Introduction to Survival Analysis

David M. Rocke

March 5, 2024

Time to Event Data

- *Survival Analysis* is a term for analyzing time-to-event data.
- This is used in clinical and epidemiological studies, where the event is often death or incidence or recurrence of disease.
- It is used in engineering reliability analysis, where the event is often failure of a device or system.
- It is used in insurance, particularly life insurance, where the event is death, disability, or damage from an accident.

Time to Event Data

- The distribution of 'failure' times is usually asymmetric and can be long-tailed.
- The base distribution is not normal, but exponential. This is the simplest distribution to model failure time data.
- There are often *censored* or *truncated* observations, which are ones in which the failure time is not observed.
- Typically, this is because the failure has not yet happened, though there are other patterns.

Time to Event Data

- Often, these are *right-censored*, meaning that we know that the event occurred after some known time t, but we don't know the actual event time, as when a patient is still alive at the end of the study.
- Observations can also be *left-censored*, meaning we know the event has already happened at time t, or *interval-censored*, meaning that we only know that the event happened between times t_1 and t_2.
- Analysis is difficult if censoring is associated with treatment or other predictors of the event in question.

Right Censoring

- Patients are in a clinical trial for cancer, some on a new treatment and some on standard of care.
- Some patients in each group have died by the end of the study. We know the survival time (measured for example from time of diagnosis—each person on their own clock).
- Patients still alive at the end of the study are right censored.
- Patients who are lost to follow-up or withdraw from the study may be right-censored.

Left and Interval Censoring

- An individual tests positive for HIV.
- If the event is infection with HIV, then we only know that it has occurred before the testing time t, so this is left censored.
- If an individual has a negative HIV test at time t_1 and a positive HIV test at time t_2, then the infection event is interval censored.

Engineering Reliability Example

- Engineering reliability studies often use parametric survival models, the simplest of which is the exponential distribution.
- The following example is based on information provide by Seagate about one of their disk drive models in terms of likelihood of failure.
- A common statistics given is MTBF = mean time between failures, which is equal to the mean lifetime under the exponential distribution.

Computer Disk Drives

Here is an example excerpt from a Product Manual, in this case for the Seagate Barracuda ES.2 Near-Line Serial ATA drive:

The product shall achieve an Annualized Failure Rate (AFR) of 0.73% (Mean Time Between Failures (MTBF) of 1.2 Million hrs) when operated in an environment that ensures the HDA case temperatures do not exceed 40° C. Operation at case temperatures outside the specifications in Section 2.9 may increase the product Annualized Failure Rate (decrease MTBF).

AFR and MTBF are population statistics that are not relevant to individual units.
AFR and MTBF specifications are based on the following assumptions for business critical storage system environments:

- *8,760 power-on-hours per year.*
- *250 average motor start/stop cycles per year.*
- *Operations at nominal voltages.*

Systems will provide adequate cooling to ensure the case temperatures do not exceed 40° C. Temperatures outside the specifications in Section 2.9 will increase the product AFR and decrease MTBF.

Computer Disk Drives

- 1.2 million hours at 8,760/hours per year (365×24) is 137 years! The exponential parameter in years is $1/137 = 0.0073$.
- How can this be tested!
- Assuming exponential failures, the average time in years until the first failure out of n units is $137/n$ and an estimate of the exponential parameter is n times the first failure time.

Computer Disk Drives

- With 1000 disk drives, the mean waiting time would be 0.137 years or less than 2 months.
- To find the mean time until failure k, we need to use the gamma distribution. The chance of 4 or more failures in 6 months is about 0.5.
- Accelerated failure time methods vs. temperature is more feasible.
- These figures are not credible to anyone who has ever had a disk drive!

Poisson Process

- A *Poisson Process* places random points in a continuous space, usually $\mathbb{R}, \mathbb{R}^2, \mathbb{R}^3$ or a subset thereof.
- The *complete independence* property of a Poisson process is that for any pre-specified collection of disjoint, bounded subregions of the space, the random variables indicating the number of points in each subregion are statistically independent.
- For a *homogeneous Poisson process*, the probability that there are *n* points in a region depends only on the measure (length, area, volume) of the region.

