Chapter 4 Nonparametric Methods: One-sample Problem

PubH 7450

Email: weip@biostat.umn.edu

Http: www.biostat.umn.edu/∼weip

§4.2 Estimate $S(t)$ for Right-Censored Data

- Given data (T_i, δ_i), $i = 1, ..., n$.
- Goal: to estimate $S(.)$ (and $H(.)$) for X_i.
  not $S(.)$ for T_i or C_i!
- Assumption: nonparametric; **independent censoring**: assume that X and C are independent $\Longrightarrow$ no difference between a subject censored at t and one surviving beyond t.–always assume this!
  Does it mean that every subject has an equal probability of being censored?
- Notation
  Define $t_1 < t_2 < ... < t_D$ as *distinct* event times;
  $d_i = \#$ events at t_i;
  $y_i = \#$ subjects at risk at $t_i^- = \#\{i : X_i \geq t_i\}$, called risk set size.

- Kaplan-Meier (K-M) estimator, Product-Limit (PL) estimator:

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_1; \\ \prod_{t_i \le t} \left(1 - \frac{d_i}{y_i}\right) & \text{otherwise.} \end{cases}$$

- Example 4.1: Table 1.1, 6-MP group, $n = 21$

  Data: 6, 6, 6, 6+, 7, 9+, 10, 10+, 11+, 13,...

  t_i: 6, 7, 10, 13, ...

  d_i: 3, 1, 1, 1,...

  y_i: 21, ?, ?, ?,...

  $t = 0$, $\hat{S}(t) = 1$;

  $t = t_1 = 6$, $\hat{S}(t) = 1 - d_1/y_1 = 1 - 3/21 = .857$;

  $t = t_2 = 7$, $\hat{S}(t) = .857(1 - 1/17) = .806$;

  $t = t_3 = 10$, $\hat{S}(t) = .806 * (1 - 1/15) = .752$;

  $t = t_4 = 13$, $\hat{S}(t) = .752(1 - 1/12) = .690$;

  ...

  Plot?

- Example 4.1: SAS; handout
- Justifications for $\hat{S}(t)$:
  - $\hat{S}(t) = S_e(t)$ if there is no censoring:

    Suppose no tied event times;

    $S(t_0) = 1;$

    $S(t_1) = 1 - \frac{1}{n};$

    $S(t_2) = \frac{n-1}{n}(1 - \frac{1}{n-1}) = 1 - \frac{2}{n};$

    $S(t_3) = \frac{n-2}{n}(1 - \frac{1}{n-2}) = 1 - \frac{3}{n};$

    ...
  - "Reduced-sample" argument:

    Because only observe events at t_i's, without any other assumption, $\hat{S}(t)$ should be a step function of t_i; think about $S_e(t)$.

$$\widehat{Pr}(X > t_i | X \geq t_i) = 1 - \widehat{Pr}(X = t_i | X \geq t_i) = ...$$

$$\begin{aligned}\hat{S}(t_i) &= \frac{\hat{S}(t_i)}{\hat{S}(t_{i-1})}\frac{\hat{S}(t_{i-1})}{\hat{S}(t_{i-2})}\ldots\frac{\hat{S}(t_1)}{\hat{S}(t_0)}\hat{S}(t_0) \\ &= Pr(X > t_i | X \geq t_i) Pr(X > t_{i-1} | X \geq t_{i-1}) \ldots 1 \\ &= \prod_{j \leq i}\left(1 - \frac{d_j}{y_j}\right).\end{aligned}$$

– Redistribution-to-the-right algorithm:

Example: $n = 6$; Probability mass:

T_i	3	4	5+	6	7+	8
Iter=1	1/6	1/6	1/6	1/6	1/6	1/6
Iter=2	1/6	1/6	0	1/6+1/18	1/6+1/18	1/6+1/18
				=4/18	4/18	4/18
Iter=3	1/6	1/6	0	4/18	0	4/18+4/18

$\tilde{S}(3) = 1 - 1/6;$

$\tilde{S}(4) = 1 - 1/6 - 1/6;$

$\tilde{S}(6) = 1 - 1/6 - 1/6 - 4/18$;
$\tilde{S}(8) = 1 - 1/6 - 1/6 - 4/18 - 8/18 = 0$.
Verify $\hat{S}(t) = \tilde{S}(t)$?

