Lecture two: Population Distributions and Non-parametric Estimations of Survival Function

1. Survival function and hazard function

- **Cumulative Function and Survival Function**

The actual survival time of an individual, *t*, can be regarded as the realization of a random variable *T*, which can take any non-negative value.

  - The distribution of the population survival is often unknown
  - Any theoretical distribution to approximate the true one?
  - Only few theoretical distributions available (e.g. exponential, Weibull, etc)
  - Empirical distribution that is solely based on data (non-parametric)

Definition of Survival Distribution:

$S(t) = pr(T > t) = 1 – pr(T <= t) = 1 – F(t)$

  - F(t): cumulative frequency distribution (of failure)
  - S(t): proportion of survivors
  - 0<= S(t) <= 1; non-increasing, right continuous and has left limits; S(0) = 1, $S(+\infty) = 0$.
  - F(t) or S(t) does not tell directly the failure rate at time t and failure rate at time t given survival

- **Density Function:**

$f(t) = dF(t) / dt$

  - Interpretation: Relative frequency distribution or failure rate
  - Limiting distribution of relative frequency when segment window has arbitrarily small length
  - $f(t) \geq 0$; integration of f(t) equals to one
  - Density function does not tell directly instantaneous risk or rate of failure GIVEN at risk

- **Hazard Function:**

$h(t) = \lim_{\delta \to 0} Pr(\text{failure in } (t, t + \delta] \mid \text{survival at } t)/\delta$

  - Instantaneous/conditional failure rate
  - Age-specific failure rate
  - Force of mortality
  - Non-negative; unbounded from above
  - Example: h(t) = constant
  - Cumulative Hazard: H(t)

- **Relationships between S(t), F(t), h(t) and H(t)**:
- **Example**: $h(t) = \lambda$ (i.e. exponential distribution)

- **Mean Residual Life time (mrl):**

  $mrl(t_0) = E[T - t_0 \mid T \geq t_0]$,

i.e.,$mrl(t_0)$ = average remaining survival time given the population has survived beyond t_0. It can be shown that

$$mrl(t_0) = \frac{\int_{t_0}^{\infty} (t - t_0) f(t) dt}{S(t_0)} = \frac{\int_{t_0}^{\infty} S(t) dt}{S(t_0)},$$

$mrl(0) = E(T)$. For exponential distribution, $mrl(t_0) = E(T)$.

2. Estimating the Survivor Function – nonparametric approach

- **Assumptions**

  - Observation on any one individual is independent of those on others
  - Random/independent censoring
  - Statistical implication of random censoring: censoring time and true survival time are independent conditioning on survival history

- **Empirical survivor function**

  In the absence of censoring

- **Instantaneous risk of failure (or conditional failure rate) between two arbitrary time points: (t_{j-1}, t_j]**

$q(t_{j-1},t_j)$ = Pr{failure in (t_{j-1}, t_j] given survival at t_{j-1}} = $(S(t_{j-1}) – S(t_j))/ S(t_{j-1})$

  - $0 <= q(t) <= 1$
  - Sample estimate:
  $\hat{q}(t_{j-1}, t_j) =$ {# of deaths in (t_{j-1}, t_j]}/{# of survivors at t_{j-1}}
  or
  $\hat{q}(t_{j-1}, t_j) =$ {# of deaths in (t_{j-1}, t_j]}/{average # of survivors during (t_{j-1}, t_j]}

- **Life-table and Kaplan-Meier/Product-Limit Estimate of S(t)**

Suppose we want to estimate S(t), the population proportion surviving beyond time t.

- Survival history can be described by the conditional probabilities or instantaneous risk, q(s, t). Let's pick up a sequence of time points leading to t (divide and conquer?):

$$t_0 = 0 < t_1 < \ldots < t_{j-1} < t$$

$$\begin{aligned} S(t) &= \Pr(T > t) = \Pr(T > t_{j-1}) * \Pr(T > t \text{ Given } T > t_{j-1}) \\ &= \Pr(\text{survive beyond } t_{j-1})*\Pr(\text{No Failure in } (t_{j-1}, t] \mid T > t_{j-1}) \\ &= S(t_{j-1})*(1 - q(t_{j-1}, t)) \\ &= S(t_{j-2})*(1-q(t_{j-2}, t_{j-1}))(1- q(t_{j-1}, t)) \\ &= (1-q(t_0, t_1))*\ldots*(1- q(t_{j-2}, t_{j-1}))(1- q(t_{j-1}, t)) \end{aligned}$$

