ESTIMATION FOR COMPLETE DATA

Introduction

- Recap from chapter 4 – Data dependent distribution versus parametric distribution
  - **Definition 11.1 (13.1)** – A **data-dependent distribution** is at least as complex as the data or knowledge that produced it, and the number of “parameters” increases as the number of data points or amount of knowledge increases.
  - **Definition 11.2 (13.2)** – A **parametric distribution** is a set of distribution functions, each member of which is determined by specifying one or more values called *parameters*. The number of parameters is **fixed** and **finite**.

- Usually we deal with parametric distributions. However two data-dependent distributions are considered: the **empirical distribution** and the kernel smoothed distribution.
    - **Definition 11.3 (13.3)** – The **empirical distribution** is obtained by assigning probability 1/*n* to each data point *in the sample.*
    - **Definition 11.4 (13.4)** – A **kernel smoothed distribution** is obtained by replacing each data point *in the sample* with a continuous random variable and then assigning probability 1/*n* to each random variable. The random variable used must be identical except for a location or scale change that is related to its associated data point (see chapter 14).
    - **Observation**: The empirical distribution is a special case of kernel smoothed distribution in which the random variable assigns probability 1 to the data point (and 0 elsewhere).
- **Data sets**: When observations are collected the "ideal" situation is to have the *exact* value for each observation ("complete individual data" as in data set B and data set D1). However, complete individual data are not always available: one reason is grouping (data set C or data set A for drivers with 5 of more claims); another reason is **censoring** and/or **truncation**.

- 4 data sets are repeatedly used:

1. Data set A – Number of accidents by one driver in one year (data presented in Dropkin, 1959).
2. Data set B – Amounts paid on workers compensation medical benefits: Random sample (artificial data) of 20 payments (full amount of the loss).
3. Data set C – Random sample of payments from 227 claims from a general liability insurance. Data classified by payment range.
4. Data set D – Time at which a five-year term insurance policy terminates (artificial data). For some policyholders termination is by death, for some others it is by surrender (cancellation of the insurance contract) and for the remainder it is at the expiration of the five-years period. Two versions of this data set are presented. The first one (data set D1) with **full information** (time of death and time of surrender when applicable) and in the second one (data set D2) only the first event is recorded.

Data sets A and B will be presented in Example 11.1 (13.1).

Data Set C

Payment range		Number
Linf	Lsup	payment
0	7500	99
7500	17500	42
17500	32500	29
32500	67500	28
67500	125000	17
125000	300000	9
300000	Infinity	3
Total number of observations		227

Data set D1

Policyholder	Time of death	Time of surrender
1		0.1
2	4.8	0.5
3		0.8
4	0.8	3.9
5	3.1	1.8
6		1.8
7		2.1
8		2.5
9		2.8
10	2.9	4.6
11	2.9	4.6
12		3.9
13	4.0	
14		4.0
15		4.1
16	4.8	
17		4.8
18		4.8
19 -30		

Data set D2

Policyholder	First observed	Last Observed	Event
1	0	0.1	s
2	0	0.5	s
3	0	0.8	s
4	0	0.8	d
5	0	1.8	s
6	0	1.8	s
…			
15	0	4.1	s
16	0	4.8	d
17	0	4.8	s
18	0	4.8	s
19 -30	0	5.0	e
31	0.3	5.0	e
32	0.7	5.0	e
33	1	4.1	d
34	1.8	3.1	d
35	2.1	3.9	s
36	2.9	5.0	e
37	2.9	4.8	s
38	3.2	4.0	d
39	3.4	5.0	e
40	3.9	5.0	e

- As we can notice the information given by data sets C to D is incomplete.
    - Data set C – grouped data
    - Data set D1 – censoring: For some observations, we only know that the time of death is greater tan a given value (the time of surrender)
    - Data set D2 – censoring and truncation: Some observations are first observed at time $c > 0$

- **Censoring** and **truncation** are problems that will be discussed in the next chapter (Estimation for modified data)

The empirical distribution for complete individual data

- Let us consider a sample of size *n*, $(x_1, x_2, \cdots, x_n)$ and let us also define the indicator function of a set A by $I_A(x) = I(x \in A) = \begin{cases} 0 & x \notin A \\ 1 & x \in A \end{cases}$

- **Definition 11.5 (13.5)** – The empirical distribution function (also known as empirical cumulative distribution function or ecdf) is

$$F_n(x) = \frac{\text{number of obs} \leq x}{n} = \frac{\sum_{i=1}^{n} I(x_i \leq x)}{n}$$

- Whatever the type (discrete, continuous, mixed) of the random variable in the "theoretical" model, the empirical distribution function behaves as a distribution function of a discrete random variable.

- Klugman *et al* (*Loss Models*) introduce the concept of empirical probability function as

$$f_n(x) = \frac{\text{number of obs} = x}{n} = \frac{\sum_{i=1}^{n} I(x_i = x)}{n}.$$

- If we are sampling from a continuous random variable, the probability that we observe a tie is 0 (some exceptions can arise due to the rounding of the observed values) and consequently in most situations $f_n(x) = 1/n$;

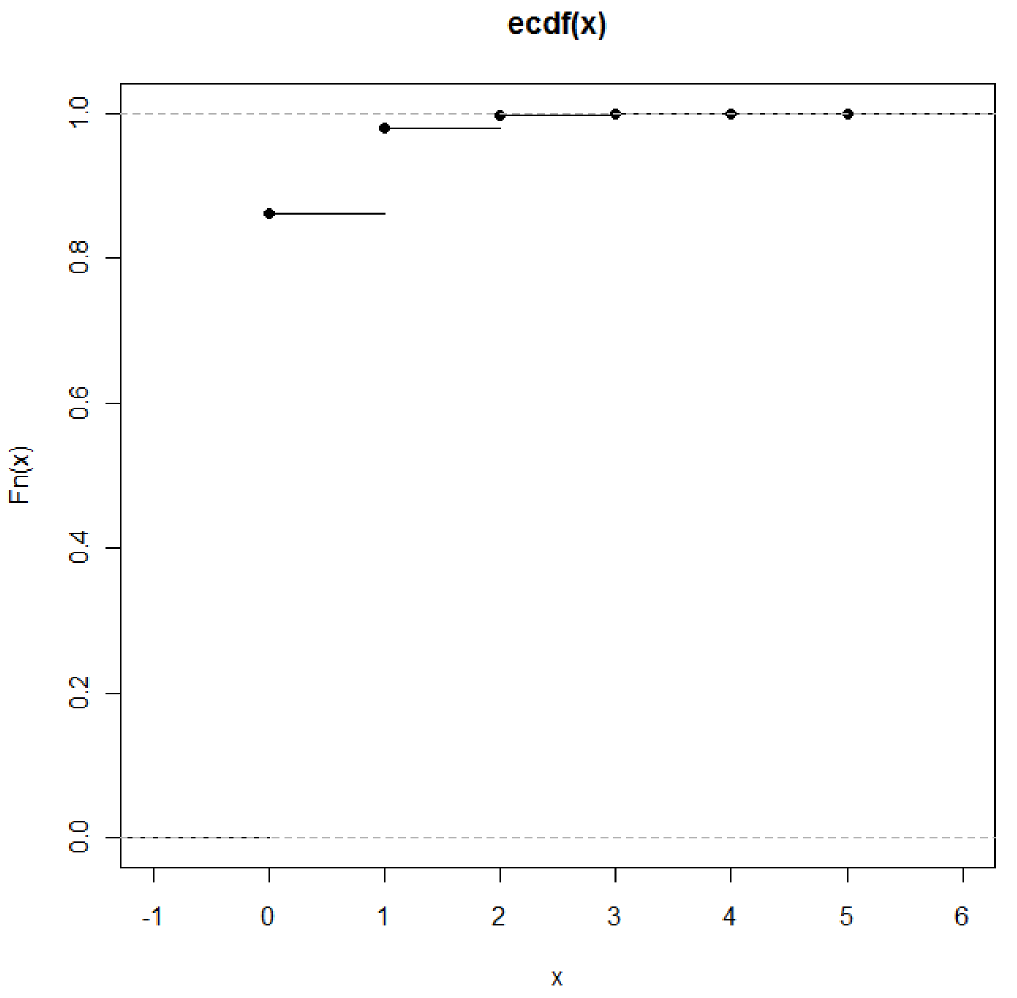
- The empirical distribution function is a much more important concept in statistical inference than the empirical probability function.

- **Example 11.1** **(13.1)** – (changed – distribution instead of probability and vice-versa) Provide the empirical distribution functions for the data in data set A and B. For data set A also provide the empirical probability function. For data set A assume that all seven drivers who had five or more accidents had exactly five accidents.

Data Set A

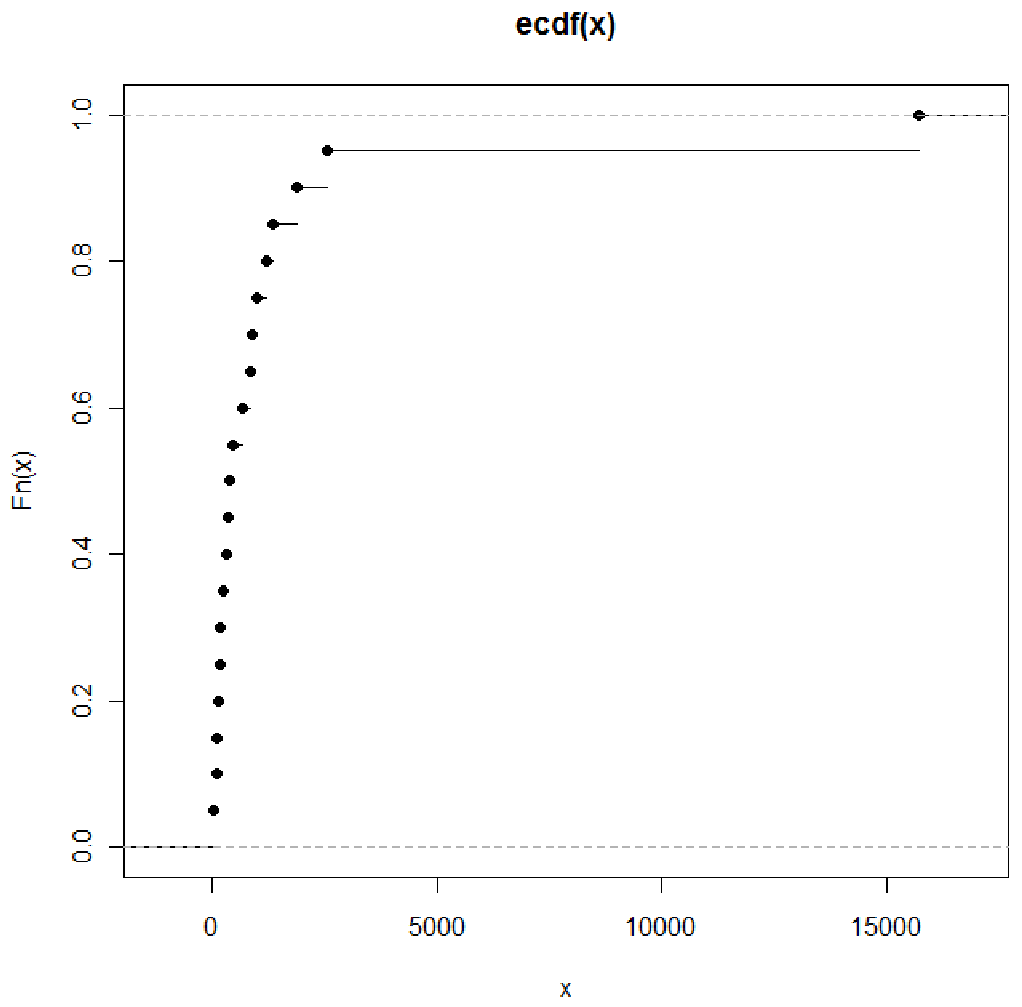
Number of Accidents	Number of drivers
0	81714
1	11306
2	1618
3	250
4	40
5 or more	7

