Counting Process and Survival Analysis

Casually Explained with Assistance from Dr. Tsiatis

SAHIL PATEL

Graduate Student at the North Carolina State University

sspate27@ncsu.edu

Contents

SUBSECTION 0.1

Things to Fix When I have Time

0.1.1 Set Up

- Fix the terminology presented in the competing risks framework to make them more clean **and** consistent

- describe little and big O notation more so that its more motivated? Probably not necessary though after the first example. Also that whole example could be explained so poorly.

- The stuff from Lemma 2.2.3 is hard to understand what it means to "cancel" the du's and leaves a lot to be understood about the actual evaluation of these things.

0.1.2 One Sample Counting Process

- The switch between random variables and data in the notation could possibly be confusing, like in the NA estimator

- define a stochastic process?

- unbiasedness of NA notation is *whack*

SUBSECTION 0.2

Counting Process

- left continuity of Y() maybe should be discussed?

0.2.1 Appendix A

- Talk about 3.13 and the generalities of stochastic integration (convergence and necessity of right continuity)

- mention the steiljes approximation and requirements of left continuity (3.13/3.14)

Set Up

Basic Survival Definitions

When doing survival analysis, we have for an observation i, we collect information about their failure or censoring time:

- T_i failure time

- U_i censoring time

- $X_i = min(T_i, U_i) = T_i \wedge U_i$ are the observation times

- $\delta_i = \mathbb{I}\{T_i \leq U_i\}$ is the failure indicator. It is 1 if we observe the failure, 0 otherwise

- Z_i is some covariate vector.

We start with T to be the failure time, and consequently a non-negative random variable with pdf $F_T(t) = P_T(T \leq t)$ that is right continuous ($F_T(t^-) = F_T(t)$) as with any cdf. We then define unique functions associated with T.

Defn 2.1.1 (Survival Function) The **survival function** $S(t)$ is the left continuous function that is defined as the probability of surviving after a point in time t.

$$S(t) := P(T \geq t)$$

Which is consequently left continuous $S(t^+) = S(t)$. We can also define a right continuous version where:

$$S(t^-) = P_T(T > t) = 1 - F_T(t)$$

Lem 2.1.1 If T is continuous then $S(t^-) = S(t) = 1 - F(t)$, and we get that:

$$f(t) = \frac{\partial}{\partial t}F(t) = -\frac{\partial}{\partial t}S(t)^1$$

and we also get that clearly $F(t) = \int_0^t f(u)du$ and that $S(t) = \int_t^\infty f(u)du$

$^1 -\frac{\partial}{\partial t}S(t) = -\frac{\partial}{\partial t}\{1 - F(t)\} = f(t)$

Defn 2.1.2 (Mortality Function) The **mortality rate**, $m(t)$, is the proportion of the population who fail between times t and $t+1$ among the people who are still alive at time t. This can be mathematically expressed as:

$$m(t) = P(t \leq T < t+1 \mid T \geq t)$$

Defn 2.1.3 (Hazard Function) We can then define the **hazard rate** is the limit of the mortality rate, or the instantaneous rate of failure at time t given the individual is alive at time

t. Expressed as:

$$\lambda(t) = \lim_{h \to 0} \frac{P(t \leq T < t + h \mid T \geq t)}{h}$$

We can then easily see that there are alternate forms to express the hazard rate based on the previous lemma.

Lem 2.1.2

$$\begin{aligned}
\lambda(t) = \lim_{h \to 0} \frac{P(t \leq T < t + h \mid T \geq t)}{h} &= \lim_{h \to 0} \frac{P(t \leq T < t + h)/h}{P(T \geq t)} \\
&= \frac{f(t)}{S(t)} = \frac{\frac{\partial}{\partial x} F(x)}{S(t)} \\
&= \frac{-\frac{\partial}{\partial t} S(t)}{S(t)} \quad \text{by Lemma 2.1.1} \\
&= -\frac{\partial}{\partial t} \log\{S(t)\}
\end{aligned}$$

Defn 2.1.4 (Cumulative Hazard Function) The **cumulative hazard function,** is just the cumulative hazard function up to a point t, or expressed as:

$$\Lambda(t) = \int_0^t \lambda(u) du$$

Lem 2.1.3 We can find an alternate form of the cumulative hazard through:

$$\begin{aligned}
\Lambda(t) &= \int_0^t \lambda(u) du \\
&= \int_0^t -\frac{\partial}{\partial t} \log\{S(t)\} \quad \text{by Lemma 2.1.2} \\
&= -\log\{S(t)\}
\end{aligned}$$

which then implies that $S(t) = \exp\{-\Lambda(t)\}$

It's important to understand that the hazard rate isn't a probability and thus it can take on any positive value, unlike mortality which is bounded by 1. With the above lemmas, we can now give a new form to the mortality rate