- **Life-table (or actuarial) estimate**:
  - Dividing the period of observation into a series of time intervals: t'_j to t'_{j+1}, j = 1, 2, …,m
  - d_j deaths, c_j censored in $(t'_j, t'_{j+1}]$ and n_j at risk at the start of the j'th interval
  - Assume censored times occur uniformly (i.e. $U(0, c_j)$) through the j'th interval, then average number of individual at risk is $n'_j = n_j - c_j / 2$
  - The probability of survival beyond time t'_k, k = 1, 2, …,m is

$$S(t) = \prod_{j=1}^{k} (n'_j - d_j) / n'_j$$

for $t'_k \le t < t'_{k+1}$, k = 1, 2, …,m

- **For KM estimate:**

Choose above sequence as the distinguish death times: Observed survival times: $t_1, t_2, \ldots, t_n$; death times: $t_{(1)} < t_{(2)} < \ldots < t_{(r)}$; n_j at risk just before $t_{(j)}$, d_j deaths at $t_{(j)}$

$$\hat{S}(t) = \prod_{j=1}^{k} (n_j - d_j) / n_j$$

for $t_{(k)} \le t < t_{(k+1)}$, k = 1,2, 3,…,r

  - The largest observation is censored? Undefined beyond that time; otherwise is zero (largest observation is event).
  - Censoring time and death time occur simultaneously? Assume censored time(s) occur(s) right after the death time(s).
- **Nelson-Aalen estimate:**

$$\tilde{S}(t) = \prod_{j=1}^{k} \exp(-d_j / n_j)$$

for $t_{(k)} \le t < t_{(k+1)}$, $k = 1,2,3,\ldots,r$.

- The above estimate can be derived from an estimate of the cumulative hazard function, using Taylor expansion of log (1-x).
- KM estimate can be regarded as approximation to the Nelson-Aalen estimate, using Taylor expansion of e^{-x}.
- The Nelson-Aalen estimate of survivor function ≥ KM estimate at any given time.
- Small-sample properties; and large-sample properties.

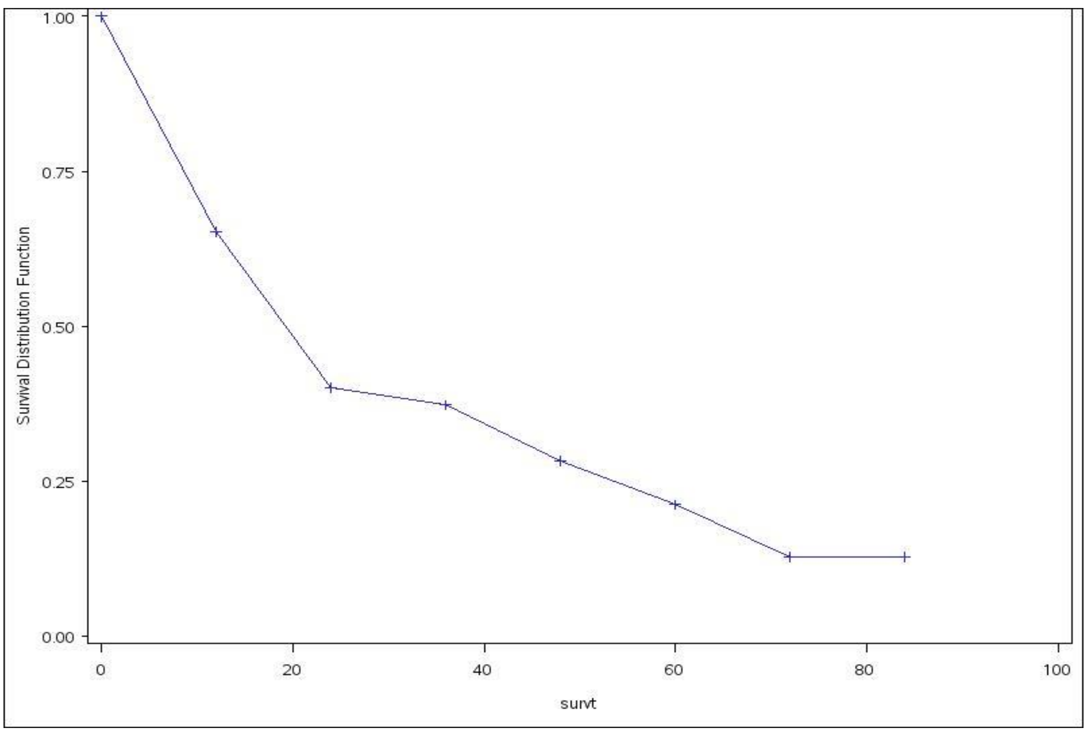
○ **Examples:**

Example 2.2 (p17):

Output:

Life Table Survival Estimates

Interval [Lower,	Upper)	Num. Failed	Num. Censored	Sample Size	Probability of Failure	Standard Error
0	12	16	4	46.0	0.3478	0.0702
12	24	10	4	26.0	0.3846	0.0954
24	36	1	0	14.0	0.0714	0.0688
36	48	3	1	12.5	0.2400	0.1208
48	60	2	2	8.0	0.2500	0.1531
60	.	4	1	4.5	0.8889	0.1481