Poisson Process

- For survival analysis, we are interested in a counting process on the positive real line. This can be defined as a random process that generates a random function $N(t), t \geq 0$, defined as the number of points in the interval $(0, t]$.
- For a homogeneous Poisson process, there is a parameter λ such that the mathematical expectation of the number of points in an interval of length t is λt.
- The probability mass function for the number of points n in an interval of length t is

$$f(n; \lambda t) = \frac{(\lambda t)^n e^{-\lambda t}}{n!}$$

which is called the *Poisson distribution*.

Exponential Distribution

If the points on the line are generated by a homogeneous Poisson process with parameter λ and $t_0 \geq 0$ is any pre-chosen point on the line, then the distance between t_0 and the point of the process that has the smallest distance forward from t_0 has a distance x defined by the exponential density

$$f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

This is the *waiting time* until the next event.

Gamma Distribution

- In cases where more than one event can happen, we can measure the time to the k^{th} event forward from a particular time. That waiting time has a gamma distribution defined by

$$f(x; \lambda, k) = \frac{x^{k-1} e^{-\lambda x} \lambda^k}{\Gamma(k)}$$

- For `R`, using the gamma distribution, scale $= \lambda$ and shape $= k$.

Disk Drive Failure

- If the failures of a particular type of disk drive form a homogeneous Poisson process on the real line with parameter λ and if we have m disk drives on test with independent failures, then the pooled failure times form a Poisson process with parameter $\lambda^* = m\lambda$.
- The probability that an interval of length T contains n points is Poisson with parameter $m\lambda$.
- The time until the k^{th} failure has a gamma distribution with scale $m\lambda$ and shape k.

Counting Process

- In general, the time to an event can be viewed as the result of a counting process, but one without necessarily the same value of λ thoughout time.
- The hazard λ can depend on characteristics of the individual and to vary over time.
- But the exponential model is still interesting as the simplest example of time to event data.

Basic Quantities and Models

The probability density function $f(x)$ is defined as with any continuous distribution. For any short interval of time, it can be thought of as the relative chance that the event will occur in that short interval. The cumulative distribution function is

$$F(x) = \Pr(X \leq x) = \int_0^x f(x)dx$$

For survival data, a more relevant quantity is the *survival function*

$$S(x) = 1 - F(x) = \Pr(X > x) = \int_x^\infty f(x)dx$$

Basic Quantities and Models

$$S(x) = 1 - F(x) = \Pr(X > x) = \int_x^\infty f(x)dx$$

The survival function $S(x)$ is the probability that the event time is later than x. If the event in a clinical trial is death, then this is the fraction of the original population at time 0 that is still alive at time x; that is, the fraction surviving to time x.

The Hazard Function

Another important function is the *hazard function*, which is the probability that the event will occur in the next very short interval, given that it has not occurred yet.

$$h(x) = \lim_{\Delta x \to 0} \frac{\Pr[x \leq X < x + \Delta x | X \geq x]}{\Delta x}$$

The expression in the numerator is the probability of survival until at least time $x + \Delta x$ conditional on surviving until time x. This might be the chance of someone who has just turned 30 still being alive one day later.

The Hazard Function

$$h(x) = \lim_{\Delta x \to 0} \frac{\Pr[x \leq X < x + \Delta x | X \geq x]}{\Delta x}$$

This might be the chance of someone who has just turned 30 still being alive one day later. You can see that this is different than the probability at birth of surviving until age 30 plus one day. The first is the ratio of the number of those who die at age 30 plus one day over the number alive at age 30. The second is a ratio with the same numerator, but with the larger denominator of the number who are born. The latter ratio is smaller.

The Hazard Function

$$\begin{aligned} h(x) &= \lim_{\Delta x \to 0} \frac{\Pr[x \le X < x + \Delta x | X \ge x]}{\Delta x} \\ &= S^{-1}(x) \lim_{\Delta x \to 0} \frac{\Pr[x \le X < x + \Delta x]}{\Delta x} \\ &= f(x)/S(x) \end{aligned}$$

The limit takes the difference quotient into a derivative (by definition of the derivative) and the result is because the density $f(x)$ is the derivative of the CDF $F(x)$.