Lem 2.1.4 A new form of the mortality rate takes the form:

$$\begin{aligned}
m(t) &= P(t \leq T < t + 1 \mid T \geq t) \quad \text{by Definition 2.1.2} \\
&= \frac{P(t \leq T < t + 1}{P(T \geq t)} \\
&= \frac{P(T \geq t) - P(T \geq t + 1)}{P(T \geq t)} \\
&= 1 - \frac{P(T \geq t + 1)}{P(T \geq t)} \\
&= 1 - \frac{S(t + 1)}{S(t)} \quad \text{by Definition 2.1.1} \\
&= 1 - \frac{\exp\{-\Lambda(t + 1)\}}{\exp\{-\Lambda(t)\}} \quad \text{by Lemma 2.1.1}
\end{aligned}$$

$$= 1 - \exp\{-\Lambda(t+1) + \Lambda(t)\}$$

$$= 1 - \exp\left\{-\int_0^{t+1} \lambda(u)du + \int_0^t \lambda(u)du\right\} \text{ by Definition 2.1.4}$$

$$= 1 - \exp\left\{-\int_t^{t+1} \lambda(u)du\right\}$$

Then note if the probability of an event happening in a single unit of time (event between t and $t+1$) then the hazard is low (if the hazard also doesn't drastically change) then we can provide an approximation for the mortality as:

$$m(t) = 1 - \exp\left\{-\int_t^{t+1} \lambda(u)du\right\}$$

$$\approx 1 - \left\{1 - \int_t^{t+1} \lambda(u)du\right\} \qquad \text{we use the approximation } e^x = 1 + x \text{ for small x}$$

$$= \int_t^{t+1} \lambda(u)du$$

$$\approx \lambda(t)$$

Competing Risks

The whole purpose of survival analysis is essentially to handle this issue of censoring. Usually, this is an issue of right censoring, where we observe the starting time for all (usually the time of randomization in an experiment) then fail to record when they "fail" (die). This happens to the person withdrawing from the study or the study ending before the person dies. There is a whole other issue of *competing risks* where the person dies from a cause not associated to what is deemed as a failure. For example, if we are testing a cancer treatment, our definition of failure is death by cancer as it would describe how well the cancer treatment prolongs life. However, if a person dies in a car accident, we are censored from their failure time due to cancer, and the car accident is deemed as a competing risk.

In this case, we observe $U_i = min\{T_1, ..., T_k\}$ where T_i represents a failure time distribution for the ith cause of failure. Since the cause of failure is known to us, we let $\Delta = j$ if $U_i = T_j$ (if we observed that the failure was cause by the jth way to fail). *It might be good to keep in mind that some T_l could represent a "censoring failure time".* From here, we can define the rest of the competing risk framework:

Defn 2.2.1 (Competing Risks Framework) Similar definitions of all of the above functions exist in the competition risk framework. We now just define a function for each failure model T_j, but technically these aren't observable because we only observe U_i

- **Cause-Specific Survival Function:** $S^{(j)}(x) = P(T_j \geq x) = \exp\left\{-\Lambda^{(j)}(x)\right\}$

- **Cause-Specific Hazard Function:** $\lambda^{(j)}(x) = -\frac{\partial}{\partial x}\log\left\{S^{(j)}(x)\right\}$

- **Cause-Specific Cumulative Hazard:** $\Lambda^{(j)}(x) = \int_0^x \lambda^{(j)}(u)du$

As mentioned, we only observe $(U_i, \Delta_i)_{i=1,...,n}$. So let's describe these distributions:

- **Observed Sub-Survival Function:** $S_U^{(j)}(t) = P(U \geq t, \Delta = j)$

- **Total Observed Survival:** $S_U(x) = P(U \geq x) = \sum_{j=1}^{k} S_U^{(j)}(x)$

- **Marginal Observed CDF:** $F_U^{(j)}(t) = P(U \leq t, \Delta = j)$

- **Cause Specific Hazard:** $\lambda_U^{(j)}(x) = \lim_{h \to 0} \left\{ \frac{P(x \leq U < x+h, \Delta = j | U \geq x)}{h} \right\}$

- **Total Observed Hazard:** $\lambda_U(x) = \sum_{j=1}^{k} \lambda_U^{(j)}(x)$

- **Observed Cumulative Hazard:** $\Lambda_U^{(j)}(x) = \int_0^x \lambda_U^{(j)}(u) du$

- **Total Observed Cumulative Hazard:** $\Lambda_U(x) = \sum_{j=1}^{k} \Lambda_U^{(j)}(x)$

Note that the *observed* survival function denotes the probability of serving beyond or at time t and eventually dying from cause j. This is compared to the *net* survival function that is the probability of surviving at or beyond time t if only cause j was acting on the population.

Theorem 2.2.1 | **(1:1 Relationship between Sub-Survival and Cause-Specific Hazard Functions)** This theorem shows that there is a one-to-one relationship between the K sub-distribution functions and the K cause specific hazard functions. This theorem is important as it allows us to say that *either* the survival or hazard function can be used to describe the joint distribution of (U, Δ)

PROOF | To prove this claim, we show that each function can be derived from the other.

