I. Introduction

A. Analysis of data that are event times

1. Examples
   a. Medicine
      i. Deaths from certain kinds of disease
      ii. Recovery from disease
      iii. Time until disease progresses to a certain point
         • sometimes a surrogate endpoint
   b. Equipment failures
   c. Mortgage prepayments
   d. etc., etc., ... SAS Code R Code

2. Life data analysis differs from other forms of data analysis
   a. Additional complications arising from missing data
      i. Missingness is in the form of censoring
      ii. To be defined later this lecture.
   b. Different probabilistic models are appropriate
      i. For ex., wouldn’t be interested in modeling times as normal
      ii. Asymmetric (skewed right)
c. Different effects of interventions are usually expected
   i. For ex., horizontal shifts are usually not of interest
d. We are probably interested in estimating not just a mean but all quantiles of a distribution

   KM: 2.1

B. Describing life distributions

   KM: 2.2

1. Most elemental description of distributions is CDF
   a. Let $F(x)$ be the CDF.
      i. Defined to be $P[X \leq x]$.
   b. Most survival applications use instead the survival function
      $$S(x) = 1 - F(x) = P[X > x]$$
c. All distributions have these
   d. Properties:
      i. Non-increasing
         ii. $S(x) = 1 \forall x < 0$
         iii. $S(0) = 1$ if $P[X = 0] = 0$
            - For some applications, like reliability, $S(0) < 1$
            - We won’t see many such applications
iv.  $S(\infty) = 0$ typically, unless individual can last for ever.

2. Density of life times, if distribution is continuous

   a. Here $S'(x) = \int_x^\infty f(s) \, ds$

   b. $f(x) = -S''(x)$

   c. Some distributions are made up of smooth parts pasted together

      i. Relation $f(x) = -S''(x)$ doesn’t necessarily hold at junction

      KM: 2.3

3. The hazard rate:

   a. Definition: Chance of failing in the next small interval conditional on lasting until now, divided by length of interval.

   b. Assume distribution continuous

   c. $h(x) = \lim_{\Delta \to 0} P [X < x + \Delta | X \geq x] / \Delta$

   d. Quantity inside limit is $h(x) \approx f(x) \Delta / [S(x) \Delta] \rightarrow$

      $f(x) / S(x) = -S''(x) / S(x) = -\frac{d}{dx} \log(S(x))$ if $S''(x)$ exist

   e. $h(x) \geq 0 \forall x$

   f. Recover $S(x) = \exp(-\int_0^x h(s) \, ds)$,

      i. by noting that $\int_0^x h(s) \, ds = -\log(S(x))$.

      ii. $H(x) = \int_0^x h(s) \, ds$ is called integrated hazard.

   g. Discrete case: suppose possible event times are $x_j$
Lecture 1

i. counts of days, months, etc.

ii. Hazard is $h(x_j) = p(x_j)/S(x_{j-1})$

   $p(x_j) = P[X = x_j] = S(x_{j-1}) - S(x_j)$.

h. Some possible hazard shapes:

i. Constant: Same “risk” of event at every time point:
   - Many electronics components
   - Event of one radioactive atom decaying

ii. Increasing

iii. Decreasing

iv. Bathtub: Highest at beginning and late
   - Human life spans
   - products under warranty

v. Bump shaped:
   - Mortgages
   - Not many others

vi. Will be used later to give model diagnostics

   KM: 2.4

4. The mean residual life:

   a. Consider distribution with finite expectation.
b. Expectation of how much life remains, conditional on no event so far.

c. For continuous distributions, is \( \text{m.r.l.}(x) = \frac{\int_x^\infty (s - x) f(s) \, ds}{S(x)} \)

d. Integrate by parts: 
\[
\int_x^\infty (s - x) f(s) \, ds = (s - x)S(s)\bigg|_x^\infty + \int_x^\infty S(s) \, ds = \int_x^\infty S(s) \, ds
\]

i. \( \int_a^b u \, dv = uv\bigg|_a^b - \int_a^b v \, du \).

ii. \( u(s) = (s - x), \, dv = f(s) \, ds \) and so \( du = ds \), \( v(s) = -S(s) \).

iii. since \( \lim_{s \to \infty} sS(s) = 0 \); otherwise the expectation is infinite.

e. Hence \( \text{m.r.l.}(x) = \frac{\int_x^\infty S(s) \, ds}{S(x)} \)

f. When calculating \( \text{m.r.l.}'(x) \), note \( \frac{d}{dx} \int_x^\infty S(s) \, ds = -S'(x) \)

i. Because
\[
\frac{d}{dx} \int_{x}^{\infty} S(s)ds = \lim_{\delta \to 0} \left[ \int_{x+\delta}^{\infty} S(s)ds - \int_{x}^{\infty} S(s)ds \right] / \delta \\
= \lim_{\delta \to 0} \left[ - \int_{x}^{x+\delta} S(s)ds \right] / \delta \\
= \lim_{\delta \to 0} \left[ -S(x^*)\delta \right] / \delta \text{ for } x^* \in [x, x + \delta] \\
= \lim_{\delta \to 0} \left[ -S(x^*) \right] \text{ for } x^* \in [x, x + \delta] \\
= -S(x) \text{ since } S \text{ continuous}
\]

\( g. \) m.r.l. (0) is original mean

\( h. \) You can give formulae for survival function, density, and hazard

in terms of mean residual life.

