chapter ③: Estimation of incomplete data.

I Complete data:

① Empirical distribution function:

Data: $x_1, \dots, x_n$.

The empirical distribution function (edf) is given by:

$$F_n(x) = \frac{\#\text{ of values } y \le x}{n}.$$

Ex: 5, 2, 5, 10, 1. Compute the edf?

1, 2, 5, 5, 10.

$$F_5(x) = \begin{cases} 0 & x < 1 \\ 1/5 & 1 \le x < 2 \\ 2/5 & 2 \le x < 5 \\ 4/5 & 5 \le x < 10 \\ 1 & x \ge 10 \end{cases}$$

* $y_1 < y_2 < \dots < y_k$ (we arrange x_i's without repetition)

s_i = number of repetitions of y_i in the sample.

$r_i = s_i + s_{i+1} + \dots + s_k$.

* we can write:

$$F_n(x) = \begin{cases} 0 & x < y_1 \\ 1 - \frac{r_i}{n} & y_{i-1} \le x < y_i, \ i = 2 \dots k. \\ 1 & x \ge y_k \end{cases}$$

Ex: ~~2, 2, 5, 10, 1~~ 5, 2, 5, 10, 1.

i	y_i	s_i	r_i
1	1	1	5
2	2	1	4
3	5	2	3
4	10	1	1

$$F_5(x) = \begin{cases} 0 & x < 1 \\ 1 - \frac{4}{5} = 1/5 & 1 \le x < 2 \\ 1 - \frac{3}{5} = 2/5 & 2 \le x < 5 \\ 1 - \frac{1}{5} = 4/5 & 5 \le x < 10 \\ 1 & x \ge 10 \end{cases}$$

(2) Nelson-Åalen estimate:

* $H(x)$ = Cumulative hazard rate function:

$$H(x) = \int_{-\infty}^{x} h(t)\,dt = \int_{-\infty}^{x} \frac{S'(t)}{S(t)}\,dt = -\ln(1 - F(x))$$

$$F(x) = 1 - e^{-H(x)}$$

* The Nelson-Åalen estimate for the Cumulative hazard rate function is:

$$\hat{H}(x) = \begin{cases} 0 & x < y_1 \\ \frac{s_1}{r_1} + \cdots + \frac{s_{i-1}}{r_{i-1}} & y_{i-1} \le x < y_i, \; i = 2 \ldots k \\ \frac{s_1}{r_1} + \cdots + \frac{s_k}{r_k} & x \ge y_k \end{cases}$$

Ex: Compute the Nelson-Åalen estimate for previous example:

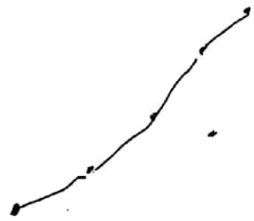
$$\hat{H}(x) = \begin{cases} 0 & x < 1 \\ \frac{1}{5} & 1 \le x < 2 \\ \frac{1}{5} + \frac{1}{4} = \frac{9}{20} & 2 \le x < 5 \\ 9/20 + \frac{2}{3} = \frac{29}{60} & 5 \le x < 10 \\ \frac{29}{60} + 1 = \frac{89}{60} & x \ge 10 \end{cases}$$

(2)

(3) Empirical distribution for grouped data:

	# of observations
$(c_0, c_1]$	n_1
$(c_1, c_2]$	n_2
$\vdots$	$\vdots$
$(c_{k-1}, c_k]$	n_k
(c_k, ∞)	0

$n = n_1 + \dots + n_k$.



* $F_n(c_j) = \frac{1}{n} \sum_{i=1}^{j} n_i$; $F_n(c_0) = 0$.

* $F_n(x) = \frac{c_j - x}{c_j - c_{j-1}} F_n(c_{j-1}) + \frac{x - c_{j-1}}{c_j - c_{j-1}} F_n(c_j)$

$c_{j-1} < x \le c_j$

Ex:

Given: Interval	# of observations	$F_{50}(\ )$
$(0, 2]$	25	$25/50 = 0.5$
$(2, 10]$	10	$35/50 = 0.7$
$(10, 100]$	10	$45/50 = 0.9$
$(100, 1000]$	5	$50/50 = 1$

$n = 50$.