- *Cause specific hazard to sub-survival:*

$$
\begin{aligned}
\lambda_U^{(j)}(x) &= \lim_{h \to 0} \left\{ \frac{P(x \leq U < x+h, \Delta = j \mid U \geq x)}{h} \right\} \text{ by Definition 2.2.1} \\
&= \frac{\lim_{h \to 0} \left\{ h^{-1} P(x \leq U < x+h, \Delta = j) \right\}}{P(U \geq x)} \text{ by defn of marginal prob} \\
&= \frac{\lim_{h \to 0} \left\{ h^{-1} \left[S_U^{(j)}(x) - S_U^{(j)}(x+h) \right] \right\}}{S_U(x)} \text{ by Definition 2.2.1} \\
&= \frac{-\frac{\partial}{\partial x} S_U^{(j)}(x)}{S_U(x)} \\
&= \frac{\frac{\partial}{\partial x} F_U^{(j)}(x)}{S_U(x)} \text{ by Lemma 2.1.1}
\end{aligned}
$$

- *Sub-survival to cause-specific hazard:*
Note that $S_U(x) = \exp\{-\Lambda_U(x)\}$ by Lemma 2.1.1. Then by the above equation

$$
\lambda_U^{(j)}(x) = \frac{\frac{\partial}{\partial x} F_U^{(j)}(x)}{S_U(x)} = \frac{\frac{\partial}{\partial x} F_U^{(j)}(x)}{\exp\{-\Lambda_U(x)\}}
$$

$$
\implies \frac{\partial}{\partial x} F_U^{(j)}(x) = \lambda_U^{(j)}(x) \exp\{-\Lambda_U(x)\}
$$

$$
\implies F_U(x) = \int_0^x \lambda_U^{(j)}(u) \exp\{-\Lambda_U(u)\} \, du \text{ after integrating}
$$

$$\equiv S_U^{(j)}(x) = \int_x^\infty \lambda_U^{(j)}(u)\exp\left\{-\Lambda_U(u)\right\} du$$

Thus we are able to recover each equation from another, meaning they are unique for whatever joint distribution of (U, Δ) that corresponds to the problem. □

Naturally in the problem of competing risks, we don't care about $S_U(x)$ or $S_U^{(j)}(x)$ as it aggregates all of the causes of death together, we instead care mostly about T_j acts, or more specifically, how $S^{(j)}(x) = P(T_j \geq x)$ acts as it would give us information about that one cause of failure. However, we only observe instances of (U_i, Δ_i) not T_j. In the case that T_j's are independent, we can actually recover $S^{(j)}(x)$

2.0.1 A Mathematical Aside on O-notation

An understanding of this notation is needed for a theorem coming up, so we switch gears to just understanding some math with a quick example to show how this works.

Defn 2.2.2 (Big $\mathcal{O}$-Notation) $\mathcal{O}$-notation describes the order of a function $f(h)$. $f(h)$ is $\mathcal{O}(h^\alpha)$ (notated as $f(h) = \mathcal{O}(h^\alpha)$) if:
$$\frac{|f(h)|}{|h|^\alpha}$$
is bounded for h in a neighbourhood of 0.

Defn 2.2.3 (Little o-Notation) This notation describes the *small order(?)* of a function $f(h)$. We say $f(h) = o(h^\alpha)$ if:
$$\lim_{h \to 0} \frac{|f(h)|}{|h|^\alpha} = 0$$

We illustrate these two definitions with some examples that will come in handy in the next proof (that's why they're called lemmas)!

Lem 2.2.1 An important example of $\mathcal{O}$-notation:

$$P\{(x \leq T_j < x + h) \cap_{i \neq j} (T_i > T_j)\} = P\{(x \leq T_j < x + h) \cap_{i \neq j} (T_i > x)\} + \mathcal{O}(h^2)$$

PROOF To show this fact is true, we equivalently (and more clear from the notation) that:

$$\underbrace{P\{(x \leq T_j < x + h) \cap_{i \neq j} (T_i > T_j)\} - P\{(x \leq T_j < x + h) \cap_{i \neq j} (T_i > x)\}}_{:=f(h)} = \mathcal{O}(h^2)$$

Which requires us to show that the following is bounded:

$$\frac{|f(h)|}{h^2}$$

I'll take a graphical approach to solving the problem. It's first important to recognize that we are bounding a difference in probability, so let's start by plotting the probability space.
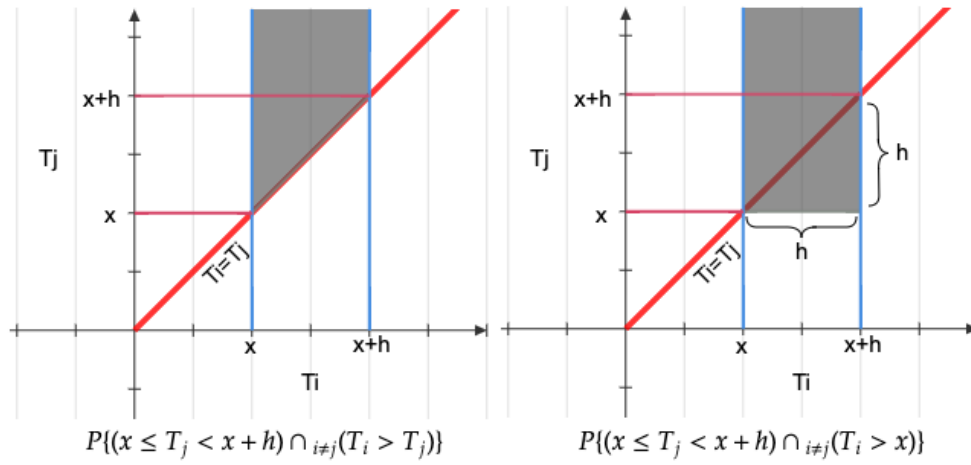
$$P\{(x \le T_j < x + h) \cap_{i \ne j}(T_i > T_j)\} \qquad P\{(x \le T_j < x + h) \cap_{i \ne j}(T_i > x)\}$$

Figure 1. Event space of both events mentioned in $f(h)$ denoted by the grey space

$\square$

Notice how the right side is larger by a triangle of length/width of h, meaning the volume differs by $\frac{1}{2}h^2$. This implies that $f(h) = P(\textit{triangle space})$ which, by the volume of a triangle, is $O(h^2)$

We now prove one more result before going back to survival.