– Self-consistency:

$$\widehat{SC}(t) = \frac{1}{n}\left(\sum_{i=1}^{n} I(T_i > t) + \sum_{i=1}^{n} \frac{\widehat{SC}(t)}{\widehat{SC}(T_i)} I(T_i \le t, \delta_i = 0)\right)$$

2nd term $= \widehat{Pr}(X_i > t | X_i > T_i) = E[I(X_i > t | X_i > T_i)]$.
K-M estimator is the unique self-consistent estimator for $t < T_{(n)}$; for a proof, see Miller (1981).

– K-M estimator is the nonparametric MLE (NPMLE):
1) No censoring. The ECDF $\hat{F}$ is the NPMLE of F, hence $1 - \hat{F} = \hat{S}$ is the NPMLE of S.

Example: observe events at x_1, x_2.

Denote $p_i = Pr(X = x_i)$, $i = 1, 2$.

$L = p_1 p_2$ with $0 \le p_1, p_2, p_1 + p_2 \le 1$.

To maximize L, we need to have $p_1 + p_2 = 1$; otherwise, p_1 and/or p_2 will be smaller, and thus L will be smaller.

So, $L = p_1(1 - p_1) \Longrightarrow L' = 1 - 2p_1 = 0 \Longrightarrow \hat{p_1} = \hat{p_2} = 1/2$.

2) With censoring. Again the NPMLE would put all probability mass at observed event times t_i's and possibly $T_{(n)}$ if $\delta_{(n)} = 0$ and $T_{(n)} > t_D$.

$$
\begin{aligned}
L &= \prod_{i=1}^{n} Pr(X = T_i)^{\delta_i} Pr(X > T_i)^{1-\delta_i} \\
&= \prod_{i=1}^{n} [p_j I(T_i = t_j)]^{\delta_i} (\sum_{T_i > t_j} p_j)^{1-\delta_i}.
\end{aligned}
$$

$\Longrightarrow$ L is maximized by $\hat{S}(t)$; see Miller 1981 for a proof.

- Variance of $S\hat{(t)}$
  $\log \hat{S}(t) = \sum_{t_i \le t} \log\left(1 - \frac{d_i}{y_i}\right)$.
  $1 - \frac{d_i}{y_i} = 1 - p_i = q_i$, $d_i \sim Bin(y_i, E(p_i))$.
  Delta method: $Var(g(X)] \approx [g'(X)]^2 Var(X)$.
  $Var(\log q_i) \approx (1/q_i)^2 Var(q_i) \approx \frac{1}{q_i^2}\frac{q_i(1-q_i)}{y_i} = \frac{d_i}{y_i(y_i-d_i)}$.
  Treating all the terms as independent (incorrectly),
  $Var \log \hat{S}(t) \approx \frac{1}{\hat{S}(t)^2} Var[\hat{S}(t)] \approx \sum_{t_i \le t} \frac{d_i}{y_i(y_i-d_i)}$.
  $Var[\hat{S}(t)] \approx \hat{S}(t)^2 \sum_{t_i \le t} \frac{d_i}{y_i(y_i-d_i)}$, —Greenwood's formula.

- For any given $t_0 < t_D$,

$$\hat{S}(t_0) \overset{a.}{\sim} N(S(t_0), Var[\hat{S}(t)]).$$

- Example 4.1.

§4.2 Estimate $H(t)$ for Right-censored Data

- Based on K-M estimator:

$$\hat{H}(t) = ...$$

- Nelson-Aalen estimator of H:

$$\tilde{H}(t) = \begin{cases} 0 & \text{if } t < t_1; \\ \sum_{t_i \le t} \frac{d_i}{y_i} & \text{otherwise.} \end{cases}$$

Interpretation of d_i/y_i:
For discrete r.v.. $h(t_i) = Pr(X = t_i | X \ge t_i) \Longrightarrow ...$