The Hazard Function

Also,

$$
\begin{aligned}
h(x) &= \lim_{\Delta x \to 0} \frac{\Pr[x \le X < x + \Delta x | X \ge x]}{\Delta x} \\
&= f(x)/S(x) \\
f(x) &= -\frac{dS(x)}{dx} \qquad \text{Because } F' = f \\
h(x) &= -\frac{d\ln(S(x))}{dx} = -S^{-1}(x)\frac{dS(x)}{dx}
\end{aligned}
$$

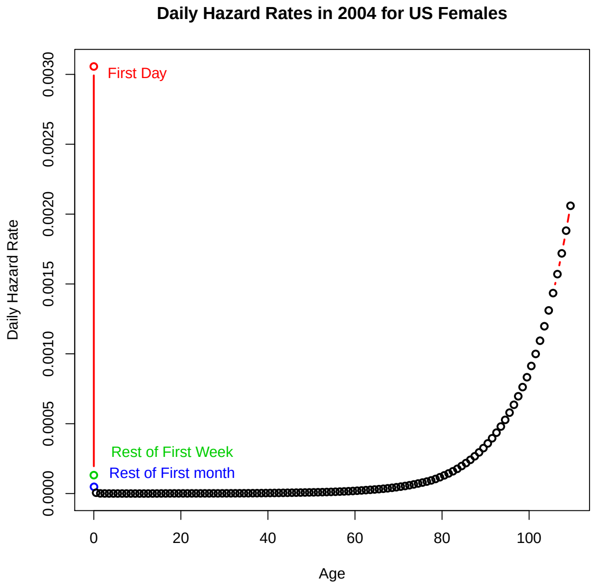
Cumulative Hazard

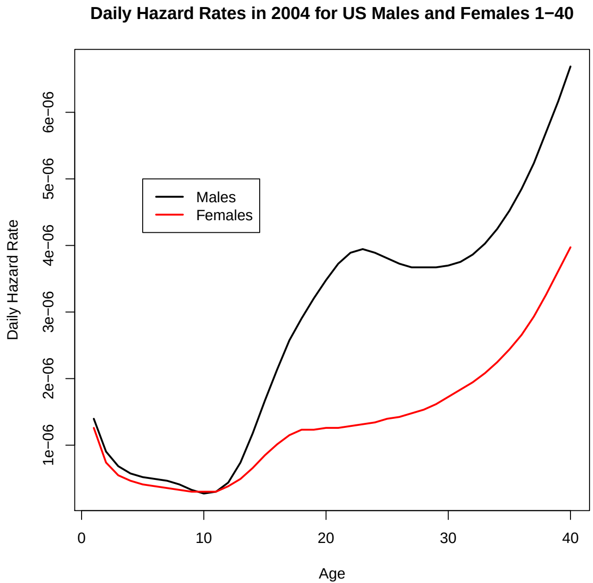
$$h(x) = -\frac{d \ln(S(x))}{dx}$$

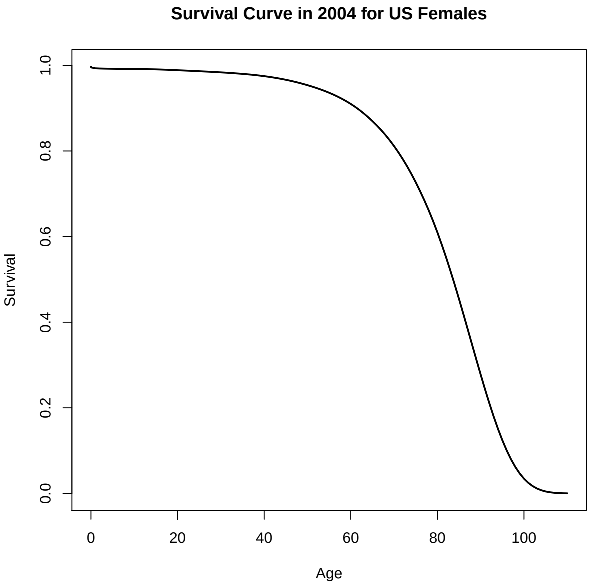
The cumulative hazard function is

$$H(x) = \int_0^x h(t)dt = -\ln(S(x))$$

This function is easier to estimate than the hazard function, and we can then approximate the hazard function by the approximate derivative of the cumulative hazard.