SAS program:

```
Options ls = 80;
libname fu '../sdata';
data fu.myeloma;
infile "../data/myeloma.dat" ;
input pid survt censor age sex bun ca hb pc bj;
proc lifetest plots=(s) method =life width=12;
        time survt*censor(0);
run;
```

Example 2.3 (p23): KM estimate

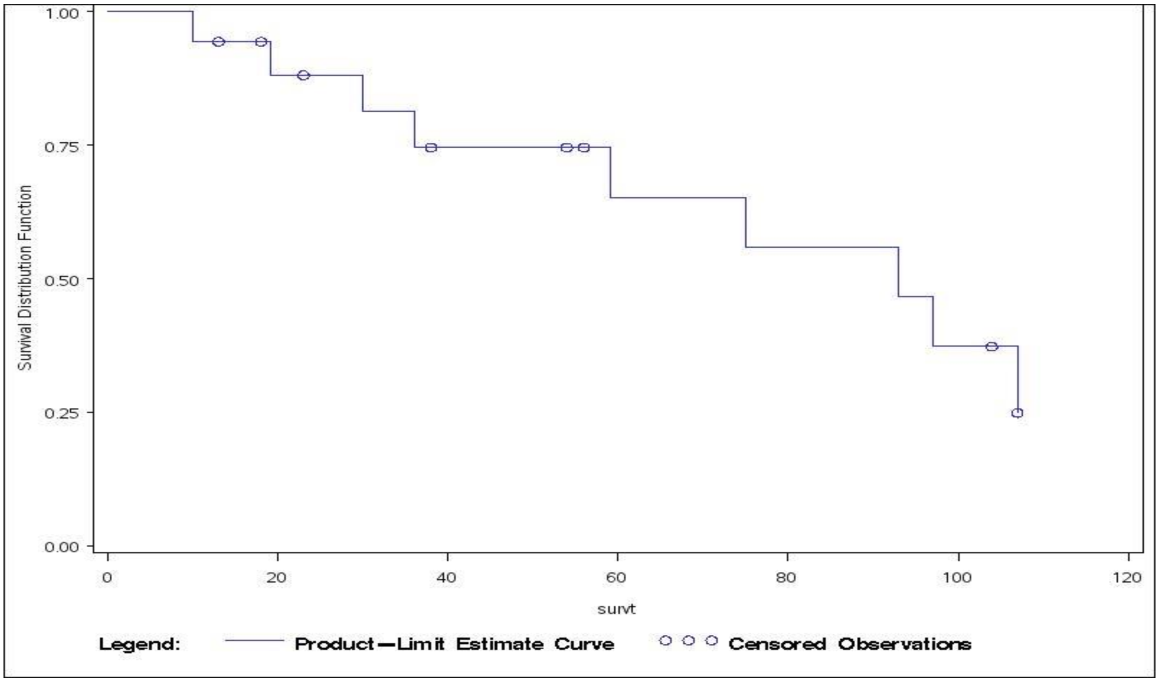
Output:

The LIFETEST Procedure

Product-Limit Survival Estimates

SURVT	Survival	Failure	Standard Error	Number Failed	Number Left
0.000	1.0000	0	0	0	18
10.000	0.9444	0.0556	0.0540	1	17
13.000*	.	.	.	1	16
18.000*	.	.	.	1	15
19.000	0.8815	0.1185	0.0790	2	14
23.000*	.	.	.	2	13
30.000	0.8137	0.1863	0.0978	3	12
36.000	0.7459	0.2541	0.1107	4	11
38.000*	.	.	.	4	10
54.000*	.	.	.	4	9
56.000*	.	.	.	4	8
59.000	0.6526	0.3474	0.1303	5	7
75.000	0.5594	0.4406	0.1412	6	6
93.000	0.4662	0.5338	0.1452	7	5
97.000	0.3729	0.6271	0.1430	8	4
104.000*	.	.	.	8	3
107.000	0.2486	0.7514	0.1392	9	2
107.000*	.	.	.	9	1
107.000*	.	.	.	9	0

NOTE: The marked survival times are censored observations.



SAS program:

```
* create sas dataset from ASCII file iud.dat;
Options ls = 80;
libname fu '../sdata';
data fu.iud;
infile "../data/iud.dat";
input survt censor;
* run lifetest procedure;
filename gsasfile 'iud.gsf';
goptions gaccess=gsasfile ROTATE=LANDSCAPE gsfmode=replace device=ps;
proc lifetest plots=(s) method =km;
        time survt*censor(0);
run;
```

Example 2.4 (p20): Nelson-Aalen estimate of survivor function - ***SAS program only:***

```
Options ls = 80;
libname fu '../sdata';
data w;
        set fu.iud;
proc lifetest method =PL NELSON;
        time survt*censor(0);
run;
```

Assignment two: Calculate Kaplan-Meier estimate of survivor function for chronic active hepatitis data set (Table B.1, p499) by hand for each treatment group (i.e. Prednisolone and Control); and verify your results using statistical software (e.g. SAS).

Reading assignment: read sections 1.5, 1.6.