- N-A estimator of S, also called Fleming-Harrington estimator:

$$\tilde{S}(t) = exp(-\tilde{H}(t)) = \prod_{t_i \le t} exp(-\frac{d_i}{y_i}).$$

Note: if d_i/y_i is small,

$\log \hat{S}(t) = \sum \log\left(1 - \frac{d_i}{y_i}\right) \approx \sum -\frac{d_i}{y_i} = -\tilde{H}(t).$

- Comparison:
  Fleming-Harrington compared the performance of $\hat{S}(t)$ and $\tilde{S}(t)$ empirically, finding that $\tilde{S}$ has smaller MSE when $S(x) \geq 0.2$, but larger MSE otherwise.
  In practice, use $\hat{S}(t)$ for $S(t)$; use $\tilde{H}(t)$ for $H(t)$.

- $Var(\tilde{H}(t))$:
  1) $d_i \sim Pois(E(d_i)) \Longrightarrow Var(\tilde{H}(t)) = \sum_{t_i \leq t} \frac{d_i}{y_i^2}$.
  2) $d_i \sim Bin(y_i, E(p_i)) \Longrightarrow Var(\tilde{H}(t)) = \sum_{t_i \leq t} \frac{y_i p_i (1-p_i)}{y_i^2} = \sum_{t_i \leq t} \frac{d_i (y_i - d_i)}{y_i^3}$.
  1) is preferred.

- Derive $Var(\tilde{S}(t))$: by $\tilde{S}(t) = exp(-\tilde{H}(t))$,
  $Var(\tilde{S}(t)) \approx (exp(-\tilde{H}(t)))^2 Var(\tilde{H}(t)) = \tilde{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{y_i^2}$, close to Greenwood's formula.

- Example 4.1b.r

§4.3 Point-wise CI for $S(t)$

- Recall $\hat{S}(t) \overset{a.}{\sim} N(S(t), Var(\hat{S}(t)))$
  $\Longrightarrow$ 95% Wald CI of $S(t)$ at any given t is
  $\hat{S}(t) \pm 1.96\sqrt{Var(\hat{S}(t))}$, –linear CI.
  Downsides:
  1) can be out of the range of [0,1];
  2) lower coverage probability: the true distribution of $\hat{S}(t)$ is typically skewed.
  $\Longrightarrow$ take some transformation!

- log transformation:
  $Var(\log(\hat{S}(t))) \approx \frac{1}{\hat{S}(t)^2} Var(\hat{S}(t))$, and 95% CI of $\log S(t)$ is

$$\log(\hat{S}(t)) \pm 1.96\sqrt{Var(\log(\hat{S}(t)))},$$

hence 95% CI of $S(t)$ is

$$\exp[\log(\hat{S}(t)) \pm 1.96\sqrt{Var(\log(\hat{S}(t)))}].$$

- log-log transformation:
  $\log(-\log\hat{S}(t))$.
  Use the Delta-method to estimate $Var[\log(-\log\hat{S}(t))]$, then

$$\exp\{\exp[\log(-\log\hat{S}(t)) \pm 1.96\sqrt{Var}]\}.$$

- Textbook gives another one based on arcsin-square root (Angular or Anscomb) transformation; p.105.
- Logit transformation
- Summary: Linear CI is not good for small samples, while log- or log-log-transformation is good enough, which works well even for $n \geq 25$ (with 50% censoring).
- Similarly, one can derive linear, log-, log-log-transformed CI for

$H(t)$ based on $\tilde{H}(t)$ and $Var(\tilde{H}(t))$; p.107.

- Example 4.1c.sas

§4.4 Confidence bands for $S(t)$

- Point-wise CI:
  valid only for a given point t_0.
  Suppose 95% $CI = [L(t_0), U(t_0)]$, we have

$$Pr\{L(t_0) \leq S(t_0) \leq U(t_0)\} = .95.$$

- Confidence bands: for t in some interval,
  $Pr\{L(t) \leq S(t) \leq U(t)\} = .95$ for all $t \in [t_L, t_U]$.