Find the edf?

* $x \le 2$: $F_{50}(x) = \frac{2 - x}{2} F_{50}(0) + \frac{x - 0}{2} F_{50}(2) = 0.25x$.

$2 \le x \le 10$: $F_{50}(x) = \frac{10 - x}{8} F_{50}(2) + \frac{x - 2}{8} F_{50}(10)$

$= \frac{0.5(10 - x) + 0.7(x - 2)}{8}$

$= \frac{0.2x + 3.6}{8} = 0.025x + 0.45$.

$10 \le x \le 100$: $F_{50}(x) = \frac{100 - x}{90} F_{50}(10) + \frac{x - 10}{90} F_{50}(100)$

$= \frac{(100 - x)0.7 + (x - 10)0.9}{90}$

$= \frac{0.2x + 61}{90} = 0.0022x + 0.67$.

(3)

$100 \le x \le 1000$:

$$F_{S_0}(x) = \frac{1000-x}{900} F_{S_0}(100) + \frac{x-100}{900} F_{S_0}(1000)$$

$$= \frac{(1000-x)\,0.9 + (x-100)\,1}{900}$$

$$= \frac{0.1x + 900}{900} = 0.0001x + 0.88.$$

$$F_{S_0}(x) = \begin{cases} 0.25x & 0 \le x \le 2 \\ 0.025x + 0.45 & 2 \le x \le 10 \\ 0.0022x + 0.67 & 10 \le x \le 100 \\ 0.0001x + 0.88 & 100 \le x \le 1000 \end{cases}$$

(II) Incomplete data :

data: $x_1, x_2, x_3, \dots, x_n$.

$\longrightarrow$ $y_1, y_2, \dots, y_k$

$$s_j = \sum_{i=1}^{n} 1_{(x_i = y_j)}$$

$$r_j = \sum_{i=1}^{n} 1_{(x_i \ge y_j)} + \sum_{i=1}^{u} 1_{(u_i \ge y_j)} - \sum_{i=1}^{n} 1_{(d_i \ge y_j)}.$$

where :

u_i = censored point for the i^{th} observation.

d_i = truncated point for the i^{th} observation.

Ex :

Ex:

i	d_i	x_i	u_i
1	0	—	4
2	0	0.5	—
3	0	—	1
4	0	—	4
5	1	—	4
6	1.2	2	—
7	1.5	2	—
8	2	—	3
9	2.5	—	4
10	3.1	3.2	—

Complete the table:

j	y_j	s_j	r_j
1	0.5	1	$4+6-6=4$
2	2	2	$3+5-3=5$
3	3.2	1	$1+4-0=5$

(1) Kaplan-Meier Method (product limit):

$$S_n(t) = \begin{cases} 1 & 0 \le t \le y_1 \\ \prod_{i=1}^{j-1} \left(1 - \frac{s_i}{r_i}\right) & y_{j-1} \le t \le y_j \;,\; j = 2, \dots, k \\ \prod_{i=1}^{k} \left(1 - \frac{s_i}{r_i}\right) \text{ or } 0 & t \ge y_k \end{cases}$$

Ex: Use the previous table to compute the Kaplan-Meier survival function:

$$S_{10} = \begin{cases} 1 & 0 \le t < y_1 \\ 1 - \frac{s_1}{r_1} = 1 - \frac{1}{4} = \frac{3}{4} & y_1 \le t < y_2 \\ \frac{3}{4}\left(1 - \frac{s_2}{r_2}\right) = \frac{3}{4}\left(1 - \frac{2}{5}\right) = \frac{9}{20} & y_2 \le t < y_3 \\ \frac{9}{20}\left(1 - \frac{s_3}{r_3}\right) = \frac{9}{20}\left(1 - \frac{1}{5}\right) = \frac{9}{25} & t \ge y_3 \end{cases}$$