Lem 2.2.2

$$f(h) = O(h^2) \implies f(h) = o(h)$$

PROOF

$$f(h) = O(h^2) \implies \exists M : \frac{|f(h)|}{h^2} \le M \text{ by Definition 2.2.2}$$
$$\implies \frac{|f(h)|}{h} \le hM$$
$$\implies \lim_{h \to 0} \frac{|f(h)|}{h} = 0$$
$$\implies f(h) = o(h) \text{ by Definition 2.2.3}$$

$\square$

Theorem 2.2.2

(**The independence of failure causes implies that the observed cause-specific hazard function is equal to the true cause-specific hazard function**) Mathematically, this is presented as:

$\forall i \ne j, T_i \perp T_j \implies$

$\lambda_U^{(j)}(x) = \lambda^{(j)}(x)\forall j$ or, equivalently

$$\lim_{h \to 0} \left\{ \frac{P(x \le U < x + h, \Delta = j \mid U \ge x)}{h} \right\} = \lim_{h \to 0} \left\{ \frac{P(x \le T_j < x + h \mid T_j \ge x)}{h} \right\}$$

Where, again (U, Δ) is the observed failure time and failure cause.

PROOF

$$\lambda_U^{(j)}(x) = \lim_{h \to 0} \left\{ \frac{P(x \le U < x + h, \Delta = j \mid U \ge x)}{h} \right\} \text{ by Definition 2.2.1}$$

$$= \frac{\lim_{h \to 0} \left\{ h^{-1} P(x \le U < x + h, \Delta = j) \right\}}{P(U \ge x)}$$

$$= \frac{\lim_{h \to 0} \left\{ h^{-1} P[(x \le T_j < x + h) \cap_{l \ne j} (T_l > T_j)] \right\}}{P\left[\cap_{l=1}^k (T_l \ge x) \right]}$$

$$= \frac{\lim_{h \to 0} \left\{ h^{-1} P[(x \le T_j < x + h) \cap_{l \ne j} (T_l > x)] + \mathcal{O}(h^2) \right\}}{P\left[\cap_{l=1}^k (T_l \ge x) \right]} \text{ by Lemma 2.2.1}$$

$$= \frac{\lim_{h \to 0} \left\{ h^{-1} P[(x \le T_j < x + h) \cap_{l \ne j} (T_l > x)] \right\}}{P\left[\cap_{l=1}^k (T_l \ge x) \right]} \text{ by Lemma 2.2.2} \qquad \text{as } O(h^2) = o(h) \to 0 \text{, as } h \to 0$$

$$= \frac{\lim_{h \to 0} \left\{ h^{-1} P[(x \le T_j < x + h)] \right\} [P(\cap_{l \ne j}(T_l > x))]}{P\left[\cap_{l=1}^k (T_l \ge x) \right]}$$

$$= \frac{\lim_{h \to 0} \left\{ h^{-1} P[(x \le T_j < x + h)] \right\}}{P[T_j \ge x]}$$

$$= \lambda^{(j)}(x) \text{ by Definition 2.2.1}$$

□

Lem 2.2.3 | If each competing risk $T_1, ..., T_k$ are mutually independent, we can rewrite cause-specific hazard as a function of the observed data.

$$S^{(j)}(x) = \exp\left\{ -\Lambda^{(j)}(x) \right\} \text{ by Definition 2.2.1}$$

$$= \exp\left\{ -\Lambda_U^{(j)}(x) \right\} \text{ by Theorem 2.2.2}$$

$$= \exp\left\{ -\int_0^x \frac{\frac{\partial}{\partial u} F_U^{(j)}(u)}{S_U(u)} du \right\} \text{ by Definition 2.2.1 and Theorem 2.2.1}$$

$$= \exp\left\{ -\int_0^x \frac{dF_U^{(j)}(u)}{S_U(u)} \right\} \text{ cancelling } du\text{'s}$$

From the above, we can also see that if competing risks are independent, we can rewrite the cumulative hazard:

$$\Lambda^{(j)}(x) = \int_0^x \frac{dF_U^{(j)}(u)}{S_U(u)}$$

The above lemma gives us a function for the cause-specific hazard, as a function of the observed data $(U_i, \Delta_i)_{i=1}^N$, thus making it estimable.

SECTION 3

Dependent Competing Risks

The problem with the estimation arises when the causes of death aren't independent. Each cause-specific survival function becomes unidentifiable, as we can no longer use Theorem 2.2.2. So far, the most we can do is provide a bound for a cause-specific hazard function, as studied by Peterson (1976)

Theorem 2.3.3 | **(Bounds for Cause-Specific Hazard in Nonidentifiable Competing Risks)**
Suppose we have 2 non-independent competing risks, then the bounds for a cause-specific survival function can be described as:

$$S_U^{(1)}(t) + S_U^{(2)}(t) \leq S^{(1)} \leq S_U^{(1)}(t) + S_U^{(2)}(0)$$

PROOF | While Dr. Tsiatis presents this proof more through diagrams and illustrations, I find that the original proof from Peterson (1976) is more "exact" and relatively easy to follow.