Total number of observations 94935



Data Set B

27	82	115	126	155
161	243	294	340	384
457	680	855	877	974
1193	1340	1884	2558	15743



Data set B and R - Empirical distribution function

```
> # read data - Data set B
> x=c(27,82,115,126,155,161,243,294,340,384,457,680,855,877,974,
1193,1340,1884,2558,15743)
> F20=ecdf(x)
> summary(F20)
Empirical CDF:    20 unique values with summary
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   27.0   159.5   420.5  1424.0  1029.0 15740.0
> plot(F20)
```

Data Set A and R - Empirical distribution function

```
> # read data
>x=c(rep(0,81714),rep(1,11306),rep(2,1618),rep(3,250),rep(4,40),
rep(5,7))
> length(x)
[1] 94935
> F94935=ecdf(x)
>  summary(F94935) # Be very careful with the results!!!!
F94935 is treated as an array with 6 observations equally distributed
Empirical CDF:    6 unique values with summary
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    1.25    2.50    2.50    3.75    5.00
# To get the empirical quartiles (all equal to 0 in this example) do
> quantile(x,c(0.25,0.5,0.75))
25% 50% 75%
  0   0   0
>plot(F94935)
```

```
> # Empirical probability function
> z=rep(1,length(x)); zz=tapply(z,x,sum)
> zz
    0     1     2     3     4     5
81714 11306  1618   250    40     7
> values=as.numeric(names(zz))
> values
[1] 0 1 2 3 4 5
> EmpProb=as.numeric(zz)/sum(as.numeric(zz))
> EmpProb
[1] 8.607363e-01 1.190920e-01 1.704324e-02 2.633381e-03 4.213409e-04
[6] 7.373466e-05
> F=cumsum(EmpProb)
> F
 [1] 0.8607363 0.9798283 0.9968715 0.9995049 0.9999263 1.0000000
```

Risk set and cumulative hazard rate

- Consider a sample of size *n*, $(x_1, x_2, \cdots, x_n)$, and let $y_1 < y_2 < \cdots < y_k$ be the *k* unique values that appear in the sample ($k \le n$). Let s_j be the number of times the observation y_j appears, $j = 1,2,\cdots,k$. Obviously $\sum_{j=1}^{k} s_j = n$.

- Let us define the **risk set** as the observations that are greater than or equal to a given value (most of the time we will use "risk set" to refer the cardinal of the risk set) and let $r_j = \sum_{i=j}^{k} s_i$ be the risk set for the value y_j.

- Notice that, for $j = 2,3,\cdots,k$,

$$r_j = r_{j-1} - s_{j-1}; \qquad n - r_j = \sum_{i=1}^{j-1} s_i; \qquad r_j = n - \sum_{i=1}^{j-1} s_i, \text{ i.e. } \quad r_1 = n; \qquad r_2 = n - s_1; \quad \ldots$$

- The empirical distribution can be written as

$$F_n(x) = \begin{cases} 0 & x < y_1 \\ 1 - \frac{r_j}{n} = \frac{\sum_{i=1}^{j-1} s_i}{n} & y_{j-1} \le x < y_j \quad j = 2,3,\cdots,k \\ 1 & y_k \le x \end{cases}$$

Note that $F_n(y_1) = 1 - \frac{r_2}{n}$, $F_n(y_2) = 1 - \frac{r_3}{n}$, …, $F_n(y_k) = 1$.

- Illustrate using previous example.
- **Definition 11.6 (13.6)** – The **cumulative hazard rate function** is defined as $H(x) = -\ln S(x)$
- Recall that $S(x) = 1 - F(x) = P(X > x)$
- Note that, if $H(x)$ is differentiable, $H'(x) = -S'(x)/S(x) = f(x)/S(x) = h(x)$ and then $H(x) = \int_{-\infty}^{x} h(y)\,dy$ where $h(x) = f(x)/S(x)$ is the hazard rate function.
- From previous definition, we get $F(x) = 1 - e^{-H(x)} \Leftrightarrow S(x) = e^{-H(x)}$

- **Definition 11.7 (13.7)** – The **Nelson-Aalen estimate** of the cumulative hazard rate function is

$$\hat{H}(x) = \begin{cases} 0 & x < y_1 \\ \sum_{i=1}^{j-1} \frac{s_i}{r_i} & y_{j-1} \leq x < y_j \quad j = 2,3,\cdots,k \\ \sum_{i=1}^{k} \frac{s_i}{r_i} & y_k \leq x \end{cases}$$

- **Comment**: Although the Nelson-Aalen estimator can be used with complete individual data, it has been established in a different framework, i.e. to be used with censored (and truncated) data. We shall return to this problem latter.

- **Examples 11.2 and 11.3 (13.2 and 13.3)** – Consider a data set containing the numbers

  1.0, 1.3, 1.5, 1.5, 2.1, 2.1, 2.1, 2.8.

  Determine the quantities described in the previous paragraph and then obtain the empirical distribution function. Determine the Nelson-Aalen estimate of the cumulative hazard function.

  Solve the example using EXCEL and **R.** Use the Nelson-Aalen estimate of the cumulative hazard function to estimate the distribution function
  Nelson-Aalen estimate using R and following definition 11.7 □

```
> # Examples 11.2 and 11.3 following definition 11.7
> x=c(1.0, 1.3, 1.5, 1.5, 2.1, 2.1, 2.1, 2.8) # The sample
> z=rep(1,length(x))               # To provide a count using tapply
> zz=tapply(z,x,sum)
> zz
  1 1.3 1.5 2.1 2.8
  1   1   2   3   1
> y=as.numeric(names(zz))                    # y_j
> s=as.numeric(zz)                           # s_j
> r=rep(length(x),length(s))
> r=r-c(0,cumsum(s)[1:length(s)-1])          # r_j
> y
[1] 1.0 1.3 1.5 2.1 2.8
> s
[1] 1 1 2 3 1
> r
[1] 8 7 6 4 1
> F=c(1-r/length(x),1)
> F # Example 11.2 finished
[1] 0.000 0.125 0.250 0.500 0.875 1.000
> H=c(0,cumsum(s/r)) # Nelson-Aalen estimate # Example 11.3
> H
[1] 0.0000000 0.1250000 0.2678571 0.6011905 1.3511905 2.3511905
```

```
> F_NA=1-exp(-H)
> F_NA                                          # another estimate of F_n
[1] 0.0000000 0.1175031 0.2349829 0.4518413 0.7410682 0.9047443
```

- **Empirical survival function**
    - Using the empirical cumulative distribution function, $F_n(x)$, it is straightforward to get the empirical survival function $S_n(x) = 1 - F_n(x)$ which can act as an estimate of the survival function.

$$S_n(x) = \begin{cases} 1 & x < y_1 \\ \dfrac{r_j}{n} & y_{j-1} \le x < y_j \quad j = 2,3,\cdots,k \\ 0 & y_k \le x \end{cases}$$

    - As we will see in the next chapter the case $S_n(x)$ for $x \ge y_k$ deserves some comments.
- We can also get an estimate of the survival function using the Nelson-Aalen estimate $\hat{S}(x) = e^{-\hat{H}(x)}$

Empirical distribution for grouped data

- For grouped data it is not possible to construct the empirical distribution function. The main idea is to approximate the empirical distribution by means of 2 points:
  - Wherever it is possible (at the groups boundaries) obtain the value of the empirical distribution;
  - Connect those points using a linear interpolation (other interpolation methods are possible)

- Let the group boundaries be $c_0 < c_1 < \cdots < c_k$, i.e. group *j* is limited by c_{j-1} and c_j (often $c_0 = 0$ and $c_k = \infty$) and let us denote by n_j the number of observations in group *j*. Obviously $\sum_{j=1}^{k} n_j = n$.

- It is straightforward to see that $F_n(c_j) = (1/n)\sum_{i=1}^{j} n_i$, $j = 1,2,\cdots,k$ and that $F_n(c_0) = 0$.

- Treatment of the group boundaries: No rule is given. If the underlying variable is continuous, as it is generally the case, there is no real problem. For other situations, the best solution is to use group boundaries such that we can guarantee that the observed values are not equal to group boundaries. Technically, in order to guarantee that $F_n(x)$ is a distribution function, the value c_{j-1} should be excluded from group *j* and c_j included.

