Everything is deception: seeking the minimum of illusion, keeping within the ordinary limitations, seeking the maximum. In the first case one cheats the Good, by trying to make it too easy for oneself to get it, and the Evil by imposing all too unfavorable conditions of warfare on it. In the second case one cheats the Good by keeping as aloof from it as possible, and the Evil by hoping to make it powerless through intensifying it to the utmost. —Franz Kafka
Last time

- Introduced parametric models commonly used in survival analysis; discussed their densities, hazards, survivor function, and CDFs; showed how to draw from these distributions using R
  - Exponential
  - Weibull
  - Gamma
  - Extreme value
  - Log-normal
  - Log-logistic
- We also discussed location-scale models and how to use the location-scale framework to incorporate covariate information
- We also discussed flexible models for the hazard function including piecewise constant and basis expansions
Last time: burning questions

- How to choose a distribution?
- How to estimate the parameters indexing a chosen distribution?
- How can we accommodate different types of censoring?
- How can I use R to do the foregoing estimation steps?
Warm-up

▶ Explain to your stat buddy
1. Hazard function
2. How a bathtub hazard might arise
3. How an increasing hazard might arise
4. How a decreasing hazard might arise

▶ True or false:
   ▶ (T/F) Minimum of independent Weibull r.v.’s are Weibull
   ▶ (T/F) Measles increases fecundity in goats
   ▶ (T/F) An exponential distribution would be good model for mortality in humans

▶ Who is generally credited with discovering maximum likelihood estimation?
Warm-up cont’d

- Some concepts and notation for today
  - Let \( a_1, \ldots, a_n \) be a sequence of constants then
    \[
    \prod_{i=1}^{n} a_i = a_1 \times a_2 \times \cdots \times a_n
    \]
  - Let \( f(x) \) denote a function from \( \mathbb{R}^p \) into \( \mathbb{R} \) then
    \[
    x^* = \arg \max_x f(x)
    \]
    satisfies \( f(x^*) \geq f(x) \) for all \( x \in \mathbb{R}^p \).
  - We use \( 1_{\text{Statement}} \) to denote the function that equals one if \( \text{Statement} \) is true and zero otherwise. Thus, \( 1_{t \leq 1} \) equals one if \( t \leq 1 \) and zero otherwise.
Observation schemes

- Why is survival analysis its own sub-field of statistics?
  - Abundance of important applications
  - Fundamental contributions to statistical theory (esp. in semi-parametrics)
  - Dealing partial information due to censoring
- Recall that when we only observe partial information about a failure time we say that it’s censored
  - $T \geq C$ (Right censored)
  - $T \leq L$ (Left censored)
  - $V \leq T < U$ (Interval censored)
If generative model is indexed by parameter $\theta$ then the likelihood is
\[ L(\theta) \propto P(\text{Data}; \theta), \]
which is viewed as a function of $\theta$ with the data being fixed.

The maximum likelihood estimator is
\[ \hat{\theta}_n = \arg \max_{\theta} L(\theta) \]

Warm-up: Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, then $\theta = (\mu, \sigma^2)$, derive $L(\theta)$, and $\hat{\theta}_n$. Check your answer with your stat buddy.
Ex. Let $T_1, \ldots, T_n$ be an iid draw from distn with density $f(t; \theta)$ indexed by $\theta$ then

$$L(\theta) \propto P(\text{Data}; \theta) = \prod_{i=1}^{n} P(T_i = t_i; \theta) = \prod_{i=1}^{n} f(t_i; \theta),$$

Let $T$ denote a generic observation distd according to $f(t; \theta)$. How can we use $L(\theta)$ to estimate:

- The mean of $T$?
- The CDF of $T$?
- The hazard of $T$?
Nonparametric maximum likelihood

- Ex. Let $T_1, \ldots, T_n$ be an iid draw from distn with density $f(t)$. Suppose now, however, we don’t put any restrictions of $f(t)$ (other than it being a density). In this case $f$ is our ‘parameter’

$$L(f) \propto P(\text{Data}) = \prod_{i=1}^{n} f(t_i),$$

how can we maximize this over densities $f$?