- Equal probability (EP) bands: proportional to point-wise CI.
  $Var(\hat{S}(t)) = \hat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{y_i(y_i - d_i)} = \hat{S}(t)^2 \sigma_S^2(t)$.
  $a_L = \frac{n\sigma_S^2(t_L)}{1+n\sigma_S^2(t_L)}$,
  $a_U = \frac{n\sigma_S^2(t_U)}{1+n\sigma_S^2(t_U)}$,
  n: sample size.
  Require $0 < a_L, a_U < 1$.
  Find a coefficient $c_\alpha(a_L, a_U)$ from Table C.3, an analog of 1.96

for N(0,1).

Linear:

$$\hat{S}(t) \pm c_\alpha(a_L, a_U)\sqrt{Var(\hat{S}(t))} = \hat{S}(t) \pm c_\alpha(a_L, a_U)\sigma_S(t)\hat{S}(t).$$

log-log transformed:

$$[\hat{S}(t)^{1/\theta}, \hat{S}(t)^{\theta}],\ \theta = \exp\left(\frac{c_\alpha(a_L, a_U)\sqrt{Var(\hat{S}(t))}}{\hat{S}(t)\log \hat{S}(t)}\right).$$

- Hall-Wellner bands

  1) not proportional to CI;

  2) allow $t_L = 0$.

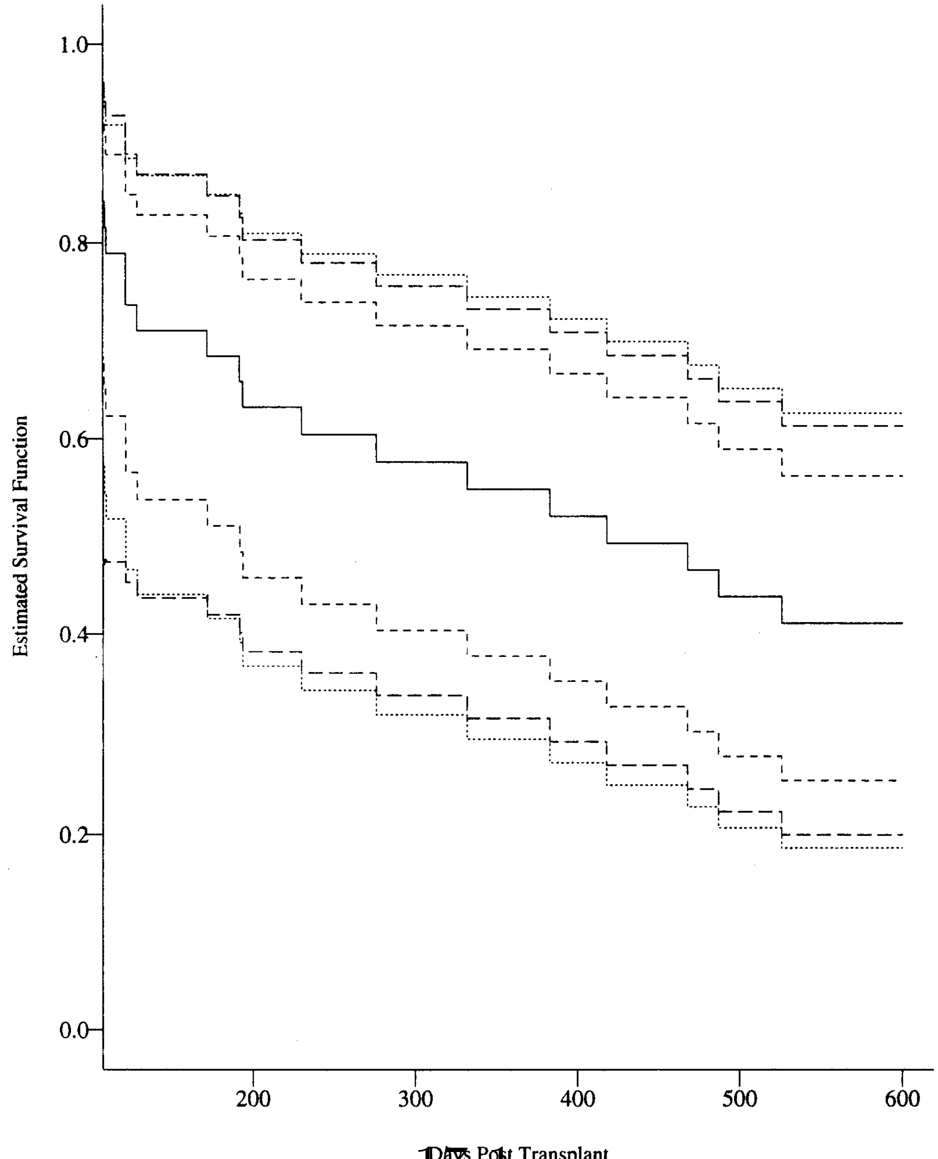
  Find coefficient $k_\alpha(a_L, a_U)$ from Table C.4.