- **Definition 11.8** **(13.8)** – For grouped data the distribution function obtained by connecting the values of the empirical distribution function at the group boundaries with straight lines is called the **ogive**. The formula is

$$F_n(x) = \frac{c_j - x}{c_j - c_{j-1}} F_n(c_{j-1}) + \frac{x - c_{j-1}}{c_j - c_{j-1}} F_n(c_j), \qquad c_{j-1} \le x < c_j$$

- Comments:
  - As this function is differentiable at all points except group boundaries, the (empirical) density function can be obtained. To specify the density function at the boundaries it is arbitrarily made right continuous.
  - We can re-write the empirical distribution function as

$$F_n(x) = \frac{c_j F_n(c_{j-1}) - c_{j-1} F_n(c_j)}{c_j - c_{j-1}} + \frac{F_n(c_j) - F_n(c_{j-1})}{c_j - c_{j-1}} x, \qquad c_{j-1} \le x < c_j$$

$$\begin{aligned} S_n(x) = 1 - F_n(x) &= 1 - \frac{c_j F_n(c_{j-1}) - c_{j-1} F_n(c_j)}{c_j - c_{j-1}} - \frac{F_n(c_j) - F_n(c_{j-1})}{c_j - c_{j-1}} x \\ &= \frac{c_j S_n(c_{j-1}) - c_{j-1} S_n(c_j)}{c_j - c_{j-1}} - \frac{S_n(c_{j-1}) - S_n(c_j)}{c_j - c_{j-1}} x \end{aligned} \quad , \; c_{j-1} \le x < c_j$$

- **Definition 11.9** **(13.9)** – For grouped data the empirical density function can be obtained by differentiating the ogive. The resulting function is called a **histogram**. The formula is

$$f_n(x) = \frac{F_n(c_j) - F_n(c_{j-1})}{c_j - c_{j-1}} = \frac{n_j}{n(c_j - c_{j-1})}, \qquad c_{j-1} \le x < c_j$$

- Histograms and computer programs – be careful when classes do not have equal length

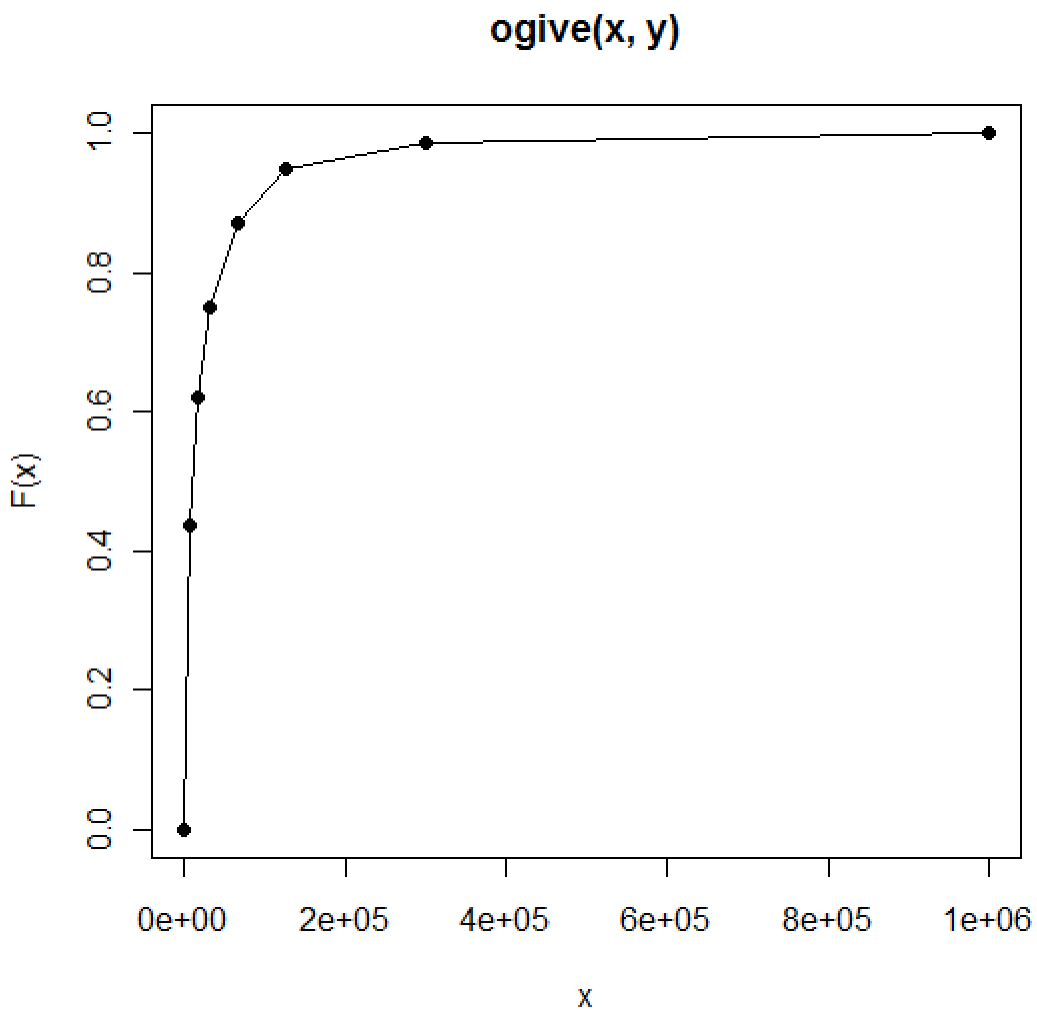
- **Example 11.5** **(13.5)** – Construct the ogive and histogram for data set C.

  Use EXCEL to define the empirical distribution function

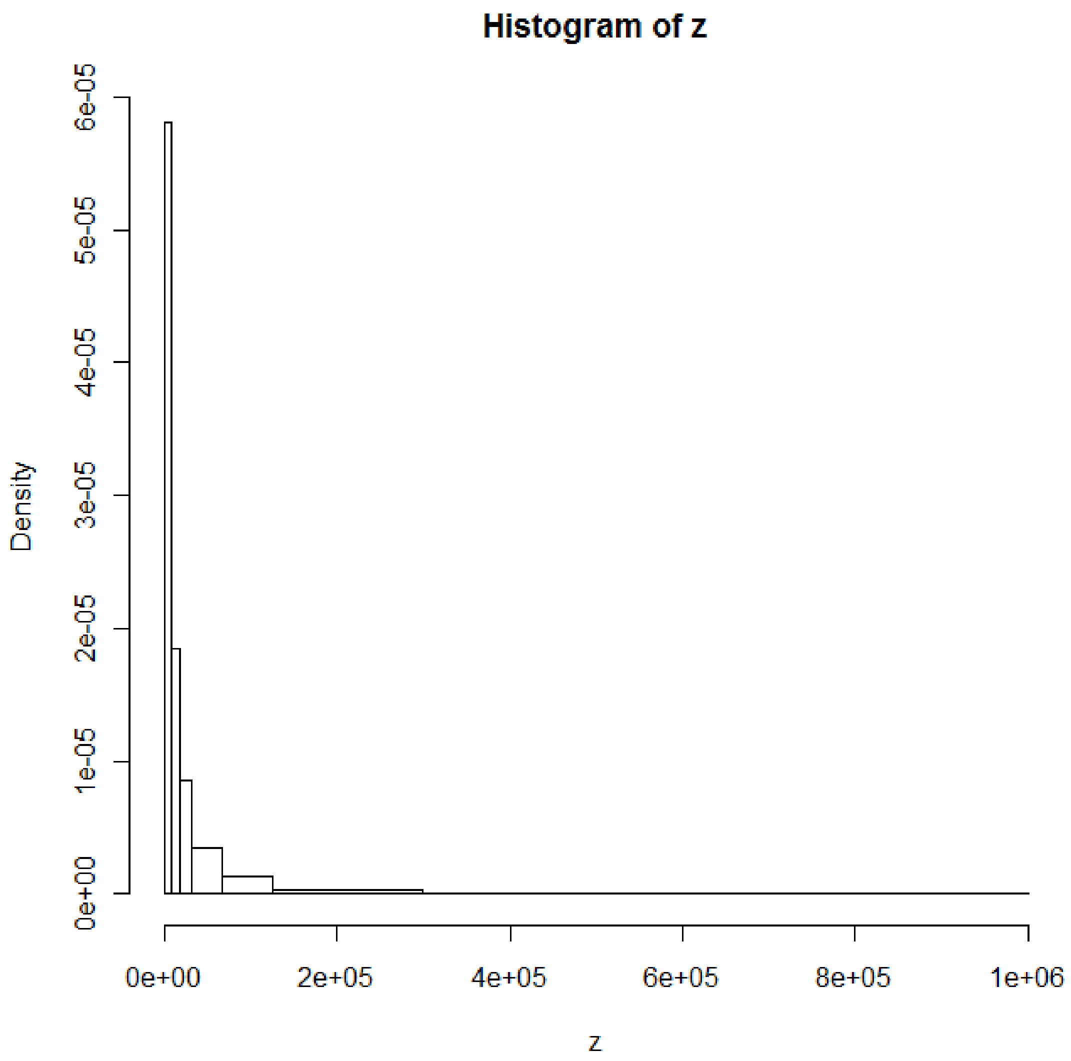
  You can also use R taking advantage of the **actuar** library or you can write your own solution

Using library actuar

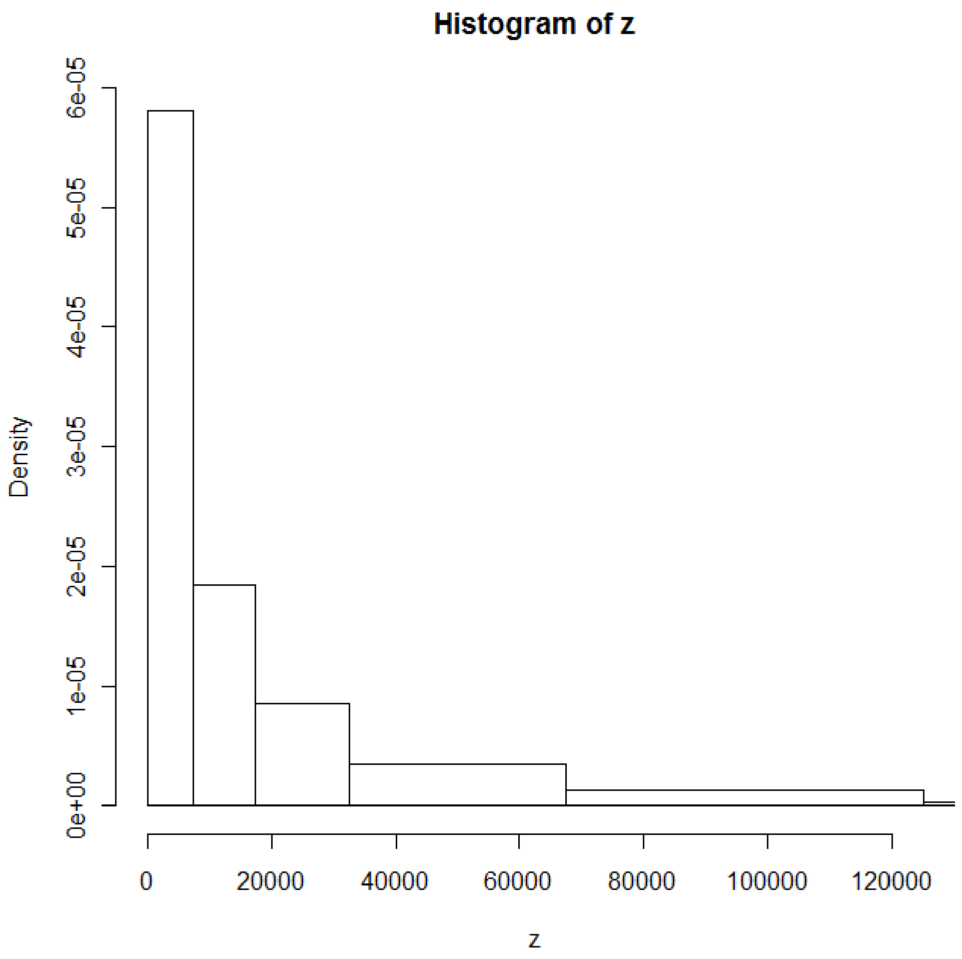
```
> library(actuar)
Attaching package: 'actuar'

The following object(s) are masked from package:grDevices : cm

> # 1000000 chosen arbitrarily
> x=c(0,7500,17500,32500,67500,125000,300000,1000000) # breaks
> y=c(99,42,29,28,17,9,3)   # counts
> a=ogive(x,y)
> a
Ogive for grouped data
Call: ogive(x, y)
    x =      0,    7500,   17500,   ...,   3e+05,   1e+06
 F(x) =      0, 0.43612, 0.62115,   ..., 0.98678,       1
> plot(a)
> a(1000)
[1] 0.05814978
> a(7500)
[1] 0.4361233
```