Ex: 2, 3, 3, 5, 5$^+$, 6, 7, 7$^+$, 9, 10$^+$.
where "+" indicates that the loss exceeded the policy limit.
Calculate $P(\text{loss} > 8)$ using Kaplan-Meier estimated edf?.

i	d_i	z_i	u_i
1	0	2	—
2	0	3	—
3	1	3	—
4	1	5	5
5	1	6	—
6	1	7	—
7	1	9	7
8	0	9	—
9	0	10	10
10		1	

j	y_j	s_j	r_j
1	2	1	7+3 = 10
2	3	2	6+3 = 9
3	5	1	4+3 = 7
4	6	1	3+2 = 5
5	7	1	2+2 = 4
6	9	1	1+1 = 2

$$S_{10}(y_2) = P(\text{loss} > y) = \begin{cases} 1 - \frac{1}{10} = 0.9 & 0 \le y < 2 \\ 0.9\left(1 - \frac{1}{9}\right) = 0.7 & 2 \le y < 3 \\ 0.7\left(1 - \frac{1}{7}\right) = 0.6 & 3 \le y < 5 \\ 0.6\left(1 - \frac{1}{5}\right) = 0.48 & 5 \le y < 6 \\ 0.48\left(1 - \frac{1}{4}\right) = 0.36 & 6 \le y < 7 \\ 0.36\left(1 - \frac{1}{2}\right) = 0.18 & 7 \le y < 9 \\ & y > 9 \end{cases}$$

$S_{10}(8) = 0.36$.

(2) Nelson – Aalen method:

$$\hat{H}(t) = \begin{cases} 0 & t < y_1 \\ \sum_{i=1}^{j-1} \frac{s_i}{r_i} & y_{j-1} \le t < y_j \ , \ j = 2 \dots k. \\ \sum_{i=1}^{k} \frac{s_i}{r_i} & t \ge y_k \end{cases}$$

$$\hat{S}(t) = e^{-\hat{H}(t)}.$$

Ex: Use the previous table to estimate the survival using Nelson – Aalen method.

$$\hat{H}(t) = \begin{cases} 0 & t < 2 \\ 0.1 & 2 \le t < 3 \\ 0.1 + \frac{2}{9} = 0.32 & 3 \le t < 5 \\ 0.32 + \frac{1}{7} = 0.46 & 5 \le t < 6 \\ 0.46 + \frac{1}{5} = 0.66 & 6 \le t < 7 \\ 0.66 + \frac{1}{4} = 0.91 & 7 \le t < 9 \\ 0.91 + \frac{1}{2} = 1.41 & t \ge 9 \end{cases}$$

$$\hat{S}(9) = e^{-\hat{H}(9)} = e^{-0.91} = 0.4.$$

(III) Mean and variance for the empirical estimators:

(1) Complete data: • Individual data.

$$\hat{S}_n(t) = \frac{\sum_{i=1}^{n} 1_{(X_i > t)}}{n} = \frac{Y}{n}.$$

$1_{(X_i > t)} \sim$ Bernoulli $(S(t))$.

$Y \sim$ Binomial $(n, S(t))$.

$E(Y) = n S(t)$; $\text{var}(Y) = n S(t)(1 - S(t))$.

$E\,\hat{S}_n(t) = S(t)$. ; $\text{var}(\hat{S}_n(t)) = \frac{\text{var}(Y)}{n^2}$

$= \frac{S(t)(1 - S(t))}{n} \to 0$

(7)

estimated variance:

$$\widehat{Var}\left(\hat{S}_n(t)\right) = \frac{\hat{S}_n(t)\left(1-\hat{S}_n(t)\right)}{n}.$$

Ex: 49, 50, 50, 50, 60, 75, 80, 120, 130.

