NON-PARAMETRIC ESTIMATION

- To avoid presenting topics that will be discussed in Survival Models and Life Contingencies (2nd semester) and Actuarial Topics (3rd semester), we will only cover parts of chapters 11 (13) and 12 (14) of *Loss Models* book:
    - From chapter 11 (13) we will cover section 2 until example 11.1 (13.1) and section 3 (when solving exercise 11.1 skip, “Nelson-Aalen estimate)
    - From chapter 12 (14) we will cover section 2 until example 12.8 (14.11) and section 3. In section 2 only the first 2 exercises are covered.

Introduction

- The main purpose is to estimate, using a **non-parametric framework**, the distribution function of a risk or its survival function, $S(x)=P(X>x)=1-F(x)$. The risk can be linked to human life (duration for instance) or to non-life insurance (claim amounts for instance).

- Which information is available? In *Loss Models* the first chapter uses **complete information** while the second uses **modified data**. In both chapters, 4 datasets 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

- 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); other reasons are **censoring** and/or **truncation**.

- **Censoring** and **truncation** are problems that will be analyzed in more detail when discussing frequentist estimation (next chapter).

- 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 than a given value (the time of surrender)
    - Data set D2 – censoring and truncation: Some observations are first observed at time 0 and others at time $c > 0$

- **Definition 12.1 (14.1)** – An observation is **truncated from below** (also called left truncated) at *d* if when it is below *d* it is not recorded, but when it is above *d* it is recorded at its observed value.

  An observation is **truncated from above** (also called right truncated) at *u* if when it is above *u* it is not recorded, but when it is below *u* it is recorded at its observed value.

  An observation is **censored from below** (also called left censored) at *d* if when it is below *d* it is recorded as being equal to *d*, but when it is above *d* it is recorded at its observed value.

  An observation is **censored from above** (also called right censored)at *u* if when it is above *u* it is recorded as being equal to *u*, but when it is below *u* it is recorded at its observed value.

- **Comments**:

  Truncation - In insurance, truncation from below can happen when there is a deductible:

  - **Ordinary deductible:** claims are paid on excess of the deductible.
  - **Franchise deductible:** claims greater than the deductible are paid by total claim amount.

  In both cases a policyholder will not report a claim whose value is below the deductible. However, the knowledge of “small” claims (number and amounts) can be important for a correct evaluation of the policy risk.

- **Censoring** – Let y be the "correct" value, c the censoring point and x the available data.
  - Censoring from below $x = \begin{cases} c & y \leq c \\ y & y > c \end{cases}$
  - Censoring from above $x = \begin{cases} y & y < c \\ c & y \geq c \end{cases}$
  - In insurance **censoring from above is quite usual**. If a policy pays no more than 10000 € for a claim and if the insurance company only records the payments made, any time a loss is above 10000 € the amount of the claim will be unknown but we will know that a payment of 10000 € has happened.
  - The censoring points could be known (defined by the insurance policy) or "random". Random censoring occurs for instance when a policyholder decides to surrender his policy (data set D1). In any case we will know the censoring points that can differ from observation to observation.

○ From a statistical point of view, **truncation is a more severe limitation than censoring**.

○ When nothing else is said, truncation will mean left truncation and censoring right censoring.

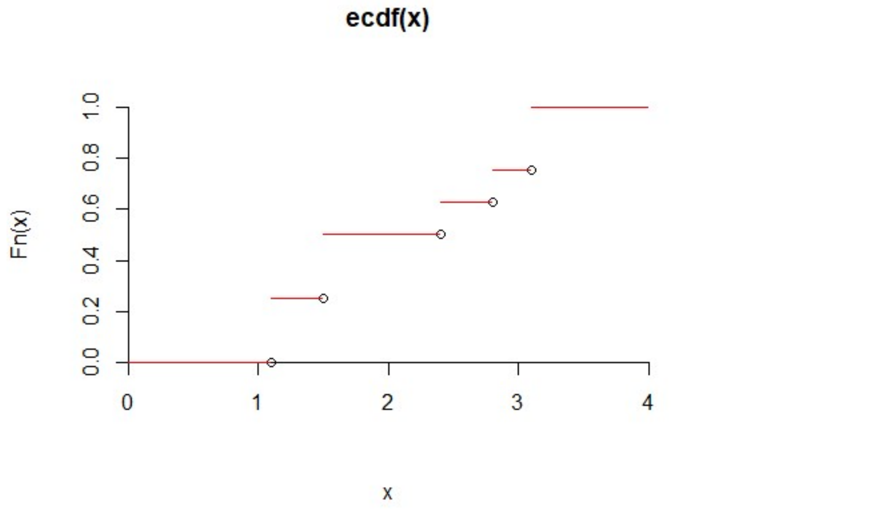
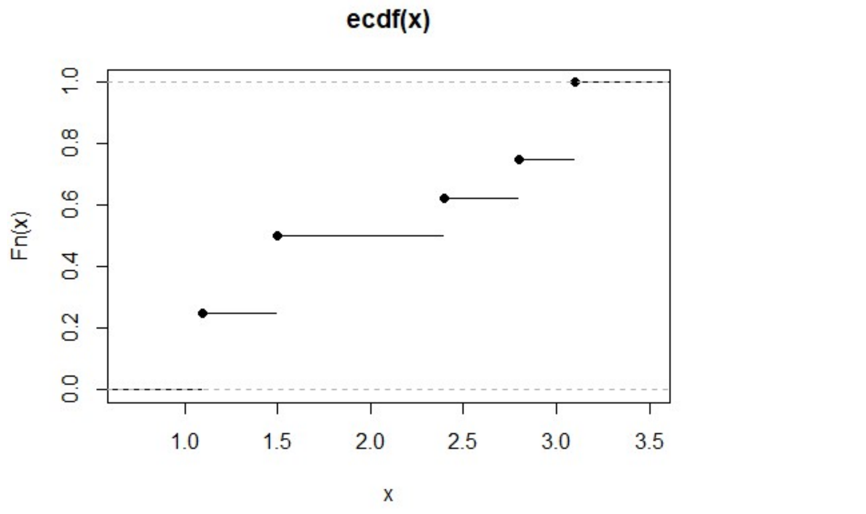
The empirical distribution for complete individual data