ogive(x, y)
F(x)
0.0
0.2
0.4
0.6
0.8
1.0
0e+00
2e+05
4e+05
6e+05
8e+05
1e+06
x

```
> lb=x[1:(length(x)-1)]; ub=c(lb[2:length(lb)],NA)
> a=cumsum(y)/sum(y);
> la=c(0,a[1:(length(a)-1)]); ua=a[1:length(a)]
> const=(ub *la-lb*ua)/(ub-lb)
> xcoef=(ua-la)/(ub-lb)
> ogive_table=data.frame(lower_bound=lb,upper_bound=ub,
constant=const,x_coef=xcoef)
> ogive_table
 lower_bound upper_bound  constant       x_coef
1           0        7500 0.0000000 5.814978e-05
2        7500       17500 0.2973568 1.850220e-05
3       17500       32500 0.4720999 8.516887e-06
4       32500       67500 0.6343612 3.524229e-06
5       67500      125000 0.7843325 1.302432e-06
6      125000      300000 0.9188169 2.265576e-07
7      300000          NA        NA           NA
> # empirical density in column 4 of ogive_table (x_coef)
> # To build array z choose an arbitrarily value in each class
> z=c(rep(5000,99),rep(10000,42),rep(20000,29),rep(50000,28),
rep(70000,17),rep(150000,9),rep(400000,3))
> b=hist(z,breaks=x)
```


Histogram of z
Density
0e+00
1e-05
2e-05
3e-05
4e-05
5e-05
6e-05
0e+00
2e+05
4e+05
6e+05
8e+05
1e+06
z

```
> hist(z,breaks=x,xlim=c(0,125000))
```



The empirical survival function (from chapter 12 (14))

- Let us consider a random sample $(X_1, X_2, \cdots, X_n)$ and let us define the **estimator** of the empirical survival function

$$S_n^*(x) = \frac{1}{n}\#\{X_i > x\} = \frac{1}{n}\sum_{i=1}^{n} I(X_i > x) = \frac{N_x}{n}, \qquad x > 0,$$

where $N_x = \#\{X_i > x\} = \sum_{i=1}^{n} I(X_i > x)$. It is straightforward to see that $N_x \sim b(n; S(x))$. If we consider an observed sample the corresponding estimate is

$$S_n(x) = \frac{1}{n}\#\{x_i > x\} = \frac{1}{n}\sum_{i=1}^{n} I(x_i > x) = \frac{n_x}{n},\ x > 0.$$

Following *Loss Models,* from now on **we will use the same notation for the estimator,** $S_n^*(x)$, **and the estimate,** $S_n(x)$. Both will be denoted by $S_n(x)$.

- **Problem 1** – How to estimate an unconditional probability like $\Pr(a < X \leq b)$?

  Noting that $\Pr(a < X \leq b) = \Pr(X > a) - \Pr(X > b) = S(a) - S(b)$ a possible **estimator** is given by

  $$\hat{\Pr}(a < X \leq b) = S_n(a) - S_n(b) = \frac{N_a - N_b}{n} = \frac{N_{(a,b]}}{n}.$$

  Defining $N_{(a,b]}$ as the number of observations in the sample that fall in the interval $(a,b]$.

  As $N_{(a,b]} \sim b(n; S(a) - S(b))$, it is straightforward to obtain the expected value and the variance of the estimator. Our estimate is

  $$\hat{\Pr}(a < X \leq b) = S_n(a) - S_n(b) = \frac{n_a - n_b}{n} = \frac{n_{(a,b]}}{n}$$

- **Problem 2** – How to estimate a conditional probability like ${}_{y-x}q_x$

  $${}_{y-x}q_x = \Pr\left(X \leq y - x + x \mid X > x\right) = \Pr\left(X \leq y \mid X > x\right) = \frac{\Pr\left(x < X \leq y\right)}{\Pr\left(X > x\right)} = \frac{S(x) - S(y)}{S(x)}$$

  The "natural" estimate is ${}_{y-x}\hat{q}_x = \dfrac{S_n(x) - S_n(y)}{S_n(x)} = \dfrac{n_x - n_y}{n_x}$, assuming that $S_n(x) > 0$.

  The corresponding estimator is ${}_{y-x}\hat{q}_x = \dfrac{N_x - N_y}{N_x}$.

  Note that **this estimator do not have neither expected value nor variance** since $\Pr(N_x = 0) > 0$.

The usual solution

Assume that $S(x)=S_n(x)$ (or equivalently that $N_x=n_x$), given that $n_x>0$.

Now the estimator is ${}_{y-x}\hat{q}_x=\dfrac{n_x-N_y}{n_x}$ but the distribution of N_y (and then the distribution of $S_n(y)$) is conditioned by $S(x)=S_n(x)$.

The estimator is still unbiased and

$$\mathrm{var}\left({}_{y-x}\hat{q}_x \mid S(x)=S_n(x)\right)=\frac{\mathrm{var}(N_y \mid N_x=n_x)}{n_x^2}=\frac{1}{n_x^2}\times n_x\times\frac{n\,S(y)}{n_x}\times\left(1-\frac{n\,S(y)}{n_x}\right)=\frac{1}{n_x^3}n\,S(y)\left(n_x-n\,S(y)\right)$$

And the estimate of the variance is

$$\hat{\mathrm{var}}\left({}_{y-x}\hat{q}_x \mid S(x)=S_n(x)\right)=\frac{1}{n_x^3}n_y\left(n_x-n_y\right)$$

How does it work?

Using the condition $S(x)=S_n(x)$ is equivalent to consider a sub-sample with all the observations greater than x and to estimate the probability of the variable being greater than y.

The sub-sample has n_x observations and we get the conditional estimator, ${}_{y-x}\hat{q}_x=\dfrac{n_x-N_y}{n_x}=1-\dfrac{N_y}{n_x}$.

Remember that, in this framework, $N_y \sim b(n_x, S(y)/S(x))$.

The variance of $\dfrac{N_y}{n_x}$, is estimated using the usual procedure applied to the sub-sample, i.e.

$$\hat{\text{var}}\left(\frac{N_y}{n_x}\right)=\frac{n_x\times\frac{n_y}{n_x}\times\left(1-\frac{n_y}{n_x}\right)}{n_x^2}=\frac{n_y\times\left(n_x-n_y\right)}{n_x^3}.$$

As it is straightforward to see, $\hat{\text{var}}\left({}_{y-x}\hat{q}_x\right)=\hat{\text{var}}\left(1-\dfrac{N_y}{n_x}\right)=\hat{\text{var}}\left(\dfrac{N_y}{n_x}\right)$.

- **Example 12.4** **(14.5)** – Using the full information of data set D1, empirically estimate q_2 and estimate the variance of this estimator.

$$x=2,\ y=3,\ n=30,\ n_2=29,\ n_3=27$$

$$\hat{q}_2=\frac{29-27}{29}=\frac{2}{29}\approx 0.06897$$

$$\hat{\text{var}}\left(\hat{q}_2 \mid S(2)=29/30\right)=\frac{27\times(29-27)}{29^3}\approx 0.002214$$

- **Example 12.5** **(14.6)** – Using data set B, empirically estimate the probability that a payment will be at least 1000 when there is a deductible of 250.

Let X be the value of a payment. Since there is a deductible of 250 we want to estimate $p=\Pr\left(X>1250 \mid X>250\right)$. Since there is a deductible we only have 13 observations

$$\hat{p}=\frac{S_n(1250)}{S_n(250)}=\frac{n_{1250}}{n_{250}}=\frac{4}{13}\approx 0.3077$$

$$\hat{\text{var}}(\hat{p})=\frac{4\times 9}{13^3}\approx 0.016386$$

Note that this variance is conditional to the existence of observations above the deductible.

Empirical estimation of probabilities

Let us consider a discrete random variable and let us assume that we want to estimate $p(x_j) = \Pr(X = x_j)$.

Let N_j be the number of times the value x_j was observed in a sample of size n. As it is straightforward to see $N_j \sim b(n; p(x_j))$.

The empirical estimator is $p_n(x_j) = N_j / n$. Consequently

$E\left(p_n(x_j)\right) = p(x_j)$, the estimator is unbiased

$\operatorname{var}\left(p_n(x_j)\right) = \dfrac{p(x_j) \times \left(1 - p(x_j)\right)}{n}$. The estimator is consistent.

The estimate of the variance is given by $\hat{\operatorname{var}}\left(p_n(x_j)\right) = \dfrac{n_j \times \left(n - n_j\right)}{n^3}$

Note that the usual approximation from the binomial to the normal distribution can be used to get a confidence interval for $p(x_j)$.

Note also that similar results can be obtained for a continuous random variable when considering the probability of a particular event.

- **Example 12.7 (14.10) –** For Data Set A determine the empirical estimate of $p(2)$ and estimate the variance of the estimator.

$n = 94935 \qquad p_n(2) = 1618/94935 \approx 0.017043$

$$\hat{\text{var}}\left(p_n(2)\right) = \frac{1618 \times (94935 - 1618))}{94935^3} \approx 1.76466 \times 10^{-7}$$

- **Example 12.8 (14.11) –** Use (10.3) and (10.4) – (12.3) and (12.4) – to construct approximate 95% confidence intervals for $p(2)$ using Data Set A

First approximation using (10.4): $\dfrac{p_n(2) - p(2)}{\sqrt{p_n(2) \times \left(1 - p_n(2)\right)/n}} \overset{\circ}{\sim} n(0;1)$

Confidence interval: $p_n(2) \pm 1.96 \times \sqrt{p_n(2) \times \left(1 - p_n(2)\right)/n}$, i.e. (0.01622; 0.01789)

Second approximation using (10.3): $\dfrac{p_n(2) - p(2)}{\sqrt{p(2) \times \left(1 - p(2)\right)/n}} \overset{\circ}{\sim} n(0;1)$

□ Confidence interval: $\dfrac{2n\,p_n(2) + 1.96^2 \pm 1.96\sqrt{1.96^2 + 4n\,p_n(2) - 4\,n\,p_n(2)^2}}{2\left(n + 1.96^2\right)}$, i.e. (0.01624; 0.01789)

Empirical survival distribution for grouped data

Let Y be the number of observations in the sample (size *n*) whose values are less than or equal to c_{j-1} and let Z be the number of observations whose value are less than or equal c_j but greater than c_{j-1}.