Find the estimated variance of the estimate $\hat{S}_9(60)$.

$$\hat{S}_9(60) = \frac{4}{9}$$

$$\widehat{Var}\left(\hat{S}_9(60)\right) = \frac{\frac{4}{9}\left(1-\frac{4}{9}\right)}{9} = \frac{20}{9^3} = 0.027.$$

- Grouped data:

$$\hat{S}_n(c_j) = 1 - \frac{n_1 + \cdots + n_j}{n} = \frac{n_{j+1} + \cdots + n_k}{n}.$$

$c_{j-1} \le x \le c_j$:

$$\hat{S}_n(x) = \frac{c_j - x}{c_j - c_{j-1}}\hat{S}_n(c_{j-1}) + \frac{x - c_{j-1}}{c_j - c_{j-1}}\hat{S}_n(c_j).$$

$$\rightarrow E\left(\hat{S}_n(x)\right) = \frac{c_j - x}{c_j - c_{j-1}}S(c_{j-1}) + \frac{x - c_{j-1}}{c_j - c_{j-1}}S(c_j).$$

$$\rightarrow Var\left(\hat{S}_n(x)\right) = \frac{(c_j - c_{j-1})^2 Var(Y) + (x - c_{j-1})^2 Var(Z) + 2(c_j - c_{j-1})(x - c_{j-1})\,Cov(Y,Z)}{n^2 (c_j - c_{j-1})^2}.$$

$$Var(Y) = n\,S(c_{j-1})\left(1 - S(c_{j-1})\right)$$

$$Var(Z) = n\left(S(c_{j-1}) - S(c_j)\right)\left(1 - S(c_{j-1}) + S(c_j)\right)$$

$$Cov(Y,Z) = -n\left(1 - S(c_{j-1})\right)\left(S(c_{j-1}) - S(c_j)\right).$$

$$\left[x = c_{j-1}: \quad Var\left(\hat{S}_n(c_{j-1})\right) = \frac{S(c_{j-1})\left(1 - S(c_{j-1})\right)}{n}\right].$$

Ex:

interval	#
$(0, 2)$	25
$(2, 10]$	10
$(10, 50]$	10
$(100, 1000]$	5

Find the estimated variance of the estimate of $\hat{S}_{50}(100)$?

$$\hat{S}_{50}(100) = \frac{5}{50} = 0.1.$$

$$\widehat{Var}\left(\hat{S}_{50}(100)\right) = \frac{0.1\,(1-0.1)}{50} = \frac{0.1 \times 0.9}{50}$$

$$= \frac{0.09}{50} = 0.0018.$$

4) Estimation of variance of the Kaplan-Meier estimator:

The Greenwood's approximation of the variance estimate for the Kaplan-Meier estimator is:

$$\text{Var}\left(S_n(y_j)\right) \simeq \left(S_n(y_j)\right)^2 \sum_{i=1}^{j} \frac{s_i}{r_i(r_i - s_i)}$$

Ex: Let consider 10 payments:

$4, 4, 5^+, 5^+, 5^+, 8, 10^+, 10^+, 12, 15$.

Find the Greenwood's approximation to the variance of the Kaplan-Meier stimate $S_{10}(11)$?

j	y_j	s_j	r_j	$S_{10}(y_j)$
1	4	2	10	$1 - \frac{2}{10}$
2	8	1	5	$(1-\frac{2}{10})(1-\frac{1}{5})$
3	12	1	2	$\cdots (1-\frac{1}{2})$
4	15	1	1	$\cdots 0$.

i	d_i	x_i	u_i
1	2	4	5
2	.	4	5
.	.	—	5
.	.	—	.
.	.	—	.
1	1	8	10
.		—	10
1		12	—
10	0	15	—

$\because 8 \le 11 < 12$.