- Claim: Our estimated $f$, say $\hat{f}$ should only put positive mass on $t_1, \ldots, t_n$. (Why?)

- If $\hat{f}$ puts mass on $t_1, \ldots, t_n$ then maximizing the likelihood is equivalent to solving

$$\max_{\alpha_1, \ldots, \alpha_n \geq 0} \prod_{i=1}^{n} \alpha_i \quad \text{subj. to} \quad \sum_{i=1}^{n} \alpha_i = 1$$

- Some painful calculus shows $\hat{f}$ is pmf with $f(t_i) = 1/n$, $i = 1, \ldots, n$
Thus $\hat{f}(t)$ is a discrete distribution with $f(t_i) = 1/n$. The estimated is given by

$$\hat{F}(t) = \frac{1}{n} \sum_{i=1}^{n} 1_{t \leq t_i},$$

this is called the empirical distribution function (ECDF)
Computing ECDF in R

```r
n = 50;
x = rnorm (n);
FHat = ecdf (x);
plot (FHat, xlab='x');
```
Observation schemes
Truncated estimation

- Suppose \( t_1, \ldots, t_n \) compose a random sample from subjects with lifetimes less than or equal to one year. The LH takes the form:

\[
\prod_{i=1}^{n} f(t_i | T_i \leq 1) = \prod_{i=1}^{n} \left\{ \frac{f(t_i)}{F(1)} \right\},
\]

Why?

- Suppose we posited a parametric model \( f(t; \theta) \) for \( T \), how can we estimate \( \theta \) using left-truncated data?

  - Deceptively difficult stat question: Suppose \( T_1, \ldots, T_n \) are drawn ind. from an \( \exp(\theta) \) distn but are truncated at one. When does the maximum likelihood estimator exist?
Right-censoring

- Recall that a censoring time is right censored at $C$ if we only observe that $T > C$

- Our goal in the next few slides is to derive the LH under different right-censoring mechanisms

- Notation: observe $\{(T_i, \delta_i)\}_{i=1}^n$ where $T_i$ is the observation time and $\delta_i$ is the censoring indicator

$$\delta_i = \begin{cases} 
1 & \text{Failure time observed} \\
0 & \text{Right censored}
\end{cases}$$

- Big result of the day: Under a variety of right-censoring mechanisms:

$$LH \propto \prod_{i=1}^n f(t_i)^{\delta_i} S(t_i^+)^{1-\delta_i}$$
Type I censoring

- In type I censoring each individual has a fixed (non-random) censoring time $C > 0$
  - If $T \leq C$ then failure time observed
  - If $T > C$ then right-censored

- Ex. Odense Malignant Melanoma Data: $n = 205$ subjects enrolled between 1962 and 1972 at Odense Dept. of Plastic Surgery had tumors and surrounding tissue removed. Patients were followed until the death or the study concluded in 1977.
  Note* This data is contained in the boot package in R.
Using the book’s notation: define \( t_i = \min(T_i, C_i) \) and \( \delta_i = 1_{T_i \leq C_i} \) then:

- If \( \delta_i = 1 \), \( t_i \) is the failure time so the information \((t_i, \delta_i)\) ‘contributes’ to the LH is \( f(t_i) \)
- If \( \delta_i = 1 \) then \( t_i \) is the censoring time so the information \((t_i, \delta_i)\) ‘contributes’ to the LH is \( S(t_i+) \)
- Thus, the LH is

\[
\prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i+)^{1-\delta_i}
\]
Type I censoring cont’d

In class: Suppose $T_1, \ldots, T_n$ are iid $\exp(\theta)$ but subject to Type I censoring, let $\delta_1, \ldots, \delta_n$ denote the censoring indicators. Derive the MLE for $\theta$. 
Independent random censoring

Assume lifetime $T$ and censoring time $C$ are random variables.
  - Often more realistic
  - Ex. Random study enrollment times
  - Ex. Subjects moving out of town
  - ...