- Then, for $c_{j-1} \le x < c_j$, we have $S_n(x) = 1 - \dfrac{(c_j - c_{j-1})Y + (x - c_{j-1})Z}{n(c_j - c_{j-1})}$

  Remember that, from definition 12.8, $F_n(x) = \dfrac{c_j - x}{c_j - c_{j-1}} F_n(c_{j-1}) + \dfrac{x - c_{j-1}}{c_j - c_{j-1}} F_n(c_j)$. Using the new setup $F_n(c_{j-1}) = \dfrac{Y}{n}$ and $F_n(c_j) = \dfrac{Y+Z}{n}$.

- Now the marginal distributions of Y and Z are still binomial – $Y \sim b(n;1 - S(c_{j-1}))$ and $Z \sim b(n; S(c_{j-1}) - S(c_j))$ – but the joint distribution is a multinomial (trinomial) distribution (Y and Z are not independent). Then

  $E(Y) = n\ (1 - S(c_{j-1}))$; $\mathrm{var}(Y) = n(1 - S(c_{j-1}))S(c_{j-1})$;

  $E(Z) = n\ (S(c_{j-1}) - S(c_j))$; $\mathrm{var}(Z) = n\ (S(c_{j-1}) - S(c_j))(1 - S(c_{j-1}) + S(c_j))$;

  $\mathrm{cov}(Y,Z) = -n(1 - S(c_{j-1}))(S(c_{j-1}) - S(c_j))$

- The Expected value and variance of the estimator are given by

$$E\left(S_n(x)\right)=\frac{\left(c_j-x\right)}{\left(c_j-c_{j-1}\right)}S(c_{j-1})+\frac{\left(x-c_{j-1}\right)}{\left(c_j-c_{j-1}\right)}S(c_j)$$

$$\operatorname{var}\left(S_n(x)\right)=\frac{\left(c_j-c_{j-1}\right)^2\operatorname{var}(Y)+\left(x-c_{j-1}\right)^2\operatorname{var}(Z)+2\left(c_j-c_{j-1}\right)\left(x-c_{j-1}\right)\operatorname{cov}(Y,Z)}{n^2\left(c_j-c_{j-1}\right)^2}$$

- For the density estimate we get

$$f_n(x)=\frac{Z}{n\left(c_j-c_{j-1}\right)}$$

Then

$$E\left(f_n(x)\right)=\frac{E(Z)}{n\left(c_j-c_{j-1}\right)}=\frac{n\left(S(c_{j-1})-S(c_j)\right)}{n\left(c_j-c_{j-1}\right)}=\frac{S(c_{j-1})-S(c_j)}{c_j-c_{j-1}}$$

$f_n(x)$ is a biased estimator for $f(x)$. The variance is

$$\operatorname{var}\left(f_n(x)\right)=\frac{\operatorname{var}(Z)}{n^2\left(c_j-c_{j-1}\right)^2}=\frac{\left(S(c_{j-1})-S(c_j)\right)\left(1-S(c_{j-1})+S(c_j)\right)}{n\left(c_j-c_{j-1}\right)^2}$$

Example 12.6 **(14.8)** – For data set C, estimate $S(10000)$, $f(10000)$ and the variance of your estimators.

Estimates

$$S_n(10000) = 1 - \frac{99 \times 10000 + 42 \times 2500}{227 \times 10000} \approx 0.51762$$

$$f_n(x) = \frac{42}{227 \times 10000} \approx 1.85022 \times 10^{-5}$$

Estimates of the variance of the estimators

$$\hat{\text{var}}(Y) = 227 \times \frac{128}{227} \times \frac{99}{227} = \frac{12672}{227} = 55.82379$$

$$\hat{\text{var}}(Z) = 227 \times \frac{42}{227} \times \frac{185}{227} = \frac{7770}{227} = 34.22907$$

$$\hat{\text{cov}}(Y, Z) = -227 \times \frac{42}{227} \times \frac{99}{227} = -\frac{4158}{227} = -18.31720$$

$$\hat{\text{var}}\left(S_n(x)\right) = \frac{10000^2 \times \frac{12672}{227} + 2500^2 \times \frac{7770}{227} - 2 \times 10000 \times 2500 \times \frac{4158}{227}}{227^2 \times 10000^2} \approx 0.000947127$$

$$\sqrt{\hat{\text{var}}\left(S_n(x)\right)} \approx 0.030775$$

A 95% confidence interval for $S(10000)$ is given by (0.45730 ; 0.57794)

Appendix 1 - More about the ecdf (see, for instance, Wasserman (2004))

- Until now we used the symbol $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \le x)$ to represent an estimate of the distribution function (we used an observed sample).
- Let us now consider a random sample $(X_1, X_2, \cdots, X_n)$ and let us define, in this new framework, the empirical distribution function

$$F_n*(x) = \frac{1}{n}\#\{X_i \le x\} = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x), \quad -\infty < x < +\infty$$

- For each $x \in \Re$, $F_n*(x)$ is a random variable. If we assume that there are no ties in the random sample (as it is the case for continuous populations) we get

$$\Pr\left[F_n*(x) = \frac{i}{n}\right] = \binom{n}{i}[F(x)]^i\,[1 - F(x)]^{n-i} \qquad i = 0,1,\cdots,n$$

- From this result it is straightforward to deduce that $n F_n*(x)$ follows, for each $x \in \Re$, a binomial distribution with parameters *n* and $p = F(x)$, (note that $0 \le p \le 1$, i.e. the binomial can be a degenerate random variable) and then $E[F_n*(x)] = F(x)$ and $\text{var}[F_n*(x)] = F(x)\,[1 - F(x)]/n$. $F_n*(x)$ is an unbiased and consistent estimator of $F(x)$.
- Note that $\text{var}[F_n*(x)]$ can be estimated by $\overset{\wedge}{\text{var}}[F_n*(x)] = F_n(x)\,[1 - F_n(x)]/n$

- Note that, unlike $F_n(x)$, the empirical distribution function of the observed sample, $F_n^*(x)$ depends on the distribution of the random variable X and does not depend on the observed values.
- Using the Central Limit Theorem we have, for each $x \in \Re$, $\dfrac{F_n(x) - F(x)}{\sqrt{F(x)(1-F(x))/n}} \overset{\circ}{\sim} n(0;1)$ and we can use this result to get approximate confidence intervals for $F(x)$.
- Using a well-known deduction we get:
  - Lower bound: $\max\left(0; \dfrac{2nF_n(x) + z_{\alpha/2}^2 - z_{\alpha/2}\sqrt{z_{\alpha/2}^2 + 4nF_n(x)(1-F_n(x))}}{2(n + z_{\alpha/2})}\right)$
  - Upper bound: $\min\left(1; \dfrac{2nF_n(x) + z_{\alpha/2}^2 + z_{\alpha/2}\sqrt{z_{\alpha/2}^2 + 4nF_n(x)(1-F_n(x))}}{2(n + z_{\alpha/2})}\right)$
- Using the property that, for each $x \in \Re$, $F_n^*(x)$ is a consistent estimator for $F(x)$, we can also use
  - Lower bound: $\max\left(0; F_n(x) - z_{\alpha/2}\sqrt{\dfrac{F_n(x)(1-F_n(x))}{n}}\right)$
  - Lower bound: $\min\left(1; F_n(x) + z_{\alpha/2}\sqrt{\dfrac{F_n(x)(1-F_n(x))}{n}}\right)$

- **Example** – Let us go back to data set B and determine the empirical cumulative distribution function as well as a confidence interval for the distribution function. Note that the sample is far from large and consequently we will obtain poor approximations

```
> x=c(27,82,115,126,155,161,243,294,340,384,457,680,855,877,974,
1193,1340,1884,2558,15743)
> n=length(x); Fn=cumsum(rep(1,n))/n
> #95% Confidence interval
> z=qnorm(0.975)
> L1=(2*n*Fn+z*z-z*sqrt(z*z+4*n*Fn*(1-Fn)))/(2*(n+z))
> L1=(L1>0)*L1
> L2=Fn-z*sqrt(Fn*(1-Fn)/n); L2=(L2>0)*L2
> L2
> U1=(2*n*Fn+z*z+z*sqrt(z*z+4*n*Fn*(1-Fn)))/(2*(n+z))
> U1=(U1<1)*U1+(U1>=1)
> U2=Fn+z*sqrt(Fn*(1-Fn)/n); U2=(U2<1)*U2+(U2>=1)
> res=data.frame(L1,Fn,U1,L2,Fn,U2)
> res

           L1   Fn        U1         L2 Fn.1        U2
1 0.009642397 0.05 0.2563625 0.00000000 0.05 0.1455168
2 0.030254037 0.10 0.3268257 0.00000000 0.10 0.2314784
3 0.056855617 0.15 0.3912990 0.00000000 0.15 0.3064906
```

```
4  0.087568283 0.20 0.4516611 0.02469549 0.20 0.3753045
5  0.121445834 0.25 0.5088584 0.06022730 0.25 0.4397727
6  0.157941504 0.30 0.5634376 0.09916346 0.30 0.5008365
7  0.196716052 0.35 0.6157379 0.14096270 0.35 0.5590373
8  0.237553529 0.40 0.6659752 0.18529670 0.40 0.6147033
9  0.280319841 0.45 0.7142837 0.23196777 0.45 0.6680322
10 0.324941386 0.50 0.7607370 0.28086936 0.50 0.7191306
11 0.371394671 0.55 0.8053586 0.33196777 0.55 0.7680322
12 0.419703190 0.60 0.8481249 0.38529670 0.60 0.8147033
13 0.469940544 0.65 0.8889624 0.44096270 0.65 0.8590373
14 0.522240827 0.70 0.9277369 0.49916346 0.70 0.9008365
15 0.576819987 0.75 0.9642326 0.56022730 0.75 0.9397727
16 0.634017267 0.80 0.9981101 0.62469549 0.80 0.9753045
17 0.694379431 0.85 1.0000000 0.69350943 0.85 1.0000000
18 0.758852682 0.90 1.0000000 0.76852162 0.90 1.0000000
19 0.829315873 0.95 1.0000000 0.85448317 0.95 1.0000000
20 0.910748306 1.00 1.0000000 1.00000000 1.00 1.0000000
```

Appendix 2 – Confidence interval using $\frac{\sqrt{n}\;[F_n*(x)-F(x)]}{(F(x)\;[1-F(x)])^{1/2}} \overset{\circ}{\sim} n(0;1)$

To simplify the notation let us define $p = F(x)$. Then $\frac{\sqrt{n}\;[F_n*(x)-F(x)]}{(F(x)\;[1-F(x)])^{1/2}} \overset{\circ}{\sim} n(0;1)$

$$-z_{\alpha/2} < \frac{\sqrt{n}\,[F_n*(x)-F(x)]}{\left(F(x)\,[1-F(x)]\right)^{1/2}} < z_{\alpha/2} \Leftrightarrow \left|\frac{\sqrt{n}\,[F_n*(x)-F(x)]}{\left(F(x)\,[1-F(x)]\right)^{1/2}}\right| < z_{\alpha/2}$$

$$\Leftrightarrow \left(\frac{\sqrt{n}\,[F_n*(x)-F(x)]}{\left(F(x)\,[1-F(x)]\right)^{1/2}}\right)^2 < z^2_{\alpha/2}$$

$$\Leftrightarrow \frac{n\,[F_n*(x)-F(x)]^2}{\left(F(x)\,[1-F(x)]\right)} < z^2_{\alpha/2}$$

$$\Leftrightarrow n\left[\,F_n*(x)^2 + F(x)^2 - 2\,F_n*(x)F(x)\right] < z^2_{\alpha/2}F(x) - z^2_{\alpha/2}F(x)^2$$

$$\Leftrightarrow \left(n + z^2_{\alpha/2}\right)F(x)^2 - \left(2n\,F_n*(x) + z^2_{\alpha/2}\right)F(x) + n\,F_n*(x)^2 < 0$$

The left hand side is a quadratic expression in order to $F(x)$ with a positive coefficient for the quadratic term and consequently the expression will be negative for $F(x)$ inside the interval defined by the roots of the corresponding equation.