$S_{10}(11) = S_{10}(8) = \frac{16}{25}$.

$$\text{var}\left(S_{10}(8)\right) = \left(\frac{16}{25}\right)^2 * \left(\frac{2}{10(10-2)} + \frac{1}{5(5-1)}\right) = 0.03072.$$

(5) Variance estimate of the Nelson-Åalen estimator:

$$\mathrm{Var}\left(\hat{H}(y_j)\right) = \sum_{i=1}^{j} \frac{s_i}{r_i^2}.$$

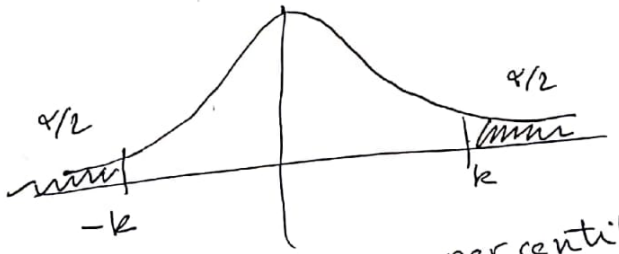
Ex: (Previous example). Find the variance estimate of the Nelson-Åalen estimator at $y = 11$.

$$\mathrm{Var}\left(\hat{H}(11)\right) = \mathrm{Var}\left(\hat{H}(8)\right) = \frac{2}{10^2} + \frac{1}{5^2} = 0.06.$$

(6) Confidence interval:

Let x a random variable with mean μ and variance σ^2.
Under the assumption that $x \sim N(\mu, \sigma^2)$, the confidence interval of μ is:

$$(\mu - k\sigma, \mu + k\sigma)$$



$k = (1 - \frac{\alpha}{2})100 - \text{percentile} = z_{\alpha/2}$.

$\longrightarrow$ $(1-\alpha)$-confidence interval:

$$(\mu - z_{\alpha/2}\sigma, \mu + z_{\alpha/2}\sigma).$$

(6.1) Kaplan-Meier estimator:

The log-transformed $100(1-\alpha)\%$ confidence interval for $S_n(t)$:

$$\left(\hat{S}_n(t)^{\frac{1}{U}} ; \hat{S}_n(t)^{U} \right),$$

$$U = \exp\left[\frac{z_{\alpha/2} \sqrt{\widehat{Var}(\hat{S}_n(t))}}{\hat{S}_n(t) \ln \hat{S}_n(t)} \right].$$

Ex: (Previous example) Find a 95% confidence interval for $S_{10}(11)$?

$$\hat{S}_{10}(11) = \frac{16}{25} ; \quad \widehat{Var}(\hat{S}_{10}(11)) = 0.03072.$$

$$z_{0.025} = 1.96$$

$$U = \exp\left(1.96 \frac{\sqrt{0.03072}}{\frac{16}{25} \ln\left(\frac{16}{25}\right)} \right) = 0.3$$

Confidence interval:

$$\left(\left(\frac{16}{25}\right)^{\frac{1}{0.3}} ; \left(\frac{16}{25}\right)^{0.3} \right) = (0.23 ; 0.87).$$

(6.2) Nelson-Åalen estimator:

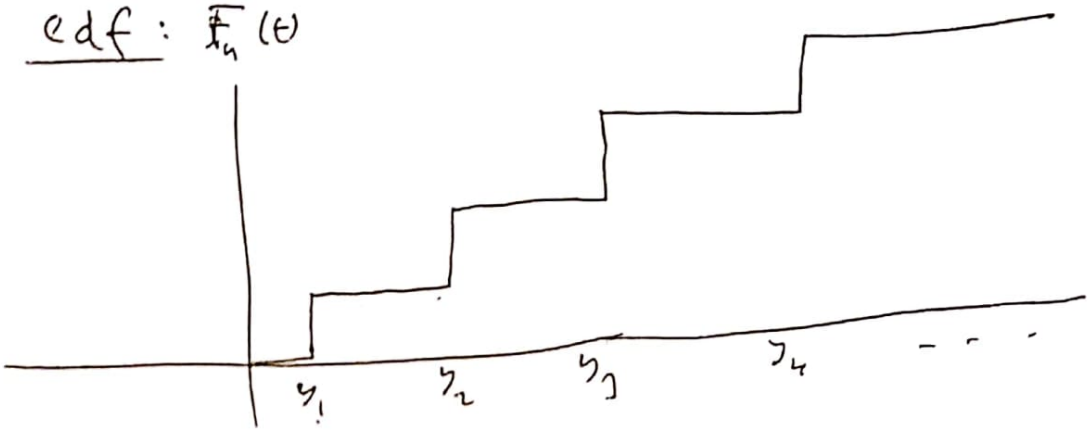
log-transformed $100(1-\alpha)\%$ confidence interval for $\hat{H}(t)$ is:

$$\left(\frac{\hat{H}(t)}{U} ; \hat{H}(t) U \right) ;$$

$$U = \exp\left(z_{\alpha/2} \frac{\sqrt{\widehat{Var}(\hat{H}(t))}}{\hat{H}(t)} \right).$$

