Empirical distribution in Section 1.2

The histogram is to the estimate of a probability density function as the empirical distribution is to the estimate of a cumulative distribution function.

We begin with a very naïve method of estimating a cumulative distribution function, $F(t)$. Let t_i be the order statistics such as $t_1 < t_2 < \cdots < t_n$. Then the naïve mean rank estimator is given by

$$\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(t_i \le t),$$

which is equivalent to

$$\hat{F}_n(t) = \begin{cases} 0 & t < t_1 \\ \dfrac{i}{n} & t_i \le t < t_{i+1} \\ 1 & t \ge t_n \end{cases}.$$

For convenience, we can write the above as

$$\hat{F}_n(t_i) = \begin{cases} 0 & i = 0 \\ \dfrac{i}{n} & i = 1, 2, \ldots, n \end{cases}.$$

Note that it is quite dangerous to write the above as

$$\hat{F}_n(t_i) = \frac{i}{n} \quad \text{for } i = 0, 1, 2, \ldots, n.$$

Anyway, the above naïve method has a pitfall in that $\hat{F}_n(t) = 1$ for any $t = t_n$. Note that a desirable empirical estimator should satisfy $0 < \hat{F}_n(t) < 1$ for $t_1 \le t \le t_n$, that is, $0 < \hat{F}_n(t_i) < 1$ for $i = 1, 2, \ldots, n$. Several improvements are introduced and these are

(i) Bernard's median rank method

$$\hat{F}_n(t_i) = \begin{cases} 0 & i = 0 \\ \dfrac{i - 0.3}{n + 0.4} & i = 1, 2, \ldots, n \end{cases},$$

(ii) Blom's median rank method (aka Gaussian plotting position)

$$\hat{F}_n(t_i) = \begin{cases} 0 & i = 0 \\ \dfrac{i - 3/8}{n + 1/4} & i = 1, 2, \ldots, n \end{cases},$$

(iii) Herd-Johnson's mean rank

$$\hat{F}_n(t_i) = \begin{cases} 0 & i = 0 \\ \dfrac{i}{n + 1} & i = 1, 2, \ldots, n \end{cases},$$

(iv) Allen Hazen's midpoint method (aka Francis Galton method)

$$\hat{F}_n(t_i) = \begin{cases} 0 & i = 0 \\ \dfrac{i - 0.5}{n} & i = 1, 2, \ldots, n \end{cases}.$$

The above four methods can be special cases of the following generalized method with

$$\hat{F}_n(t_i) = \begin{cases} 0 & i = 0 \\ \dfrac{i-a}{n+1-2a} & i = 1, 2, \ldots, n \end{cases}. \tag{1}$$

Note that we have (i) Bernard's median rank method with $a = 0.3$, (ii) Blom's median rank method with $a = 3/8$, (iii) Herd-Johnson's mean rank with $a = 0$, and (iv) Allen Hazen's midpoint with $a = 0.5$.

The empirical reliability function is then easily obtained using the relation $R(t) = 1 - F(t)$,

$$\hat{R}_n(t_i) = \begin{cases} 1 & i = 0 \\ \dfrac{n+1-i-a}{n+1-2a} & i = 1, 2, \ldots, n \end{cases}. \tag{2}$$

We can also estimate the pdf $f(t)$ considering

$$f(t) = \frac{d}{dt}F(t) \approx \frac{\Delta F(t)}{\Delta t}.$$

For convenience, we let $\Delta t_i = t_{i+1} - t_i$ for $i = 1, 2, \ldots, n-1$. Note that in reliability engineering, we usually have $t_0 = 0$ and thus $\Delta t_0 = t_1$ in general. Then the pdf $f(t)$ is approximated by

$$\hat{f}_n(t_i) = \frac{\hat{F}_n(t_{i+1}) - \hat{F}_n(t_i)}{\Delta t_i},$$

where $i = 1, 2, \ldots, n-1$. Using Equation (1), we have

$$\hat{f}_n(t_i) = \begin{cases} \dfrac{1}{\Delta t_0} \cdot \left(\dfrac{1-a}{n+1-2a}\right) & i = 0 \\ \dfrac{1}{\Delta t_i} \cdot \dfrac{1}{n+1-2a} & i = 1, \ldots, n-1 \\ \text{NA} & i = n \end{cases}. \tag{3}$$

The hazard rate $h(t)$ can also be estimated using the relation $h(t) = f(t)/R(t)$. Thus, it is immediate upon using Equations (2) and (3) that

$$\hat{h}_n(t_i) = \begin{cases} \dfrac{1}{\Delta t_0} \cdot \left(\dfrac{1-a}{n+1-2a}\right) & i = 0 \\ \dfrac{1}{\Delta t_i} \cdot \dfrac{1}{n+1-i-a} & i = 1, \ldots, n-1 \\ \text{NA} & i = n \end{cases}. \tag{4}$$

Since $h(t) = f(t)/R(t) = -\frac{d}{dt}\log R(t)$, the cumulative hazard function, $H(t)$, is given by $H(t) = -\log R(t)$. Thus, we can estimate the cumulative hazard function using $\hat{R}_n(t_i)$ in Equation (2). That is, we have

$$\hat{H}_n(t_i) = -\log \hat{R}_n(t_i).$$

Example 1.3 (when $a = 0$) 1

2

i	t_i	t_{i+1}	Δt_i	$\boxed{\hat{F}_n(t_i) = \frac{i-a}{n+1-2a}}$ for $i \geq 1$	$\hat{R}_n(t_i) =$ $1 - \hat{F}_n$	$\hat{f}_n(t_i) =$ $\frac{\hat{F}_n(t_{i+1}) - \hat{F}_n(t_i)}{\Delta t_i}$	$\hat{h}_n(t_i) =$ $\frac{\hat{f}_n}{\hat{R}_n}$	$\hat{H}_n(t_i) =$ $-\log \hat{R}_n$
0	**0**	70	70	**0**(always)	**1**(always)	$\frac{0.1}{70}$		
1	**70**	150	80	$\frac{1}{9+1} = 0.1$	0.9	$\frac{0.1}{80}$		
2	**150**	250	100	$\frac{2}{9+1} = 0.2$	0.8	$\frac{0.1}{100}$		
3	**250**	360	110	$\frac{3}{9+1} = 0.3$	0.7	$\frac{0.1}{110}$		3
4	**360**	485	125	$\frac{4}{9+1} = 0.4$	0.6	$\frac{0.1}{125}$		
5	**485**	650	165	$\frac{5}{9+1} = 0.5$	0.5	$\frac{0.1}{165}$		
6	**650**	855	205	$\frac{6}{9+1} = 0.6$	0.4	$\frac{0.1}{205}$		
7	**855**	1130	275	$\frac{7}{9+1} = 0.7$	0.3	$\frac{0.1}{275}$		
8	**1130**	1540	410	$\frac{8}{9+1} = 0.8$	0.2	$\frac{0.1}{410}$		
$n = 9$	**1540**	**NA**	**NA**	$\frac{9}{9+1} = 0.9$	0.1	**NA**	**NA**	