Equation: $\left(n+z_{\alpha/2}^{2}\right)F(x)^{2}-\left(2n\,F_{n}*(x)+z_{\alpha/2}^{2}\right)F(x)+n\,F_{n}*(x)^{2}=0$

Roots:

$$\frac{2n\,F_{n}*(x)+z_{\alpha/2}^{2}\pm\sqrt{\left(2n\,F_{n}*(x)+z_{\alpha/2}^{2}\right)^{2}-4\left(n+z_{\alpha/2}^{2}\right)n\,F_{n}*(x)^{2}}}{2\left(n+z_{\alpha/2}^{2}\right)}$$

But

$$\begin{aligned}\sqrt{\left(2n\,F_{n}*(x)+z_{\alpha/2}^{2}\right)^{2}-4\left(n+z_{\alpha/2}^{2}\right)n\,F_{n}*(x)^{2}} &= \sqrt{4n^{2}F_{n}*(x)^{2}+z_{\alpha/2}^{4}+4n\,z_{\alpha/2}^{2}F_{n}*(x)-4n^{2}\,F_{n}*(x)^{2}-4\,z_{\alpha/2}^{2}n\,F_{n}*(x)^{2}}\\ &= \sqrt{z_{\alpha/2}^{4}+4n\,z_{\alpha/2}^{2}F_{n}*(x)\left(1-F_{n}*(x)\right)}\\ &= z_{\alpha/2}\sqrt{z_{\alpha/2}^{2}+4n\,F_{n}*(x)\left(1-F_{n}*(x)\right)}\end{aligned}$$

And then the roots are

$$\frac{2n\,F_{n}*(x)+z_{\alpha/2}^{2}\pm z_{\alpha/2}\sqrt{z_{\alpha/2}^{2}+4n\,F_{n}*(x)\left(1-F_{n}*(x)\right)}}{2\left(n+z_{\alpha/2}^{2}\right)}$$

And we get the confidence interval for $F(x)$

$$\left(\frac{2nF_n*(x)+z_{\alpha/2}^2 - z_{\alpha/2}\sqrt{z_{\alpha/2}^2+4nF_n*(x)\left(1-F_n*(x)\right)}}{2\left(n+z_{\alpha/2}^2\right)} ; \frac{2nF_n*(x)+z_{\alpha/2}^2 + z_{\alpha/2}\sqrt{z_{\alpha/2}^2+4nF_n*(x)\left(1-F_n*(x)\right)}}{2\left(n+z_{\alpha/2}^2\right)} \right)$$

To get an observed value of the interval we replace $F_n^*(x)$ by $F_n(x)$ and we take into account that $0 \le F(x) \le 1$

Appendix 3 – Multinomial (trinomial) distribution

Y – number of observations in our sample (size *n*) whose values are less or equal to c_{j-1}

Z – number of observations whose value are less or equal c_j but greater than c_{j-1}.

Let

$p_Y = \Pr(X \le c_{j-1})$ and $p_Z = \Pr(c_{j-1} < X \le c_j)$

$(Y, Z) \sim \text{multinomial}(n, p_Y, p_Z)$

Appendix 4 – Examples 13.2 and 13.3 using R

Example 13.2 using definition 13.7

```
> # Example 13.2
> x=c(1.0, 1.3, 1.5, 1.5, 2.1, 2.1, 2.1, 2.8) # The sample
> z=rep(1,length(x))               # To provide a count using tapply
> zz=tapply(z,x,sum)
> zz
  1 1.3 1.5 2.1 2.8
  1   1   2   3   1
> y=as.numeric(names(zz))                    # y_j
> s=as.numeric(zz)                           # s_j
> r=rep(length(x),length(s))
> r=r-c(0,cumsum(s)[1:length(s)-1])          # r_j
                    Initial values of r             8   8   8   8   8
                    Values of s                     1   1   2   3   1
                    cumsum(s)                       1   2   4   7   8
                    c(0,cumsum(s)[1:length(s)-1])   0   1   2   4   7
                    values of r                     8   7   6   4   1

> y
[1] 1.0 1.3 1.5 2.1 2.8
```

```
> s
[1] 1 1 2 3 1
> r
[1] 8 7 6 4 1
```

Now we can use y s and r to obtain F_n(x), the Nelson-Aalen estimates of H(x) and of F(x)

```
> F=c(1-r/length(x),1)
```

add and additional element whose value is 1

```
> F # Example 13.2 finished
[1] 0.000 0.125 0.250 0.500 0.875 1.000
> H=c(0,cumsum(s/r)) # Nelson-Aalen estimate
> H
[1] 0.0000000 0.1250000 0.2678571 0.6011905 1.3511905 2.3511905
> F_NA=1-exp(-H)
> F_NA                                    # another estimate of F_n
[1] 0.0000000 0.1175031 0.2349829 0.4518413 0.7410682 0.9047443
>
```

Example 13.2 taking advantage of the Cox proportional hazard model (to be seen later)

The variables named *event* has all its values at 1 to indicate that there is no censoring in the sample. The difference between the Breslow and the Efron (the default) methods is based on how to deal with ties. To obtain the results presented in definition 13.7 one must use the Breslow method. The third available method (referred as “exact”) is based on a logistic regression and did not apply in our problem.

```
> x=c(1.0, 1.3, 1.5, 1.5, 2.1, 2.1, 2.1, 2.8) # The sample
> library(survival)
Loading required package: splines
> event=rep(1,8)
> fit1 <- coxph( Surv(x,event) ~ 1, method="breslow")
> h1 <- basehaz(fit1)
> h1          #  0.2679 is the cumulative hazard rate for $1.3 \le x < 1.5$
                 The first row is ommited ($\hat{H}(x) = 0$ for $x < 1.0$)
     hazard time
1 0.1250000  1.0
2 0.2678571  1.3
3 0.6011905  1.5
4 1.3511905  2.1
```

```
5 2.3511905  2.8
```

The default (Efron) method – Different results (there are ties)

```
> fit1 <- coxph( Surv(x,event) ~ 1, method="efron")
> h1 <- basehaz(fit1)
> h1
     hazard time
1 0.1250000  1.0
2 0.2678571  1.3
3 0.6345238  1.5
4 1.7178571  2.1
5 2.7178571  2.8
```

If we consider a sample without ties Efron's and Breslow's methods originate same results

```
> y=c(1.0,1.3,1.51,1.52,2.11,2.12,2.13,2.8)
> fit1 <- coxph( Surv(y,event) ~ 1, method="efron")
> h1 <- basehaz(fit1)
> h1
     hazard time
1 0.1250000 1.00
```

```
2 0.2678571 1.30
3 0.4345238 1.51
4 0.6345238 1.52
5 0.8845238 2.11
6 1.2178571 2.12
7 1.7178571 2.13
8 2.7178571 2.80
> fit1 <- coxph( Surv(y,event) ~ 1, method="breslow")
> h1 <- basehaz(fit1)
> h1
     hazard time
1 0.1250000 1.00
2 0.2678571 1.30
3 0.4345238 1.51
4 0.6345238 1.52
5 0.8845238 2.11
6 1.2178571 2.12
7 1.7178571 2.13
8 2.7178571 2.80
```

Appendix 5 – Example 13.4 using R

```
># Example 13.4
> # last value (row) has no direct meaning - value 5 for those who
did not die
> # read data
> x=c(4.8,0.8,3.1,2.9,2.9,4.0,4.8,rep(5,23))
> z=rep(1,30)
> # Empirical survival function
> zz=tapply(z,x,sum)
> y=as.numeric(names(zz))                    # y_j
> freq=as.numeric(zz)                        # s_j
> r=rep(length(x),length(y))
> r=r-c(0,cumsum(freq)[1:length(freq)-1])    # r_j
> freq
[1]  1  2  1  1  2 23
> S1=1-cumsum(freq)/sum(freq)
> S1
[1] 0.9666667 0.9000000 0.8666667 0.8333333 0.7666667 0.0000000
>
> # Nelson-Aalen without library survival - 0 for values below 0.8
omitted
```

```
> H=cumsum(freq/r)
> H
[1] 0.03333333 0.10229885 0.13933589 0.17779743 0.25779743 1.25779743
>
> # Nelson-Aalen using library survival
> library(survival)
> event=rep(1,30)
> fit1 <- coxph( Surv(x,event) ~ 1, method="breslow")
> h1 <- basehaz(fit1)
> h1
      hazard time
1 0.03333333  0.8
2 0.10229885  2.9
3 0.13933589  3.1
4 0.17779743  4.0
5 0.25779743  4.8
6 1.25779743  5.0
> S2=exp(-h1$hazard)
> S2
[1] 0.9672161 0.9027597 0.8699358 0.8371120 0.7727518 0.2842795
```

Appendix 6 – Examples 13.5 using R

You can also use R taking advantage of the actuar library or you can write your own solution

Without library actuar

```
> ogiv=function(F,b,x){
+   r=-1; k=length(F); a=0;
+   if (x<b[1]) {r=0; a=1}
+   for(i in 2:k) if (x<=b[i]) {
+     if (a==0) {
+       r=(b[i]-x)*F[i-1]/(b[i]-b[i-1])+(x-b[i-1])*F[i]/(b[i]-
b[i-1]);
+       a=1;
+       }
+     }
+   return(r); #-1 is equivalent to undefined
+   }
>
> # 1000000 chosen arbitrarily
> x=c(0,7500,17500,32500,67500,125000,300000,1000000) # breaks
> y=c(99,42,29,28,17,9,3)  # counts
> F=c(0,cumsum(y)/sum(y))
```

```
>
> # plot the ogive
> plot(x,F,type="l")
> points(x,F)
>
> ogiv(F,x,1000); ogiv(F,x,7500)
[1] 0.05814978
[1] 0.4361233
```

Appendix 7 – Variance of the conditional survival function

Let us now assume that we want to estimate $S(x)$, given that we knew that $S(y)=s,\ x>y$.