Daily Hazard Rates in 2004 for US Females
First Day
Rest of First Week
Rest of First month
Daily Hazard Rate
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
0.0030
Age
0
20
40
60
80
100


Daily Hazard Rates in 2004 for US Males and Females 1−40
Daily Hazard Rate
1e−06
2e−06
3e−06
4e−06
5e−06
6e−06
Males
Females
0
10
20
30
40
Age


Survival Curve in 2004 for US Females
Survival
Age
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100

Exponential Distribution

- The exponential distribution is the base distribution for survival analysis.
- The distribution has a constant hazard λ which makes it the simplest survival distribution in that sense.
- The mean survival time is λ^{-1}

$$
\begin{aligned}
f(x;\lambda) &= \lambda e^{-\lambda x} \quad \text{density} = \text{likelihood} \\
\ln(f(x;\lambda)) &= \ln\lambda - \lambda x \quad \text{log likelihood} \\
\frac{\partial}{\partial\lambda}\ln(f(x;\lambda)) &= \lambda^{-1} - x \\
F(x) &= 1 - e^{-\lambda x} \\
S(X) &= e^{-\lambda x} \\
\ln(S(x)) &= -\lambda x \\
h(x) &= -\frac{d}{dx}\ln(S(x)) \\
&= -\frac{d}{dx}(-\lambda x) \\
&= \lambda
\end{aligned}
$$

Estimation of λ

- Suppose we have m exponential survival times of $t_1, t_2, \dots, t_m$ and k right-censored values at $u_1, u_2, \dots, u_k$.
- A survival time of $t_i = 10$ means that subject i died at time 10. A right-censored time $u_i = 10$ means that at time 10, subject i was still alive and that we have no further follow-up.
- For the moment we will assume that the survival distribution is exponential and that all the subjects have the same parameter λ.

Estimation of λ

- A naive estimate of λ is the average of the survival times of the of those subjects who died, $m^{-1} \sum t_i$, but this is not correct because it ignores the k subjects that are still alive.
- Suppose one subject died at 1 day, and the rest were still alive at 10 years. One day is a poor estimate of average survival (although this is often the first thing that statistically naive investigators think of).
- This estimate of average survival could be too small or too large.

Estimation of λ

- Another naive estimate of λ is the average of the times of all the subjects, $(m+k)^{-1}[\sum t_i + \sum u_i]$, but this is not correct either because it treats the subjects who are still alive as though they had just died.
- This estimate of average survival is too small if any of the subjects are censored.

Estimation of λ

- Suppose we have m exponential survival times of $t_1, t_2, \ldots, t_m$ and k right-censored values at $u_1, u_2, \ldots, u_k$.
- A survival time of $t_i = 10$ means that subject i died at time 10. A right-censored time $u_i = 10$ means that at time 10, subject i was still alive and that we have no further follow-up.
- For the moment we will assume that the survival distribution is exponential and that all the subjects have the same parameter λ.

Estimation of λ

We have m exponential survival times of $t_1, t_2, \dots, t_m$ and k right-censored values at $u_1, u_2, \dots, u_k$. The log-likelihood of an observed survival time t_i is

$$\ln\left(\lambda e^{-\lambda t_i}\right) = \ln\lambda - \lambda t_i$$

and the likelihood of a censored value is the probability of that outcome (survival greater than u_j) so the log-likelihood is

$$\log(e^{-\lambda u_j}) = -\lambda u_j.$$

Let $T = \sum t_i$ and $U = \sum u_j$. Then the log likelihood is

$$\sum_{i=1}^{m}(\ln\lambda - \lambda t_i) + \sum_{j=1}^{k}(-\lambda u_j) = m\ln\lambda - \lambda(T + U)$$

$$m \ln \lambda - \lambda(T + U)$$

is maximized when the derivative wrt λ is 0, that is when

$$\begin{aligned} 0 &= m/\hat{\lambda} - (T + U) \\ \hat{\lambda} &= m/(T + U) \\ 1/\hat{\lambda} &= (T + U)/m \end{aligned}$$

Thus, the estimated mean survival is the total of the times, exact and censored, divided by the number of exact times. It can be show that the variance of $\hat{\lambda}$ is asymptotically λ^2/m, depending only on the number of uncensored observations. This is generally true.