We start by noticing the following:

$$S^{(1)}(t) = P(T_1 \geq t) = \underbrace{P(T_1 \geq t, T_2 \geq 0, T_1 \leq T_2)}_{:=A} + \underbrace{P(T_1 \geq t, T_2 \geq 0, T_1 > T_2)}_{:=B}$$

Where $T_2 > 0$ is always true, as failure time must be a non-negative random variable, and we then split up the probability into the two cases by the last term.
We start by finding the *upper bounds* for A and B:

A.

$$P(T_1 \geq t, T_2 \geq 0, T_1 \leq T_2) = P(T_1 \geq t, T_1 \leq T_2) \text{ as } T_2 > 0 \text{ is vacuously true}$$
$$= P(T_1 \geq t, \Delta = 1) \text{ as } T_1 \leq T_2 \text{ implies failure from cause 1}$$
$$= S_U^{(1)}(t) \text{ By Definition 2.2.1}$$

B.

$$P(T_1 \geq t, T_2 \geq 0, T_1 > T_2) \leq P(T_2 \geq 0, T_1 > T_2) \text{ less restrictions} \implies \text{greater probability}$$
$$= P(T_2 \geq 0, \Delta = 2) \text{ similar to above}$$
$$= S_U^{(2)}(0) \text{ by Definition 2.2.1}$$

We then find the *lower bounds* for A, and B

A.

$$P(T_1 \geq t, T_2 \geq 0, T_1 \leq T_2) \geq P(T_1 \geq t, T_2 \geq t, T_1 \leq T_2) \text{ more restrictive} \implies \text{less prob}$$
$$= P(T_1 \geq t, T_1 \leq T_2) \text{ as } T_2 > t \text{ is vacuously true if } T_2 > T_1 \geq t$$
$$= P(T_1 \geq t, \Delta = 1) \text{ if } T_2 > T_1 \text{ then failure is from cause 1}$$
$$= S_U^{(1)}(t) \text{ by Definition 2.2.1}$$

B.

$$P(T_1 \geq t, T_2 \geq 0, T_1 > T_2) \geq P(T_1 \geq t, T_2 \geq t, T_1 > T_2) \text{ more restrictive} \implies \text{less prob}$$
$$= P(T_2 \geq t, T_1 > T_2) \text{ as } T_1 > t \text{ is vacuously true if } T_1 > T_2 \geq t$$
$$= P(T_2 \geq t, \Delta = 2) \text{ if } T_1 > T_2 \text{ then failure is from cause 2}$$
$$= S_U^{(2)}(t) \text{ by Definition 2.2.1}$$

Then combine all of the solved bounds to obtain the desired result. □

One-Sample Counting Processes

We start looking at the one-sample problem of estimating a survival distribution from a sample of possibly right-censored observations. Censored data can be thought of as a case of competing risks, where the time to failure is a competition of the time to death, T, and censoring time C. Our primary goal is to estimate $S_T(x) = P(T \geq x)$.

Defn 3.3.1 | (Censored Observation) A censored observation is based on the random variables of time to death, T, and time to failure, C. An observation in a case of possible censoring is:

$$(U, \Delta):$$
$$U = min(T, C)$$
$$\Delta = \begin{cases} 1 & T \leq C \text{ failure} \\ 0 & T > C \text{ censoring} \end{cases}$$

Where we typically observe n observations that are iid $(U_i, \Delta_i)_{i=1}^n$

SECTION 4

Nelson-Aalen Estimator for $\Lambda^{(1)}(t)$

It is typical, in the one sample case, for us to want to find what the cumulative hazard is for death time. Recall from Lemma 2.2.3, that in the case that $C \perp T$, that the cumulative hazard, derived from the data, can be written as:

$$\Lambda^{(1)}(t) = \int_0^t \frac{dF_U^{(1)}(k)}{S_U(k)}$$

It is from this that we can get a natural estimator for the cumulative hazard function based on empirical distribution functions:

Defn 3.4.1 | (Nelson-Aalen Estimator for Cumulative Hazard) Based on the data, and independence of censoring and failure time (though I think this can be extended to any number of independent competing risks), we get the **Nelson-Aalen estimator for cumulative hazard**:

$$\widehat{\Lambda}_{NA}^{(1)}(t) := \int_0^t \frac{d\widehat{F}_U^{(1)}(k)}{\widehat{S}_U(k)} = \int_0^t \frac{dN_U^{(1)}(k)}{Y_U(k)}$$

Where, by the definition of what an empirical CDF is:

$$\widehat{F}_U^{(1)}(k) := \frac{1}{n} \underbrace{\sum_{i=1}^{n} \mathbb{I}\{u_i \leq k, \Delta_i = 1\}}_{:=N_U^{(1)}(k)}$$

$$\widehat{S}_U(k) := \frac{1}{n} \underbrace{\sum_{i=1}^{n} \mathbb{I}\{u_i \geq k\}}_{:=Y_U(k)}$$

As for any verbal explanation of the meaning of these newly defined functions, $N_U(k)$ *counts* the number of deaths that are observed in the sample up to (and including) time k. On the contrary, $Y_U(k)$ denotes the number of people in the sample who are still at risk (neither censored nor dead) at time k. Both of these give us a first idea of a *counting process*, as they are both *counting* the number of events that happens over time.