In this setup we consider only the observations where $x_i > y$ and we estimate the value of the survival function $S(x)$ which must fall between $S(y)$ and 0. We get the estimator $\hat{S}(x\,|\,S(y)=s)=\dfrac{N_x}{N_y}\times s$ and the estimate $\hat{S}(x\,|\,S(y)=s)=\dfrac{n_x}{n_y}\times s$.

Note that, in this framework, $N_x\,|\,S(y)=s \sim b\big(N_y\,;S(x)/s\big)$

- To be more specific, let us assume that $S(y)=S_n(y)$, $S_n(y)$ being the estimate of $S(y)$ obtained using the observed sample. In this setup we will have $\hat{S}(x\,|\,S(y)=S_n(y))=\dfrac{N_x}{n_y}\times\dfrac{n_y}{n}=\dfrac{N_x}{n}$. The estimate will be the same as in the unconditional case. Now $N_x\,|\,S(y)=S_n(y)\sim b\big(n_y\,;n\,S(x)/n_y\big)$. The conditional estimator will still be unbiased but with a smaller variance, as expected.

$$E(N_x\,|\,S(y)=S_n(y))=n_y\times n\,S(x)/n_y\ =n\,S(x)$$

$$E\big(\hat{S}(x\,|\,S(y)=S_n(y))\big)=E\left(\frac{N_x}{n}\right)=S(x)$$

$$\operatorname{var}\left(\hat{S}(x \mid S(y)=S_n(y))\right)=\operatorname{var}\left(\frac{N_x}{n} \mid S(y)=S_n(y)\right)=\frac{1}{n^2}\times n_y\times\frac{n\,S(x)}{n_y}\times\left(1-\frac{n\,S(x)}{n_y}\right)$$

$$=\frac{S(x)}{n}\times\left(1-\frac{n\,S(x)}{n_y}\right)$$

It is straightforward to see that the conditional variance is smaller than the unconditional variance since

$$\operatorname{var}(S_n(x))=\frac{S(x)\times(1-S(x))}{n}$$

and

$$\operatorname{var}(S_n(x))\geq\operatorname{var}\left(\hat{S}(x \mid S(y)=S_n(y))\right)\Leftrightarrow\frac{S(x)\times(1-S(x))}{n}\geq\frac{S(x)}{n}\times\left(1-\frac{n\,S(x)}{n_y}\right)$$

$$\Leftrightarrow 1-S(x)\geq 1-\frac{n\,S(x)}{n_y}\Leftrightarrow\frac{n\,S(x)}{n_y}\geq S(x)\Leftrightarrow n\geq n_y$$

Which is obvious.

The variance estimate will be $\hat{\operatorname{var}}\left(\hat{S}(x \mid S(y)=S_n(y))\right)=\frac{n_x}{n^2}\times\left(1-\frac{n_x}{n_y}\right)$

Appendix 8 – Expected value and variance for grouped data

Y – number of observations in our sample (size *n*) whose values are less or equal to c_{j-1}

Z – number of observations whose values are less or equal c_j but greater than c_{j-1}.

$$S_n(x) = 1 - \frac{\left(c_j - c_{j-1}\right)Y + \left(x - c_{j-1}\right)Z}{n\left(c_j - c_{j-1}\right)}, \quad c_{j-1} \le x < c_j$$

- Prove that

$$E\left(S_n(x)\right) = \frac{\left(c_j - x\right)}{\left(c_j - c_{j-1}\right)} S(c_{j-1}) + \frac{\left(x - c_{j-1}\right)}{\left(c_j - c_{j-1}\right)} S(c_j)$$

$$\operatorname{var}\left(S_n(x)\right) = \frac{\left(c_j - c_{j-1}\right)^2 \operatorname{var}(Y) + \left(x - c_{j-1}\right)^2 \operatorname{var}(Z) + 2\left(c_j - c_{j-1}\right)\left(x - c_{j-1}\right)\operatorname{cov}(Y, Z)}{n^2\left(c_j - c_{j-1}\right)^2}$$

where

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

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

$$\operatorname{cov}(Y; Z) = -n \times \left(S(c_{j-1}) - S(c_j)\right) \times \left(1 - S(c_{j-1})\right)$$

- It is straightforward to see that

$Y \sim b(n;F(c_{j-1}))$, i.e. $Y \sim b\left(n;1-S(c_{j-1})\right)$.

$$E(Y) = n\ F(c_{j-1}) = n(1-S(c_{j-1}))$$
$$\operatorname{var}(Y) = n\,F(c_{j-1})\,S(c_{j-1}) = n\,S(c_{j-1})\left(1-S(c_{j-1})\right)$$

$Z \sim b\left(n;F(c_j)-F(c_{j-1})\right)$, i.e. $Z \sim b\left(n;S(c_{j-1})-S(c_j)\right)$.

$$E(Z) = n\ \left(F(c_j)-F(c_{j-1})\right) = n\ \left(S(c_{j-1})-S(c_j)\right)$$
$$\begin{aligned}\operatorname{var}(Z) &= n\ \left(F(c_j)-F(c_{j-1})\right)\left(1-F(c_j)+F(c_{j-1})\right)\\ &= n\ \left(S(c_{j-1})-S(c_j)\right)\left(1-S(c_{j-1})+S(c_j)\right)\end{aligned}$$

- Using these results it is straightforward to prove the statement about the expected value. To prove the second statement (the one about the variance) we use the usual formula about the variance of a sum of random variables and then we must prove that $\operatorname{cov}(Y;Z) = -n\times\left(S(c_{j-1})-S(c_j)\right)\times\left(1-S(c_{j-1})\right)$.

$$\operatorname{cov}(Y,Z) = E\left(Y\,Z\right)-E\left(Y\right)E\left(Z\right)$$
$$E\left(Y\,Z\right) = E\left(E\left(Y\,Z \mid Y\right)\right) = E\left(Y\,E\left(Z \mid Y\right)\right)$$

But

$$Z \mid Y \sim b\left(n-Y\,; \frac{S(c_{j-1})-S(c_j)}{S(c_{j-1})} \right)$$

$$E(Z \mid Y) = (n-Y)\times\frac{S(c_{j-1})-S(c_j)}{S(c_{j-1})}$$

$$\operatorname{var}(Z \mid Y) = (n-Y)\times\frac{S(c_{j-1})-S(c_j)}{S(c_{j-1})}\times\left(1-\frac{S(c_{j-1})-S(c_j)}{S(c_{j-1})}\right) = (n-Y)\times\frac{S(c_{j-1})-S(c_j)}{S(c_{j-1})}\times\frac{S(c_j)}{S(c_{j-1})}$$

Then

$$
\begin{aligned}
E\left(Y\,Z\right) &= E\left(Y\,E\left(Z \mid Y\right)\right) = E\left(Y \times \left(n-Y\right) \times \frac{S(c_{j-1})-S(c_j)}{S(c_{j-1})}\right) \\
&= \frac{S(c_{j-1})-S(c_j)}{S(c_{j-1})} \times E\left(Y \times \left(n-Y\right)\right) = \frac{S(c_{j-1})-S(c_j)}{S(c_{j-1})} \times \left(n\,E(Y) - E(Y^2)\right) \\
&= \frac{S(c_{j-1})-S(c_j)}{S(c_{j-1})} \times \left(n\,E(Y) - \operatorname{var}(Y) - E^2(Y)\right) \\
&= \frac{S(c_{j-1})-S(c_j)}{S(c_{j-1})} \times \left(n^2\,(1-S(c_{j-1})) - n\,S(c_{j-1})\left(1-S(c_{j-1})\right) - n^2\,(1-S(c_{j-1}))^2\right) \\
&= \frac{S(c_{j-1})-S(c_j)}{S(c_{j-1})} \times \left(n^2\,(1-S(c_{j-1}))\,S(c_{j-1}) - n\,S(c_{j-1})\left(1-S(c_{j-1})\right)\right) \\
&= \left(S(c_{j-1})-S(c_j)\right) \times \left(1-S(c_{j-1})\right) \times n \times \left(n-1\right)
\end{aligned}
$$

$$
\begin{aligned}
\operatorname{cov}(Y\,;Z) &= E\left(Y\,Z\right) - E\left(Y\right)E\left(Z\right) \\
&= \left(S(c_{j-1})-S(c_j)\right) \times \left(1-S(c_{j-1})\right) \times n \times \left(n-1\right) - n\,(1-S(c_{j-1})) \times n\,\left(S(c_{j-1})-S(c_j)\right) \\
&= \left(S(c_{j-1})-S(c_j)\right) \times \left(1-S(c_{j-1})\right) \times \left(n^2 - n - n^2\right) \\
&= -n \times \left(S(c_{j-1})-S(c_j)\right) \times \left(1-S(c_{j-1})\right)
\end{aligned}
$$

Note that

$$E(Z)=E(E(Z\mid Y))=E\left((n-Y)\times\frac{S(c_{j-1})-S(c_j)}{S(c_{j-1})}\right)=\frac{S(c_{j-1})-S(c_j)}{S(c_{j-1})}\times(n-E(Y))$$

$$=\frac{S(c_{j-1})-S(c_j)}{S(c_{j-1})}\times n\times S(c_{j-1})=n\times\left(S(c_{j-1})-S(c_j)\right)$$

And, as expected,

$$\operatorname{var}(Z)=\operatorname{var}(E(Z\mid Y))+E(\operatorname{var}(Z\mid Y))$$

$$=\operatorname{var}\left((n-Y)\times\frac{S(c_{j-1})-S(c_j)}{S(c_{j-1})}\right)+E\left((n-Y)\times\frac{S(c_{j-1})-S(c_j)}{S(c_{j-1})}\times\frac{S(c_j)}{S(c_{j-1})}\right)$$

$$=\left(\frac{S(c_{j-1})-S(c_j)}{S(c_{j-1})}\right)^2\times\operatorname{var}(n-Y)+\frac{S(c_{j-1})-S(c_j)}{S(c_{j-1})}\times\frac{S(c_j)}{S(c_{j-1})}\times E(n-Y)$$

$$=\left(\frac{S(c_{j-1})-S(c_j)}{S(c_{j-1})}\right)^2\times n\,S(c_{j-1})\left(1-S(c_{j-1})\right)+\quad+\frac{S(c_{j-1})-S(c_j)}{S(c_{j-1})}\times\frac{S(c_j)}{S(c_{j-1})}\times\left(n-n\,(1-S(c_{j-1}))\right)$$

$$
\begin{aligned}
\operatorname{var}(Z) &= n\times\frac{\left(S(c_{j-1})-S(c_j)\right)^2}{S(c_{j-1})}\times\left(1-S(c_{j-1})\right)+n\times\frac{S(c_{j-1})-S(c_j)}{S(c_{j-1})}\times S(c_j)\\
&= n\times\frac{\left(S(c_{j-1})-S(c_j)\right)}{S(c_{j-1})}\times\left(\left(S(c_{j-1})-S(c_j)\right)\times\left(1-S(c_{j-1})\right)+S(c_j)\right)\\
&= n\times\frac{\left(S(c_{j-1})-S(c_j)\right)}{S(c_{j-1})}\times\left(S(c_{j-1})-S(c_j)-S(c_{j-1})^2+S(c_j)\times S(c_{j-1})+S(c_j)\right)\\
&= n\times\left(S(c_{j-1})-S(c_j)\right)\times\left(1-S(c_{j-1})+S(c_j)\right)
\end{aligned}
$$