Suppose that we have two groups with m items in each group, where the mean time in each group is $\bar{x}$. If the times in group 1 are failures and the times in group 2 are censored, vs. both are failures, then

$$\begin{aligned}
1/\hat{\lambda} &= (m\bar{x} + m\bar{x})/m \\
&= \bar{x} + \bar{x} = 2\bar{x} \\
\hat{V}(\hat{\lambda}) &= \hat{\lambda}^2/m
\end{aligned}$$

$$\begin{aligned}
1/\hat{\lambda} &= (m\bar{x} + m\bar{x})/(2m) \\
&= (\bar{x} + \bar{x})/2 = \bar{x} \\
\hat{V}(\hat{\lambda}) &= \hat{\lambda}^2/(2m)
\end{aligned}$$

The Score and the Fisher Information

The log likelihood is

$$\ell\ell = m \ln \lambda - (T + U)\lambda$$

and its derivative, called the *score*, is

$$\ell\ell' = m/\lambda - (T + U)$$

Under certain conditions, the negative derivative of the score, called the *Fisher Information*, estimates the reciprocal of the variance of the MLE.

The score is

$$\ell\ell' = m/\lambda - (T + U)$$

(which is 0 evaluated at the MLE) and the observed Fisher information is

$$-\ell\ell'' = m/\hat{\lambda}^2$$

and its reciprocal is

$$\hat{\lambda}^2/m$$

Although the value of $\hat{\lambda}$ depends on both the uncensored data and the censored data, the variance depends only on the uncensored sample size.

- The (expected value of the) score statistic is zero when evaluated at the MLE.
- The larger the second derivative of the log likelihood is, the steeper the fall-off from the MLE and the more certainly we know the true parameter.
- The multivariate generalization of the Fisher information is most times the method of determining the variance covariance matrix for Wald tests.
- Or we can use the likelihood ratio chi-squared test and interval from inverting this test (profile likelihood).

Multivariate Generalization

- If there are p parameters, then the score is the gradient vector of length p of partial derivatives of the log likelihood. This determines the estimates by solving p equations in p unknowns setting the score vector to the zero vector.
- The Hessian H is the matrix of second partials and its negative inverse evaluated at the MLE's estimates the variance covariance matrix of the estimated parameters.

- If we have a null hypothesis for the exponential parameter λ

$$H_0 : \lambda = \lambda_0$$

  then the log likelihood at the MLE is

$$\ell\ell = m \ln(m/(T+U)) - m$$

  and at the null hypothesis is

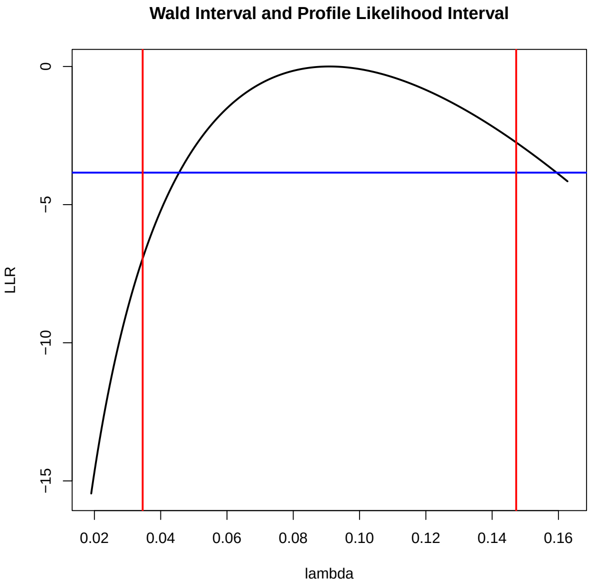
$$\ell\ell = m \ln \lambda_0 - (T+U)\lambda_0$$

- The likelihood ratio statistics is the negative of twice the difference between the log likelihood at the null and the log likelihood at the MLE.

- We can construct a confidence interval for λ in two ways: using the asymptotic normal approximation or the likelihood ratio statistics.
- The plot on the next slide is for $m = 10, k = 5, T = 100, U = 10$ with

$$\begin{aligned} \hat{\lambda} &= 0.0909 \\ \hat{se}(\lambda) &= 0.02875 \end{aligned}$$

- The red lines are at ± 1.96 standard errors away from the MLE.
- The blue line is at the chisquare statistic with 5% in the tail and 1 df and intersects the likelihood curve to form another interval.


Wald Interval and Profile Likelihood Interval
LLR
0
−5
−10
−15
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
lambda

- The Wald test/interval and the LLR test/profile likelihood interval are both asymptotically accurate subject to assumptions.
- Frequently, the convergence of the LLR procedures to asymptopia is faster than that of the Wald procedures.
- We could check this by simulation under the assumptions.
- Also, the profile likelihood procedures are unchanged by transformations in the parameter—the same for λ as for the mean λ^{-1}; this is not true of the Wald procedures.