Defn 3.4.2 ("General" Counting Process) Non-mathematically, a **counting process** is a stochastic process which has sample paths which are right continuous with left-hand limits (cad lag), and are non-decreasing step functions taking jumps of size 1.
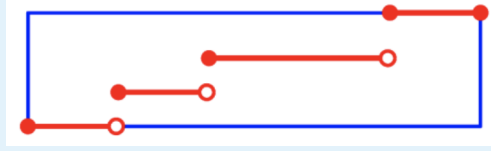

Figure 2. Example Counting Process, each jump is of size 1

Now that we have figured out what $\widehat{\Lambda}_{NA}^{(1)}$, and how to write it as counting processes, we need to understand how to evaluate this integral and turn it from something abstract into something computable. The requires *a lot* of math on integration. I will assume that the reader doesn't have much/any knowledge of measure theory and I will, hopefully, try to build up all of the relevant, calculus, then analysis, and measure to solve this integrand with respect to a counting process that I have laid out in Definition 3.4.1. **If you aren't sure how these integrals convert to sums, and want a mathematical understanding head to** Part V (not anywhere close to being done).

Lem 3.4.1 Because $\widehat{\Lambda}_N A(t)$ is an integral with respect to a counting process of jump sizes of 1, we can rewrite the integral as a sum whenever the counting process takes a jump:

$$\widehat{\Lambda}_{NA}(t) = \int_0^t \frac{1}{Y_U(k)} dN_U^{(1)}(k) \text{ by Definition 3.4.1}$$

$$= \sum_{i:u_i \leq t, \Delta_i = 1} \frac{1}{Y_U(u_i)}$$

$$= \sum_{i:u_i \leq t, \Delta_i = 1} \left\{ \sum_{j=1}^{n} \mathbb{I}\{u_j \geq u_i\} \right\}^{-1}$$

The first sum (over i's) represents the sum over all death times less than t. This is

the sum that was created by the counting process. If you didn't want to read Part V, this is just because the counting process only takes values of 1 at each death time, so we have a non-zero weight at each death time.

The second sum is the number of people still alive after the *ith* person dies...inverted.

Theorem 3.4.4 (0-Difference of Expectation of Nelson-Aalen Estimator and True Cumulative Hazard) Assuming the independence of the failure and censoring times, we can show that $\mathbb{E}\left\{\widehat{\Lambda}_{NA}(t) - \Lambda(t)\right\} = 0$.

PROOF We start by writing difference from the definition of both of the functions, then rewrite the difference of integrals as a sum over arbitrarily small intervals, then introduce a binomial distribution for the cumulative hazard (?) before finally taking an expectation.

Through the Steiljes approximation, we can rewrite the difference as a sum. Suppose we split $[0, t]$ into m intervals, $0 = t_0 < t_1 < ... < t_m = t$, by the fact that $Y_U(k)$ is left continuous:

$$\left\{\widehat{\Lambda}_{NA}(t) - \Lambda(t)\right\} = \int_0^t \frac{1}{Y_U(k)} dN_U^{(1)}(k) - \lambda(k)dk$$

$$= \int_0^t \frac{dN_U^{(1)}(k) - \lambda(k)Y_U(k)dk}{Y_U(k)}$$

$$\approx \sum_{j=0}^{m-1} \frac{\Delta N_U^{(1)}(t_j) - \lambda(t_j)Y_U(t_j)\Delta(t_j)}{Y_U(t_j)}$$

Where $\Delta N_U^{(1)}(t_j)$ represents the number of deaths in $(t_j, t_{j+1}]$, and $\Delta(t_j) = t_{j+1} - t_j$

Pivoting, let $\mathcal{F}_j$ represent all the information (number of people who died, who are censored, who are still at risk) up till time t_j. $\{\mathcal{F}_j\}_{j=1}^m$ is called a *filtration*. More specifically, this filtration is defined as:

$$\mathcal{F}_j = \{(u_i, \Delta_i) : u_i \leq t_j\}$$

From this definition, $Y_U(t_j^+) \mid \mathcal{F}_j$ is well known, and is a constant. This is because $Y_U(t_j^+)$ represents the number of people still alive after time t_j, so with a fixed sample size n and with the knowledge of how many people died from $\mathcal{F}_j$, $Y_U(t_j^+) \mid \mathcal{F}_j$ is just the number of people still alive right after time t_j

We then see:
$$\Delta N_U^{(1)}(t_j) \mid \mathcal{F}_j \sim \mathsf{Bin}(Y_U(t_j^+), \pi_j)$$

where

$$\pi_j = P(t_j < U \leq t_{j+1}, \Delta = 1 \mid U > t_j)$$
$$= P(t_j < U \leq t_j + (t_{j+1} - t_j), \Delta = 1 \mid U > t_j)$$
$$= P(t_j < U \leq t_j + \Delta(t_j), \Delta = 1 \mid U > t_j)$$
$$= \frac{P(t_j < U \leq t_j + \Delta(t_j), \Delta = 1 \mid U > t_j)}{\Delta(t_j)}\Delta(t_j)$$
$$\approx \lambda_U^{(1)}(t_j)\Delta(t_j) \text{ for small } \Delta, \text{ and by } Definition\ 2.2.1$$
$$= \lambda^{(1)}(t_j)\Delta(t_j) \text{ by Theorem 2.2.2}$$

