain in the risk set until their potential censoring time.

Cox Model Notation. Let X denote the event time, let C denote the censoring time, and let $T = \min(X, C)$ denote the observed time. Let δ denote the censoring indicator, with $\delta = 1$ if $T = X$, and $\delta = 0$ otherwise. Let $\boldsymbol{Z}$ denote time-independent covariates. Then the hazard function for a Cox Model is expressed as $h(t|Z) = h_0(t)c(\boldsymbol{Z}; \boldsymbol{\beta})$, where $h_0(t)$ is an unspecified baseline hazard function (nuisance parameter function). A common choice of $c(\boldsymbol{Z}; \boldsymbol{\beta}) = \exp(\boldsymbol{\beta Z}) = \exp(\beta_1 z_1 + \cdots + \beta_p Z_p)$. The *Linear Model formulation for the covariate effects* is obtained by taking a log transformation: $\log \frac{h(t|\boldsymbol{Z})}{h_0(t)} = \boldsymbol{\beta Z} = \beta_1 z_1 + \cdots + \beta_p z_p$. Also, since the numerator and denomonator share the same baseline hazard function, the hazard ratio is given by

$$\frac{h(t|\boldsymbol{Z})}{h(t|\boldsymbol{Z}^*)} = \frac{h_0(t)\exp[\sum_{k=1}^{p} \beta_k Z_k]}{h_0(t)\exp[\sum_{k=1}^{p} \beta_k Z_k^*]} = \exp\left[\beta_k (Z_k - Z_k^*)\right]$$

Note that the Hazard ratio is similar to the odds ratio of Logisitic Regression.

Since there is no time involved for the exponential function, the cumulative hazard function is given by $H(t|\boldsymbol{Z}) = H_0(t)\exp(\boldsymbol{\beta Z})$. Then we have $S(t|\boldsymbol{Z}) = \exp[-H(t|\boldsymbol{Z})]$, which can be written as

$$S(t|\boldsymbol{Z}) = \exp\left[-H_0(t)\exp \boldsymbol{\beta Z}\right] = [S_0(t)]^{e^{\boldsymbol{\beta Z}}}$$

Comments. Weibull is a PH model: $h(t|z_1) = \lambda p t^{p-1}$, $\lambda = e^{\beta z_1}$, $h_0(t) = p t^{p-1}$.

Interaction Between Two Categorical Covariates. Assume two binary covariates Z_1 and Z_2, each of which belonging to $\{0, 1\}$. Consider the Cox model

$$h(t|Z_1, Z_2) = h_0(t)\exp(\beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_1 Z_2)$$

We say that β_1 and β_2 are the coefficients for the main effects Z_1 and Z_2, and that β_3 is the interaction effect between Z_1 and Z_2. What is $\exp(\beta_3)$? Then

$$\exp(\beta_3) = \frac{h(t|Z_1 = Z_2 = 0) \times h(t|Z_1 = Z_2 = 1)}{h(t|Z_1 = 1, Z_2 = 0) \times h(t|Z_1 = 0, Z_2 = 1)}$$

	$Z_1 = 0$	$Z_1 = 1$
$Z_2 = 0$	g1	g2
$Z_2 = 1$	g3	g4

Note: there is a different way to specify this model. Since there are four groups, indicators for three of the groups (g2, g3, g4) and leave the other group (g1) as the reference group.

Cox Model. Estimation. Suppose the Hazard function is given by $h(t|\boldsymbol{Z}) = h_0(t)\exp(\boldsymbol{\beta Z})$. Consider the probability that an individual dies at t_j, given that there is one death at t_j. Then this leads to the partial likelihood $L = \prod_{j=1}^{n} L_j$, where

$$L_j = \frac{h(t_j|Z_j)}{\sum_{i \in R(t_j)} h(t_j|Z_j)} = \frac{h_0(t_j)\exp(\boldsymbol{\beta Z}_j)}{\sum_{i \in R(t_j)} \exp(\boldsymbol{\beta Z}_i)} = \frac{\exp(\boldsymbol{\beta Z}_j)}{\sum_{i \in R(t_j)} \exp(\boldsymbol{\beta Z}_i)}$$

Here, we do not consider probabilities for censored events. We compute the Hazard Ratio (HR) θ by $HR = \exp(\boldsymbol{\beta})$, the HR estimate by $\hat{HR} = \exp(\hat{\beta})$, the 95% CI for β by $\hat{\beta} \pm 1.96\text{se}(\hat{\beta})$, and the 95% CI for θ by $[\exp(\hat{\beta} - 1.96 \times \text{se}\hat{\beta}), \exp(\hat{\beta} + 1.96 \times \text{se}\hat{\beta})]$. Note that the baseline hazard function is not involved in the HR formula.

Assumptions of Cox Model. We assume that the HR is independent of time. The Hazard ratio for two Z's are proportional. An example of when the PH assumption is not satisfied: the hazard functions cross.

How to Evaluate the Predictability of a Cox Model? Suppose that we have a Cox model $h(t) = h_0(t)\exp(\boldsymbol{\beta Z})$ Let $g(\boldsymbol{Z}) = \boldsymbol{\beta Z}$ denote the estimated risk score for subjects with $\boldsymbol{Z}$. Then, for $T_2 > T_1$, what would be considered as concordant in terms of $g(\boldsymbol{Z})$? It would be $g(\boldsymbol{Z_1}) > g(\boldsymbol{Z_2})$