(7) Kernel density estimation:

edf: $\hat{F}_n(t)$



let $y_1 < y_2 < \dots < y_k$ and $p(y_j) = \frac{s_j}{n}$.

$p(y_j) = F_n(y_j) - F_n(y_{j-1})$.

let the Kernel density estimator $K_{y_j}(x)$.

we define the kernel smoothed estimate of the cdf by:

$$\hat{F}(x) = \sum_{j=1}^{k} p(y_j) K_{y_j}(x).$$

and the density function:

$$\hat{f}(x) = \sum_{j=1}^{k} p(y_j) k_{y_j}(x).$$

(7.1) Uniform kernel with bandwith b:

$$k_y(x) = \begin{cases} 0 & ; \text{elsewhere.} \\ \frac{1}{2b} & ; y-b \le x \le y+b \end{cases}$$

→ is the density function of Uniform $(y-b, y+b)$.

$$K_y(x) = \begin{cases} 0 & x < y-b \\ \frac{x-y+b}{2b} & y-b \le x < y+b \\ 1 & x \ge y+b \end{cases}$$

Ex: 25, 30, 35, 35, 37, 39, 45, 47, 49, 55

Use a uniform kernel of bandwidth 10 to estimate the density function at $x = 40$.

$$\hat{f}(40) = \sum_{j=1}^{9} p(y_j)\, k_{y_j}(40)$$

$$= \frac{1}{10} k_{30}(40) + \frac{2}{10} k_{35}(40) + \frac{1}{10} k_{37}(40) + \frac{1}{10} k_{39}(40) + \frac{1}{10} k_{45}(40) + \frac{1}{10} k_{47}(40) + \frac{1}{10} k_{49}(40)$$

$$= \frac{8}{10} \cdot \frac{1}{20} = 0.04.$$

(7.2) Triangular kernel with bandwidth b.

$$k_y(x) = \begin{cases} \frac{x-y+b}{b^2} & y-b \le x \le y \\ \frac{y+b-x}{b^2} & y \le x \le y+b \\ 0 & \text{elsewhere.} \end{cases}$$

$$K_y(x) = \begin{cases} 0 & x \le y-b \\ \frac{(x-y+b)^2}{2b^2} & y-b \le x < y \\ 1 - \frac{(y+b-x)^2}{2b^2} & y \le x < y+b \\ 1 & x \ge y+b \end{cases}$$

Ex: data: 25, 30, 35, 35, 37, 39, 45, 47, 49, 55.

Use triangular kernel of bandwidth 4 to estimate the pdf at $x = 40$.

$$\hat{f}(40) = \frac{1}{10} k_{37}(40) + \frac{1}{10} k_{39}(40)$$

$$= \frac{1}{10}\frac{1}{16} + \frac{1}{10}\frac{2}{16} = \frac{1}{40} = 0.025.$$

(7.3) Gamma kernel:

The Gamma kernel $k_y(x)$ is the density function of Gamma $(\alpha, x/y)$.

$$\text{i.e. } k_y(x) = \left(\frac{x}{y}\right)^{\alpha} \frac{x^{\alpha-1} e^{-x/y}}{\Gamma(\alpha)}.$$

Ex: Same data as before:

Use Gamma kernel with $\alpha = 1$ to estimate the pdf at $x = 40$.

$$\hat{f}(40) = \frac{1}{10} k_{25}(40) + \frac{1}{10} k_{30}(40) + \cdots\cdots$$

$$= \frac{1}{10} \frac{1}{25} e^{-40/25} + \cdots\cdots$$

$$= 0.049.$$