  Linear:

$$\hat{S}(t) \pm \frac{k_\alpha(a_L, a_U)[1 + n\sigma_S^2(t)]}{\sqrt{n}}\hat{S}(t).$$

Log-log transformed: ...

- Fig 4.6.
- Similarly, construct confidence bands for $H(t)$; p.114-116.
- EP bounds: the linear one not good; log-transformed good even for $n \geq 20$.
- H-W bounds: both linear and log-transformed seem fine for $S(t)$; linear not good for $H(t)$; log-transformed good for $H(t)$.
- Example 4.1d.sas



17-1

Figure 4.6 *Comparison of 95% pointwise confidence interval, EP confidence band and Hall-Wellner confidence band for the disease free survival function found using the log transformation for ALL patients. Estimated Survival (————); Pointwise confidence interval (— — —); EP confidence band (------); Hall-Wellner band (———-——)*

§4.5 Estimates of the mean or median survival time

- mean/median: a good summary of $S(t)$.
- Recall $\mu = E(X) = \int_0^\infty S(t)dt$,
  $\Longrightarrow \hat{\mu} = \int_0^\infty \hat{S}(t)dt$.
  But $\hat{S}(t) =?$ for $t > T_{(n)}$ if the largest observed time $T_{(n)}$ is a censoring time (i.e. $\delta_{(n)} = 0$) —-reasonable?
  $\Longrightarrow$ can or cannot calculate $\int_0^\infty \hat{S}(t)dt$?
- Solution 1): define $\hat{S}(T_{(n)}) = 0$, —-Efron's tail correction.
  Another one by R Gill: $\hat{S}(t) = \hat{S}(T_{(n)})$ for $t > T_{(n)}$.
  Either one leads to biased or unbiased $\hat{\mu}$?
- Solution 2): estimate μ with the restriction that $t \in [0, \tau]$ for some $\tau \leq T_{(n)}$,
  $\hat{\mu}_\tau = \int_0^\tau \hat{S}(t)dt$.

$Var(\hat{\mu}_\tau) = \sum_{i=1}^{D} \left(\int_{t_i}^{\tau} \hat{S}(t) \right)^2 \frac{d_i}{y_i(y_i - d_i)}$.
$\hat{\mu}_\tau \overset{a.}{\sim} N(\mu_\tau, Var(\hat{\mu}_\tau))$.
$\Longrightarrow$ 95% CI: $\hat{\mu}_\tau \pm 1.96\sqrt{Var(\hat{\mu}_\tau)}$.

- Read example 4.1-4.2, p.118-119.

- Recall the pth quantile $x_p = inf\{t : S(t) \leq 1 - p\}$.
$\Longrightarrow \hat{x}_p = inf\{t : \hat{S}(t) \leq 1 - p\}$.
$x_{1/2}$ is the median.
But $Var(\hat{x}_p)$ is tough to estimate.

- $100(1 - \alpha)\%$ CI for x_p: all t satisfying

$$-z_{\alpha/2} \leq \frac{\hat{S}(t) - (1 - p)}{\sqrt{Var(\hat{S}(t))}} \leq z_{\alpha/2},$$

or

$$-z_{\alpha/2} \le \frac{\log[-\log \hat{S}(t)] - \log[-\log(1-p)]}{\hat{S}(t)\log \hat{S}(t)\sqrt{Var(\hat{S}(t))}} \le z_{\alpha/2}.$$