As the number of people who fail in $(t_j, t_{j+1}]$ (aka $\Delta N_U^{(1)}(t_j)$) is a random variable

based on the number of people left, $Y_U(t_j^+)$, and their probability for failing, π_j.
Now back to the regularly scheduled proof:

$$\left\{\widehat{\Lambda}_{NA}(t) - \Lambda(t)\right\} \approx \sum_{j=0}^{m-1} \frac{\Delta N_U^{(1)}(t_j) - \lambda^{(1)}(t_j)Y_U(t_j)\Delta(t_j)}{Y_U(t_j)}$$

$$\approx \sum_{j=0}^{m-1} \frac{\Delta N_U^{(1)}(t_j) - \pi_j Y_U(t_j^+)}{Y_U(t_j)}$$

$$:= \sum_{j=0}^{m-1} \frac{\Delta M_U(t_j)}{Y_U(t_j)}$$

Where for convenience we let: $\Delta M_U(t_j) := \Delta N_U(t_j) - \pi_j Y_U(t_j^+)$
Finally, we can show the expected difference is 0:

$$\mathbb{E}\left\{\widehat{\Lambda}_{NA}(t) - \Lambda(t)\right\} = \mathbb{E}\left\{\sum_{j=0}^{m-1} \frac{\Delta M_U(t_j)}{Y_U(t_j)}\right\} \text{ by above}$$

$$= \sum_{j=0}^{m-1} \mathbb{E}\left\{\mathbb{E}\left[\frac{\Delta M_U(t_j)}{Y_U(t_j)} \mid \mathcal{F}_j\right]\right\} \text{ by law of total E}$$

$$= \sum_{j=0}^{m-1} \mathbb{E}\left\{\frac{\mathbb{E}\left[\Delta N_U^{(1)}(t_j) \mid \mathcal{F}_j\right] - \pi_j Y_U(t_j^+)}{Y_U(t_j)}\right\} \text{ is } \pi_j \text{ constant wrt } \mathcal{F}_j?$$

$$= 0$$

$\square$

Counting and Martingale Processes

Counting Processes

Defn 4.5.1 | ("General" Counting Process) Non-mathematically, a **counting process** is a stochastic process with sample paths that are right continuous with left-hand limits (cad lag), and are non-decreasing step functions taking jumps of size 1.
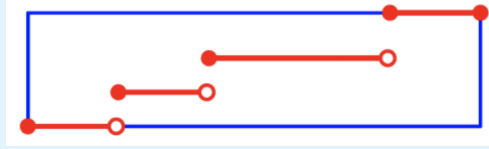

Figure 3. Example Counting Process, each jump is of size 1

Typically denoted $N(u) = \sum_{i=1}^{n} \mathbb{I}\{u_i \leq u\}$ as it counts the number of observations observed up to time t.

Defn 4.5.2 | (Information) We let $\mathcal{F}(u)$ notate all the information from the data up to point u, including the information at point u. Formally, this is called a sub-σ-algebra and can be defined as:

$$\mathcal{F}(u) = \sigma\{(u_i, \Delta_i) : u_i \leq u\}$$

The family of $\mathcal{F}(u)$ is called a *filtration*, and is *adapted* to a counting process $N(u)$ if $N(u) \mid \mathcal{F}(u)$ is known.

Defn 4.5.3 | (Intensity Process) The **intensity process** of a counting process $N(u)$ is defined as a stochastic process $A(u)$:

$$A(u) := \lim_{h \to 0} \left\{ \frac{\mathbb{E}\left[N(u) - N(u-h) \mid \mathcal{F}(u-h)\right]}{h} \right\}$$

Overall this definition describes the set up where we know the information up to right now $(u - h)$, and asks how many people we expect to see in the next instant (u).

 Usually we have the assumption that the probability we see more than one observation in the next instant ("in the next instant" being described as the time range $(u - h, u]$) is negligible. Thus we arrive at a redefinition of an intensity processes under this assumption

Lem 4.5.1 | Suppose the probability of observing more than 1 event in the next instant is negligible. In that case, the intensity process is just the probability of observing a new count given what you've seen already. Mathematically, this is expressed as:

$$\underbrace{P\{[N(u) - N(u-h) \geq 2] \mid \mathcal{F}(u-h)\} = O(h^2)}_{\text{negligible probability of more than 1 count}} \implies$$

$$A(u) = \lim_{h \to 0} \left\{ \frac{P[N(u) - N(u-h) = 1 \mid \mathcal{F}(u-h)]}{h} \right\}$$

PROOF | This simply follows from rewriting the expectation:

$$\mathbb{E}\{N(u) - N(u-h) \mid \mathcal{F}(u-h)\} = \underbrace{0 \cdot P(N(u) - N(u-h) = 0 \mid \mathcal{F}(u-h))}_{=0} +$$

$$1 \cdot P(N(u) - N(u-h) = 1 \mid \mathcal{F}(u-h)) +$$

$$\underbrace{2 \cdot P(N(u) - N(u-h) = 2 \mid \mathcal{F}(u-h)) + \ldots}_{=O(h^2) \text{ by assumption}}$$

$$= 1 \cdot P(N(u) - N(u-h) = 1 \mid \mathcal{F}(u-h))$$

Then we substitute back into the original $A(u)$ above to get the desired results. □