Appendix 9 – Example 14.11

- **Example 14.11 –** Use (12.3) and (12.4) to construct approximate 95% confidence intervals for $p(2)$ using Data Set A

Second approximation using (12.3)

$$\frac{p_n(2)-p(2)}{\sqrt{p(2)\times(1-p(2))/n}} \overset{\circ}{\sim} n(0;1)$$

$$\left|\frac{p_n(2)-p(2)}{\sqrt{p(2)\times(1-p(2))/n}}\right| < 1.96$$

$$\frac{(p_n(2)-p(2))^2\, n}{p(2)\times(1-p(2))} < 1.96^2 \Leftrightarrow$$

$$n\, p_n(2)^2 + n\, p(2)^2 - 2n\, p_n(2)\, p(2) - 1.96^2\, p(2) + 1.96^2\, p(2)^2 < 0 \Leftrightarrow$$

$$p(2)^2\left(n+1.96^2\right) - p(2)\left(2n\, p_n(2)+1.96^2\right) + n\, p_n(2)^2 < 0$$

$$p(2)^2\left(n+1.96^2\right)-p(2)\left(2n\,p_n(2)+1.96^2\right)+n\,p_n(2)^2<0$$

$$\Delta=\left(2n\,p_n(2)+1.96^2\right)^2-4\left(n+1.96^2\right)n\,p_n(2)^2$$

$$=4n^2\,p_n(2)^2+1.96^4+4n\,p_n(2)1.96^2-4\,n^2\,p_n(2)^2-4\,n\,p_n(2)^2\,1.96^2$$

$$=1.96^2\left(1.96^2+4n\,p_n(2)-4\,n\,p_n(2)^2\right)$$

Confidence interval:

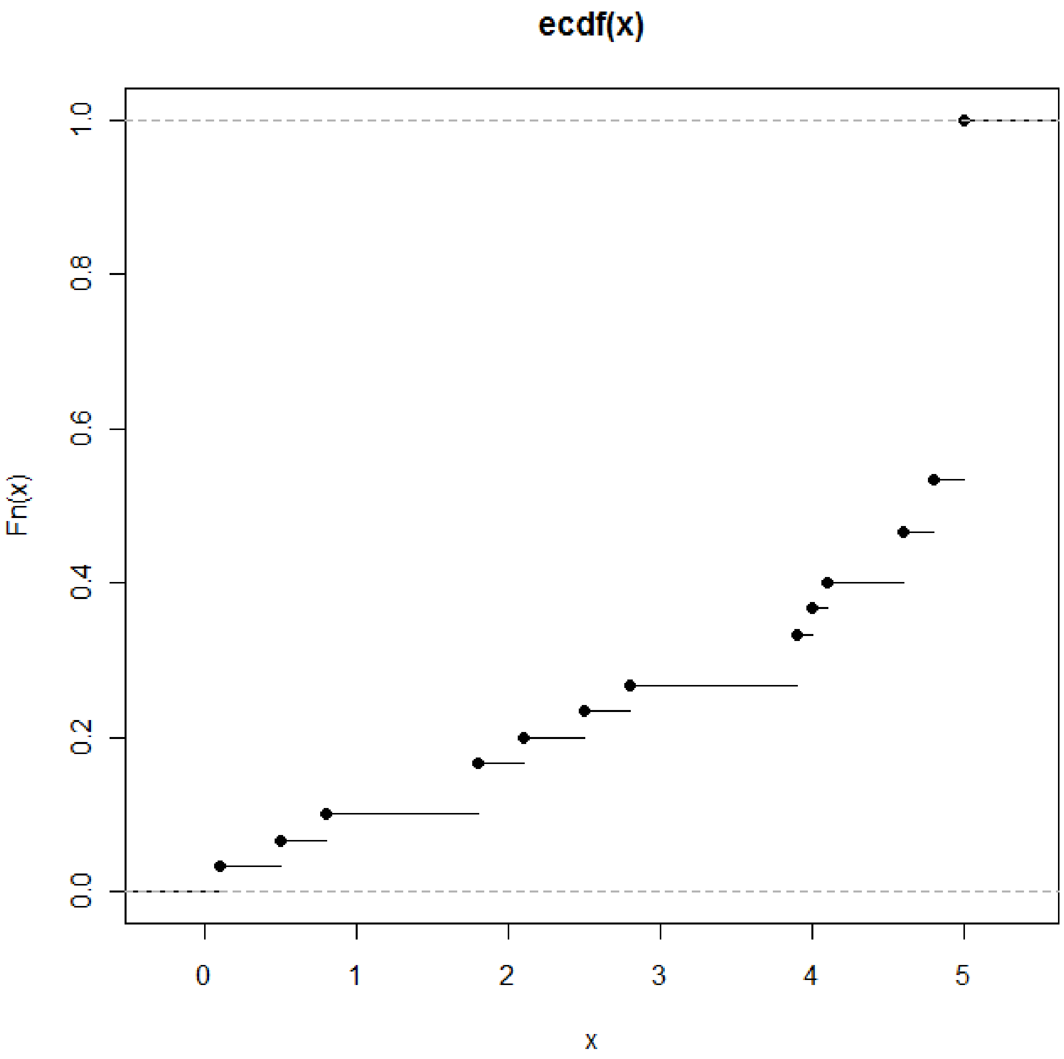
$$\frac{2n\,p_n(2)+1.96^2\pm1.96\sqrt{1.96^2+4n\,p_n(2)-4\,n\,p_n(2)^2}}{2\left(n+1.96^2\right)}$$, i.e. (0.016239; 0.017886)

Appendix 10 – Solution of chapter 13 exercises using R

13.1 –

```
> # read data
>
x=c(.1,.5,.8,3.9,1.8,1.8,2.1,2.5,2.8,4.6,4.6,3.9,5.0,4.0,4.1,5.0,4.8,
4.8,rep(5,12))
>
> # Empirical distribution function
> F30=ecdf(x)
> plot(F30)
> z=rep(1,30)
> tapply(z,x,sum)
0.1 0.5 0.8 1.8 2.1 2.5 2.8 3.9   4 4.1 4.6 4.8   5
  1   1   1   2   1   1   1   2   1   1   2   2  14
> values=as.numeric(names(zz))
> values
[1] 0 1 2 3 4 5
> freq=as.numeric(zz)
> F=cumsum(freq)/sum(freq)
> F
```

```
[1] 0.8607363 0.9798283 0.9968715 0.9995049 0.9999263 1.0000000
>
> # Nelson-Aalen estimate
> library(survival)
> status=rep(1,30)
> fit1 <- coxph( Surv(x,status) ~ 1, method="breslow")
> h1 <- basehaz(fit1)
> h1
       hazard time
1  0.03333333  0.1
2  0.06781609  0.5
3  0.10353038  0.8
4  0.17760445  1.8
5  0.21760445  2.1
6  0.25927112  2.5
7  0.30274938  2.8
8  0.39365847  3.9
9  0.44365847  4.0
10 0.49629005  4.1
11 0.60740116  4.6
12 0.73240116  4.8
13 1.73240116  5.0
> FF=1-exp(-h1$hazard)
> FF
 [1] 0.03278390 0.06556769 0.09835137 0.16272646 0.19555643
0.22838620
 [7] 0.26121577 0.32541559 0.35831546 0.39121496 0.45523521
0.51924676
[13] 0.82314077
```


ecdf(x)
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
x

13.2 –

```
> # read data
>
year=c(64,68,71,56,61,66,55,58,74,59,71,76,64,49,59,50,54,73,80,64,55
,67,57,79,75,72,64,60,61,69,54,54,70,79,65)
>
loss=c(6766,7123,10562,14474,15351,16983,18383,19030,25304,29112,3014
6,33727,40596,41409,47905,49397,52600,59917,63123,77809,102942,103217
,123680,140136,192013,198446,227338,329511,361200,421680,513586,54577
8,750389,863881,1638000)
>
> # a
> summary(loss)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   6766   27210   59920  204900  212900 1638000
> ?var
> sd(loss)
[1] 330563.7
> sd(loss)/mean(loss)
[1] 1.613290
```

```
> mean(loss)
[1] 204900.4
> library(help="Deducer")
> skewness(loss)
[1] 2.664174
>
># b ????
```

13.3

13.10

Left limit	0	50	150	250	500
Right limit	50	150	250	500	1000
Nº Observations	36	x	y	84	80

Let n be the sample size.

$n = 36 + x + 84 + 80 = 200 + x + y$

We also know that $F_{n(}(90) = 0.21$ and $F_{n(}(210) = 0.51$. Since

$F_n(x) = \dfrac{c_j - x}{c_j - c_{j-1}} F_n(c_{j-1}) + \dfrac{x - c_{j-1}}{c_j - c_{j-1}} F_n(c_j)$, $c_{j-1} \le x < c_j$, we have

$$F_n(90) = \frac{150-90}{150-50} F_n(50) + \frac{90-50}{150-50} F_n(150)$$

$$= 0.6 \times \frac{36}{n} + 0.4 \times \frac{36+x}{n} = \frac{36+x}{n}$$

$$F_n(210) = \frac{250-210}{250-150} F_n(150) + \frac{210-150}{250-150} F_n(250)$$

$$= 0.4 \times \frac{36+x}{n} + 0.6 \times \frac{36+x+y}{n}$$

$$= 0.4 \times \frac{36+x}{n} + 0.6 \times \frac{n-164}{n}$$

$$= \frac{14.4 + 0.4x - 98.4}{n} + 0.6 = 0.6 - \frac{84 - 0.4x}{n}$$

Let us solve the system

$$\begin{cases} 0.21 = \dfrac{36+0.4x}{n} \\ 0.51 = 0.6 - \dfrac{84-0.4x}{n} \end{cases} \Leftrightarrow \begin{cases} 0.21 = \dfrac{36+0.4x}{n} \\ 0.09 = \dfrac{84-0.4x}{n} \end{cases} \Leftrightarrow \begin{cases} 0.4x = 0.21n - 36 \\ 0.09 = \dfrac{84-0.21n+36}{n} \end{cases}$$

$$\Leftrightarrow \begin{cases} 0.4x = 0.21n - 36 \\ 0.3n = 120 \end{cases} \Leftrightarrow \begin{cases} 0.4x = 84 - 36 \\ n = 400 \end{cases} \Leftrightarrow \begin{cases} 0.4x = 48 \\ n = 400 \end{cases} \Leftrightarrow \begin{cases} x = 120 \\ n = 400 \end{cases}$$