- Example 4.2, Table 4.7, p.121.
- ex4.1.sas

TABLE 4.7

Construction of a 95% Confidence Interval for the Median

t_i	$\hat{S}(t_i)$	$\sqrt{\hat{V}[\hat{S}(t_i)]}$	*Linear* (4.5.4)	*Log* (4.5.5)	*Arcsine* (4.5.6)
1	0.9737	0.0260	18.242	3.258	7.674
55	0.9474	0.0362	12.350	3.607	6.829
74	0.9211	0.0437	9.625	3.691	6.172
86	0.8947	0.0498	7.929	3.657	5.609
104	0.8684	0.0548	6.719	3.557	5.107
107	0.8421	0.0592	5.783	3.412	4.645
109	0.8158	0.0629	5.022	3.236	4.214
110	0.7895	0.0661	4.377	3.036	3.806
122	0.7368	0.0714	3.316	2.582	3.042
129	0.7105	0.0736	2.862	2.334	2.679
172	0.6842	0.0754	2.443	2.074	2.326
192	0.6579	0.0770	2.052	1.804	1.981
194	0.6316	0.0783	1.681	1.524	1.642
230	0.6041	0.0795	1.309	1.220	1.290
276	0.5767	0.0805	0.952	0.909	0.945
332	0.5492	0.0812	0.606	0.590	0.604
383	0.5217	0.0817	0.266	0.263	0.266
418	0.4943	0.0819	−0.070	−0.070	−0.070
468	0.4668	0.0818	−0.406	−0.411	−0.405
487	0.4394	0.0815	−0.744	−0.759	−0.741
526	0.4119	0.0809	−1.090	−1.114	−1.078
609	0.3825	0.0803	−1.464	−1.497	−1.437
662	0.3531	0.0793	−1.853	−1.886	−1.798
2081	0.3531	0.0793	−1.853	−1.886	−1.798

§4.6 left-truncated and right-censored data

- Given data: (L_i, T_i, δ_i), $i = 1, 2, ..., n$.
- Goal: to estimate $S(t)$ and $H(t)$ for X.
- Notation: as before,
  i) define $t_1 < t_2 < ..., t_D$ as ordered distinct event times;
  ii) $d_i = \#$ (events at t_i);
  iii) $y_i = \#$ (subjects at risk at t_i)=$\# \{j : L_j \leq t_i \leq T_j\}$; i.e. # of subjects who entered the study at/before t_i, and died at/after t_i.
- **All** the estimators discussed earlier for R-C'ed data are applicable here (with modified y_i). e.g.

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{y_i}\right).$$

- Note 1). Suppose $L = \min_i L_i$, then it's obvious that the data contain only information for those who can survive beyond L; that is, $\hat{S}(t)$ estimate

$$Pr(X > t|X \geq L) = S(t)/S(L).$$

If $L \approx 0 \Longrightarrow S(L) \approx 1$ and this $\hat{S}(t)$ is roughly estimate $S(t)$.

- Note 2). Truncation introduces difficulty in estimating $S(t)$ (or more precisely, $S(t)/S(L)$).
  e.g., if $d_1 = y_1 = 1 \Longrightarrow \hat{S}(t) = 0$ for **any** $t \geq t_1$!
  More generally, results may not be reliable if some early y_i's are small.

- Example 4.3; Figs 4.10-4.11.

- Channing House data: the male group; order the subjects by their L_i's:
  1st subject entered at month 751;
  2nd subject entered at month 759;

these two died at month 777 and 781, respectively;
3rd subject entered at month 782;
$\Longrightarrow t_1 = 777$, $d_1 = 1$, $y_1 = 2 \Longrightarrow \hat{S}(t_1) = 1/2$;
$\Longrightarrow t_2 = 781$, $d_2 = 1$, $y_2 = 1 \Longrightarrow \hat{S}(t_2) = 1/2 * (1 - 1/1) = 0$!

- How to fix?