\( \text{KM: 2.5} \)

C. Parametric Models for Life Distributions

1. Normal

   a. lifetime is sum of large number of independent contributions

   b. Bad one: negative values illegal

2. What distribution has constant hazard?

   a. Exponential

   i. \[-\frac{d}{dx} \log(S(x)) = \lambda \]

   ii. \[\log(S(x)) - \log(S(0)) = -\lambda x\]

   iii. \[\log(S(x)) = -\lambda x\]
Lecture 1

iv. \( S(x) = \exp(-\lambda x) \)

v. \( f(x) = \lambda \exp(-\lambda x) \)

vi. Distribution of remaining life is always the same: Memoryless

- Take \( s > x \).
- Note joint probability \( P [X > s \text{ and } X > t] = P [X > s] \)
- \( P [X > s | X > x] = \frac{\exp(-\lambda s)}{\exp(-\lambda x)} = \exp(- (s - x) \lambda) = P [X > s - x] \)

3. Generalization of Exponential

a. Sum of \( k \) independent exponentials with same (hazard) rate \( \lambda \) is \( \Gamma(\lambda, k) \).

b. Density is \( \lambda^k x^{k-1} \exp(-\lambda x) / \Gamma(k) \)

c. Survival function given by incomplete gamma function

d. \( h(x) = \lambda + \frac{1 - k}{x} \).

i. See Fig. 1.


a. Suppose that \( Y \) has an exponential distribution with rate \( \lambda \).

b. Let \( X = Y^{1/\alpha} \) for some \( \alpha > 0 \).

c. Survival function for \( X \) is \( S(x) = \exp(-\lambda x^\alpha) \)

4. Density \( \exp(-\lambda x^\alpha) \alpha \lambda x^{\alpha-1} \)
e. Hazard Function $\alpha \lambda x^{\alpha - 1}$
   
i. Increasing if $\alpha > 1$ and decreasing if $\alpha < 1$. See Fig. 2.

5. Log Normal
   
a. Maybe $T$ is product of large number of independent contributions
   
b. Log is normal
   
c. Gives the log normal distribution
   
d. $S(x) = 1 - \Phi((\log(x) - \mu)/\sigma) = 1 - \Phi(\log(x/\exp(\mu))/\sigma)$
      some $\mu, \sigma$
Hence $\mu$ is equivalent to a scale parameter.

e. $f(x) = \phi((\log(x) - \mu)/\sigma)/(\sigma x)$

f. $\phi(x) = \exp(-x^2/2)/\sqrt{2\pi}$ the standard normal density,

$$\Phi(x) = \int_{-\infty}^{x} \exp(-y^2/2) \, dy/\sqrt{2\pi}$$

the standard normal CDF.

g. MRL is pretty complicated.

h. Hazard is no simpler. See Fig. 3.

6. Generalization of Gamma:

a. $f(x) = \alpha \lambda^k x^{\alpha k - 1} \exp(-\lambda x^\alpha)/\Gamma(k)$
b. Integral $S(x)$ involves incomplete gamma function

c. Hazard is complicated.

d. Contains Weibull ($k = 1$) and log normal as special cases.

i. See Fig. 4.

7. **Pareto distribution:**

a. $h(x) = \theta/x$ for $x > \lambda$: decreasing.

b. \[
\frac{S(x)}{S(\lambda)} = \exp\left(-\int_\lambda^x \frac{\theta}{s} ds\right) = \exp(\theta \log(\lambda) - \theta \log(x)) = \frac{\lambda^\theta}{x^\theta}
\]

c. $f(x) = - \frac{d}{dx} S(x) = \theta x^{-\theta-1} \lambda^\theta$
II. Censoring and Truncation

A. Types of censoring

1. Right censoring

   a. know that a realization of $X$ exceeds some value, rather than
b. Observe $\min(X, C)$ and indicator for $X \geq C$.

c. Probabilistic structure relevant

i. If censoring mechanism has nothing to do with event you are trying to study, censored events give no additional information
   • For ex., life of a car before theft censored because of a serious accident doesn’t tell you anything

ii. If censoring mechanism is related to point in life, you know more than just that life exceeds some value.
   • For ex., scrapping a car because of poor condition might tell you that no one would bother to steal it.

d. Taxonomy

i. *Type I censoring*: $C$ fixed and known
   • Ex., medical study designed to follow people for a year censors them after a year
   • Censoring times might not all be the same
   • Might be the time between a enrollment and the fixed end of a study. See Fig. 5.
   • Makes *time on study* more relevant than *calendar time*. See
Type II censoring: Study proceeds until \( r < n \) events.

- \( C_j = X_{(r)} \)

**Random censoring:**

- For each \( X_i \) associate a censoring time \( C_i \)
- easiest is when censoring is \( \perp \) mechanism under study

iv. May have a mixture of these mechanisms.
v. Display on a *Lexis diagram*: Time on study by calendar time. See Fig. 7.

2. *Left censoring*:
   a. Knowledge that failure time is less than some value replaces knowledge of exact time
   b. Ex., disease onset age for those who have disease at first examination
   c. Ex., disease onset age for those who forgot when it came on
      KM: 3.3

3. More exotic censoring mechanisms
   a. *double censoring* if either may happen
b. *interval censoring* if your information is an interval.
   i. Ex., if a person is periodically screened for a disease.

   KM: 3.4

4. *Truncation:*
   a. Certain subjects omitted from data set.
   b. *left truncation:*
      i. result of delayed entry
      ii. Those who have event before start are not recorded
   c. *right truncation: *Those who haven’t had event are not recorded
      i. Ex.: data from death records.