We also can also rewrite the relationship between the intensity process and the expectation using calculus notation:

Lem 4.5.2 | If we let $df(u)$ represent the change $f(u) - f(u-h)$ for a limitingly small h, then we get that

$$\mathbb{E}\{dN(u) \mid \mathcal{F}(u^-)\} = A(u)du$$

PROOF | Note that $du = u - (u-h) = h$, and $dN(u) = N(u) - N(u-h)$. This gets us:

$$A(u) = \lim_{h \to 0} \left\{ \frac{\mathbb{E}[N(u) - N(u-h) \mid \mathcal{F}(u-h)]}{h} \right\}$$

$$= \frac{\mathbb{E}\{dN(u) \mid \mathcal{F}(u^-)\}}{du}$$

Where the substitution was made from the note, and the $\mathcal{F}(u^-)$ just represents the limit from the left of $\mathcal{F}(u)$. □

We now move to a larger result about the distribution of the number of number of events that happens in du, which in the limit gives us a distribution of the instantaneous failure. The following theorem clarifies.

Theorem 4.5.5 | (Distribution of Number of Failures) If we assume individuals to be independent then the distribution of the number of people who fail on $(u-h, u]$ given how many people we know who have already failed is binomial, where the number of trials is the number of people who remain (the risk set size) at the start of the interval, times the probability that a person fails on that interval given that they've at least survived to the start of the interval. Mathematically:

$$dN(u) \mid \mathcal{F}(u^-) = \{N(u) - N(u-h)\} \mid \mathcal{F}(u-h) \sim \mathsf{Bin}(Y_U((u-h)^+), \pi_{h,U}(u))$$

- Where $Y_U((u-h)^+) = \sum_{i=1}^n \mathbb{I}\{u_i \geq (u-h)^+\}$ represents the number of people who are left and still at risk of failure. It is equivalent to $n - N(u-h)$.

- Where $\pi_{h,U}(u) = P(u - h < U_i \leq u, \Delta_i = 1 \mid U_i > u - h)$ represents the probability of a given individual, who has survived to at least time $u - h$ to fail in the next instant (du).

There really isn't a proof of this result, its mostly a result of observing that each person can either fail or not in the next du time $((u, u - h])$ so it is inherently bernoulli with the probability of failure being the probability of failing in that interval.

    For the independence of failure times, and the above results, we can simplify and rewrite the intensity process.

Lem 4.5.3 | From the independence of failure times between individuals, we get that:

$$A(u) = \lambda_U^{(1)}(u)Y(u)$$

PROOF | First note that $\mathbb{E}\left\{N(u) - N(u - h) \mid \mathcal{F}(u - h)\right\} = \pi_{h,U}(u)Y\{(u - h)^+\}$ as a consequence of it being binomial distributed by Theorem 4.5.5.

$$
\begin{aligned}
A(u) &= \lim_{h \to 0} \left\{ \frac{\mathbb{E}\left[N(u) - N(u - h) \mid \mathcal{F}(u - h)\right]}{h} \right\} \\
&= \lim_{h \to 0} \frac{\pi_{h,U}(u)Y\{(u - h)^+\}}{h} \quad \text{by note} \\
&= \left\{ \lim_{h \to 0} \frac{\pi_{h,U}(u)}{h} \right\} \cdot \left\{ \lim_{h \to 0} Y\left[(u - h)^+\right] \right\} \\
&= \left\{ \lim_{h \to 0} \frac{P(u - h < U_i \leq u, \Delta_i = 1 \mid U_i > u - h)}{h} \right\} Y(u) \quad \text{by defn of } \pi_{h,U}, \text{ and lim} \\
&= \lambda_U^{(1)}(u)Y(u) \quad \text{by Definition 2.2.1}
\end{aligned}
$$

$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

Appendix A: Integrals

This isn't mathematically rigorous, and some definitions are flat out *wrong*. The purpose of this is to get some kind of intuitive grasp of the simple cases of how these things work

SUBSECTION 5.1

Reimann to Lebesgue Integrals

All of the following "theory" is done very loosely. In reality these things are more complicated and require a lot more math. A consequence of doing survival analysis is that we know our data lives in $\mathbb{R}$, so that's all I will be considering.

Intuitively, the "dx" is a little width of the input space (in this case, some x) of your function f. So to get a rectangle you multiply that little width by the height of the function somewhere in that little width. You then add up those integrals and you get the area under the curve. This formulation provides us a volume (or area in a 2d case, and purely in a geometric sense) of how much "stuff" lies underneath a curve.

To Lebesgue integration: why do we need it? We need a generalization of volume. Maybe that isn't terribly motivating, but it provides a good intuition as to *how* we can use it.

So how do we use the idea of volume in our world? Let start by letting $x \in \mathcal{F}$ being an *event* in our *event space*. In our world, we can define the volume of an event to be $p(x)$. ***PROBABILITY IS VOLUME!***. More precisely, probability is the volume of some event. This allows us to notate things more generally. It also gives us things like countable additivity pretty easy. Say you have two disjoint pieces, wouldn't you expect the addition of their volume to be equal to the volume of those pieces together? Maybe that intuition is only useful for me.

Maybe some illustrations using this idea will be clear. Usually, we denote the CDF of a distribution to be a capital $F_X(x)$