- 1) To estimate $S_a(t) = S(t)/S(a) = Pr(X > t|X > a)$ for some large (but not so large) a (around which y_i's are reasonably large): for $t \geq a$,

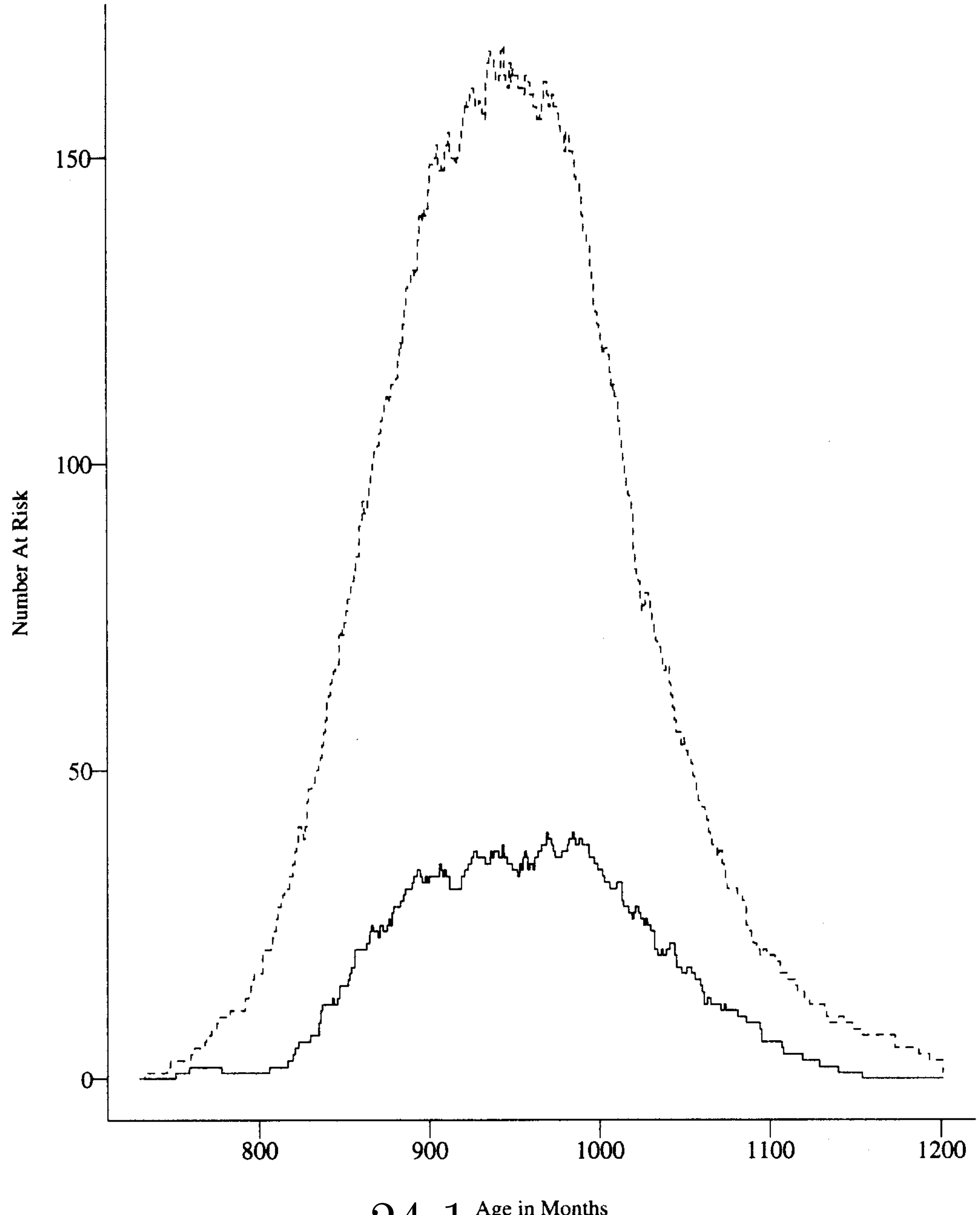
$$\hat{S}_a(t) = \prod_{a \leq t_i \leq t} \left(1 - \frac{d_i}{y_i}\right).$$

Fig 4.11.

- 2) Lai-Ying's estimator:

$$\hat{S}(t) = \prod_{t_i \le t} \left(1 - \frac{d_i}{y_i} I(y_i \ge cn^{\alpha})\right),$$

where $c > 0$, $0 < \alpha < 1$ are some constants.
Asymptotically equivalent to PL estimator, but ad hoc for finite samples; more importantly, how to choose c and α?