Let $G(t)$ and $g(t)$ denote the survivor and density function for $C$ resp., define $t_i = \min(T_i, C_i)$, $\delta_i = 1_{T_i \leq C_i}$, then

$$f(t_i, \delta_i) = [f(t_i)G(t_i+)]^{\delta_i} [S(t_i+)g(t_i)]^{1-\delta_i}.$$ Why?
Independent random censoring cont’d

- The LH for \( n \) iid observations is

\[
\left( \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i+)^{1-\delta_i} \right) \left( \prod_{i=1}^{n} g(t_i)^{1-\delta} G(t_i+)^{\delta_i} \right),
\]

note that if \( g(t) \) and \( G(t) \) does contain information about \( f(t) \) or \( S(t) \) then the LH is proportional to

\[
\prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i+)^{1-\delta_i}
\]

It’s the same LH as before!!!
Type II censoring

- Observe individuals until the $r$th failure is observed, so that we observe the $r$ smallest lifetimes $t(1) \leq \cdots \leq t(r)$.
  - All $n$ units start at the same time
  - Follow-up stops at the time of the $r$th failure
  - Follow-up time is random
- Using properties of order statistics, the LH is

$$
\frac{n!}{(n-r)!} \left( \prod_{i=1}^{r} f(t(i)) \right) S(t(r)+)^{n-r} \propto \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i+)^{1-\delta_i}
$$
Code break

Go over mle.R in R Studio
For those who are adventurous
Counting process notation

- Goal: Show the form the LH for right-censoring applies in very general settings
- For clarity we assume discrete time \( t = 0, 1, \ldots \)
- Let \( h_i(t) \) and \( S_i(t) \) denote the hazard and survivor function for \( ith \) subject resp.; further define

\[
Y_i(t) \triangleq 1_{T_i \geq t, \text{ith subj not censored}} = \begin{cases} 
1 & \text{ith subj. hasn’t failed or been censored at } t \\
0 & \text{Otherwise} 
\end{cases}
\]

if \( Y_i(t) = 1 \) then we say the \( ith \) subj. is at risk at time \( t \)
Counting process notation cont’d

- Define

\[ dN_i(t) \triangleq Y_i(t)1_{T_i=t} \]

\[
= \begin{cases} 
1 & \text{if at risk and fails at } t \\
0 & \text{Otherwise}
\end{cases}
\]

\[ dC_i(t) \triangleq Y_i(t)1_{\text{ith subj. censored at } t} \]

\[
= \begin{cases} 
1 & \text{if at risk and censored at } t \\
0 & \text{Otherwise}
\end{cases}
\]

- Claim: \( \{dN_i(t), dC_i(t), t \geq 0\} \) has a single 1 and the rest zeros
Counting process notation cont’d

▶ Even more definitions:

\[ d\mathbb{N}(t) \triangleq (dN_1(t), \ldots, dN_n(t)) \]

\[ d\mathbb{C}(t) \triangleq (dC_1(t), \ldots, dC_n(t)) \]

\[ \mathcal{H}(t) \triangleq \{(d\mathbb{N}(s), d\mathbb{C}(s), s = 0, 1, \ldots, t - 1) \} \]

we say \( \mathcal{H}(t) \) is the history of the survival process up to time \( t \)
Counting process notation: the likelihood

Note that \( \lim_{t \to \infty} \mathcal{H}(t) \) contains all the information in the collected data (why?), thus

\[
P(\text{Data}) = P(d\mathbb{N}(0)) P(d\mathbb{C}(0)|d\mathbb{N}(0)) \\
\times P(d\mathbb{N}(1)|\mathcal{H}(1)) P(d\mathbb{C}(1)|d\mathbb{N}(1), \mathcal{H}(1)) \times \cdots \\
= \prod_{t=0}^{\infty} P(d\mathbb{N}(t)|\mathcal{H}(t))P(d\mathbb{C}(t)|d\mathbb{N}(t), \mathcal{H}(t))
\]
Counting process notation: the likelihood cont’d

- To make the horrible expression tractable we’ll assume conditional independence across subjects given $\mathcal{H}(t)$ and

$$P(dN_i(t) = 1|\mathcal{H}(t)) = Y_i(t)h_i(t),$$

explain this expression to your stat buddy

- We will also assume that terms inside $P(dC(t)|dN(t), \mathcal{H}(t))$ are not informative for the parameters in $h_i(t)$