24-1

Figure 4.10 *Number at risk as a function of age for the 97 males (———) and the 365 females (-----) in the Channing house data set*

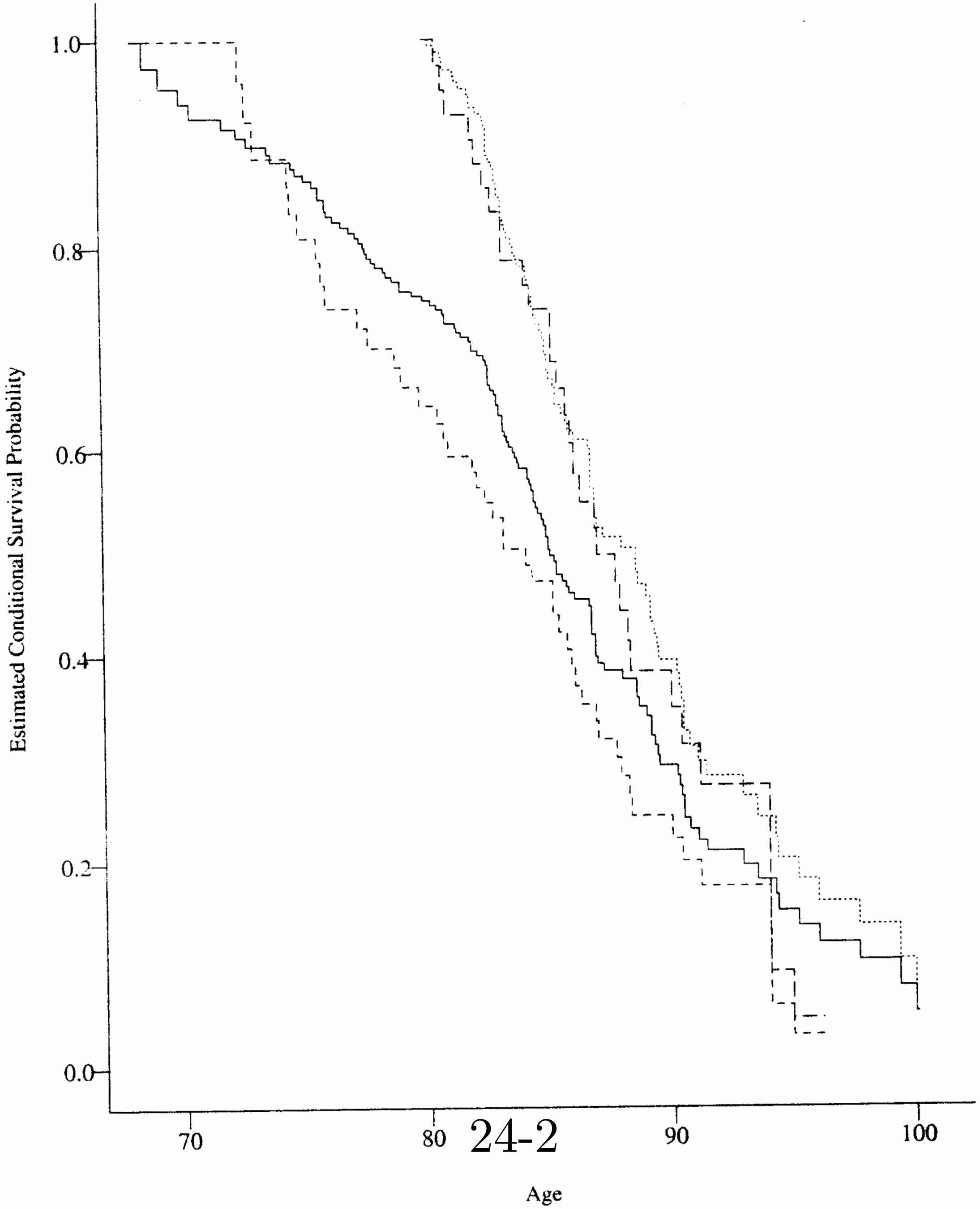


Figure 4.11 *Estimated conditional survival functions for Channing house residents. 68 year old females (————); 80 year old females (------); 68 year old males (— — —); 80 year old males (—— ——).*