- When the above assumptions hold we say the censoring is non-informative
Counting process notation: the likelihood cont’d

- Under the foregoing assumption the LH is given by

\[ \prod_{i=1}^{n} \prod_{t=0}^{\infty} h_i(t)^{dN_i(t)}(1 - h_i(t)) Y_i(t)(1 - dN_i(t)) \]

- To see this, we’ll consider two cases:
  - Case 1: *ith* subject’s failure time is observed at \( t_i \), then they’re at risk at \( t = 0, 1, \ldots, t_i \), and \( dN_i(t) = 1_{t=t_i} \), thus

\[ \prod_{t=0}^{\infty} h_i(t)^{dN_i(t)}(1 - h_i(t)) Y_i(t)(1 - dN_i(t)) = h_i(t_i) \prod_{s=0}^{t_i-1} (1 - h_i(s)) = f_i(t_i) \]

  - Case 2: *ith* subject is censored at time \( t_i \), then they’re at risk at times \( t = 0, 1, \ldots, t_i \), and \( dC_i(t) = 1_{t=s_i} \), thus

\[ \prod_{t=0}^{\infty} h_i(t)^{dN_i(t)}(1 - h_i(t)) Y_i(t)(1 - dN_i(t)) = \prod_{t=0}^{t_i} (1 - h_i(t)) = S_i(t_i+1) \]
Counting process notation: the likelihood cont’d

- Putting it all together shows the LH is proportional to

\[
\prod_{i=1}^{n} f_i(t_i) \delta_i S_i(t_i + 1)^{1-\delta_i}
\]

- Limiting arguments show the LH is the same (with \(S_i(t_i + 1)\) replaced by \(S_i(t_i +)\)) in the continuous case

- Note* we’ll see later that using the framework of partial likelihood that maximizing the above LH is appropriate in even more general settings
LH-based inference

- For parametric models the LH provides an efficient framework for estimation and inference.

- Let $\theta \in \Theta \subseteq \mathbb{R}^p$ index the survival distribution of interest, define:
  - $\mathcal{L}(\theta)$ the LH
  - $\ell(\theta) = \log \mathcal{L}(\theta)$ the log-LH
  - $u(\theta) = \frac{d}{d\theta} \ell(\theta)$ the score function
  - $I(\theta) = -\frac{d^2}{d\theta d\theta^\top} I(\theta)$ the Fisher information
Recall maximum LH estimator

\[ \hat{\theta}_n = \arg \max_{\theta \in \Theta} \mathcal{L}(\theta), \]

solves \( u(\theta) = 0 \)

Under mild regularity conditions

\[ \sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{\text{d}} N(0, I^{-1}(\theta^*)) \]
LH-based inference example

- Assume $T_1, \ldots, T_n$ are iid $\exp(\lambda)$ and subject to noninformative censoring. Let $\delta_1, \ldots, \delta_n$ denote the censoring indicator. Find the MLE for $\lambda$, the Fisher information matrix, and a 95% confidence interval for $\lambda$.
  - The LH, $L(\lambda)$ is

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

$$
= \lambda \sum_{i=1}^{n} \delta_i \exp \left\{-\lambda \sum_{i=1}^{n} t_i\right\}
$$

- The log-LH, $\ell(\lambda)$, is

$$
\left( \sum_{i=1}^{n} \delta_i \right) \log \lambda - \lambda \sum_{i=1}^{n} t_i
$$
LH-based inference example cont’d

The score function \( u(\lambda) \) is given by

\[
\begin{align*}
  u(\lambda) &= \frac{\sum_{i=1}^{n} \delta_i}{\lambda} - \sum_{i=1}^{n} t_i \\
  \end{align*}
\]

setting this to zero and solving yields

\[
\lambda_n = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} t_i}
\]
LH-based inference example cont’d

- We take the negative derivative of $u(\lambda)$ to get
  $$I(\lambda) = \sum_{i=1}^{n} \delta_i / \lambda^2$$
- How do we get a 95% confidence interval for $\lambda$?