- Let us define the indicator function of a set A by $I_A(x) = I(x \in A) = \begin{cases} 0 & x \notin A \\ 1 & x \in A \end{cases}$
- Now, let us assume that a sample of size *n*, $(x_1, x_2, \cdots, x_n)$, from a given population has been observed
- **Definition 11.5 (13.5)** – The empirical distribution function (also known as empirical cumulative distribution function or ecdf) is

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

- Comments:
  1. 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. We will return to this topic later, when discussing KERNEL estimation.
  2. If our focus is the survival function, we can define $S_n(x) = 1 - F_n(x)$;

- **Example**: Define the empirical cumulative distribution function when the following random sample has been observed (1.1; 1.1; 2.8; 1.5; 2.4; 1.5; 3.1; 3.1)

	
Open circle points	**Closed circle points**

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

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

- **Example**: using the previous example we get

x	1.1	1.5	2.4	2.8	3.1
$f_n(x)$	2/8	2/8	1/8	1/8	2/8

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

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

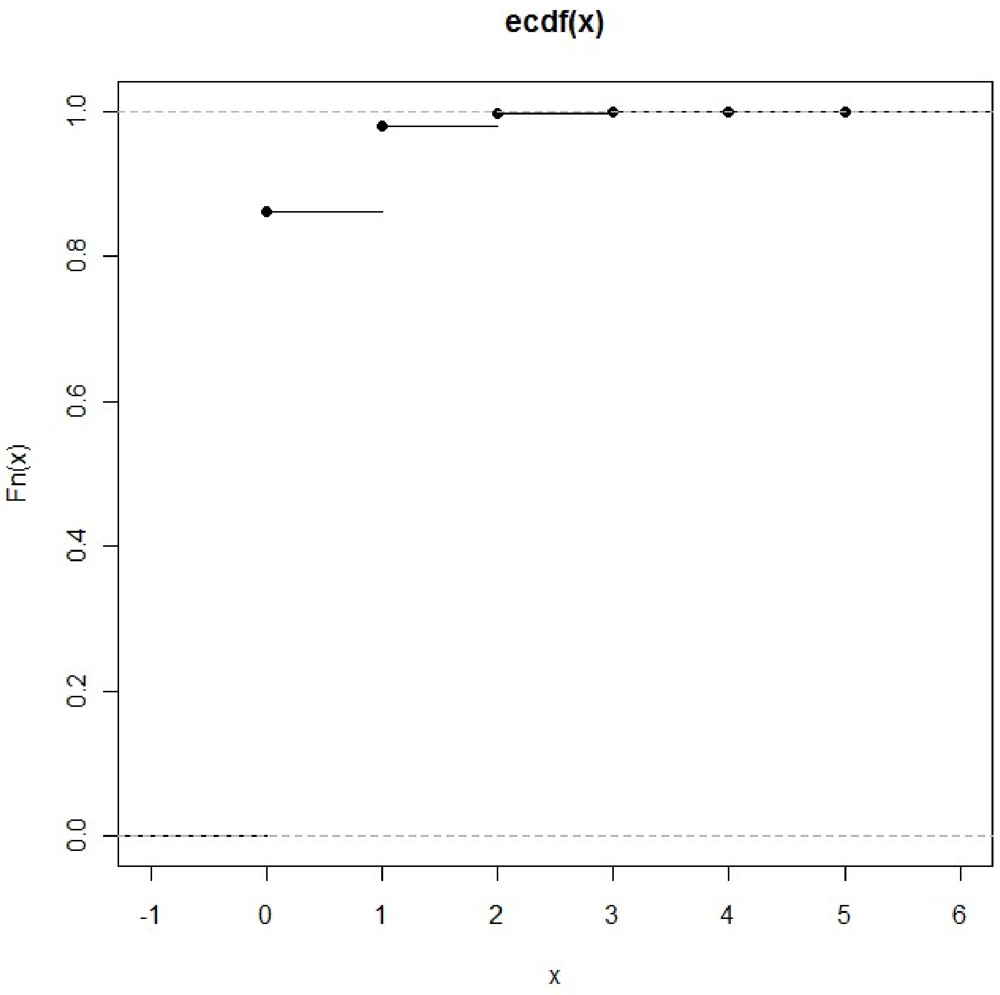
- **Example 11.1** **(13.1)** – Provide the empirical distribution functions for the data in data sets A and B. For data set A also provide the empirical probability function. For data set A assume that all seven drivers who had five or more accidents had exactly five accidents.

Data Set A

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

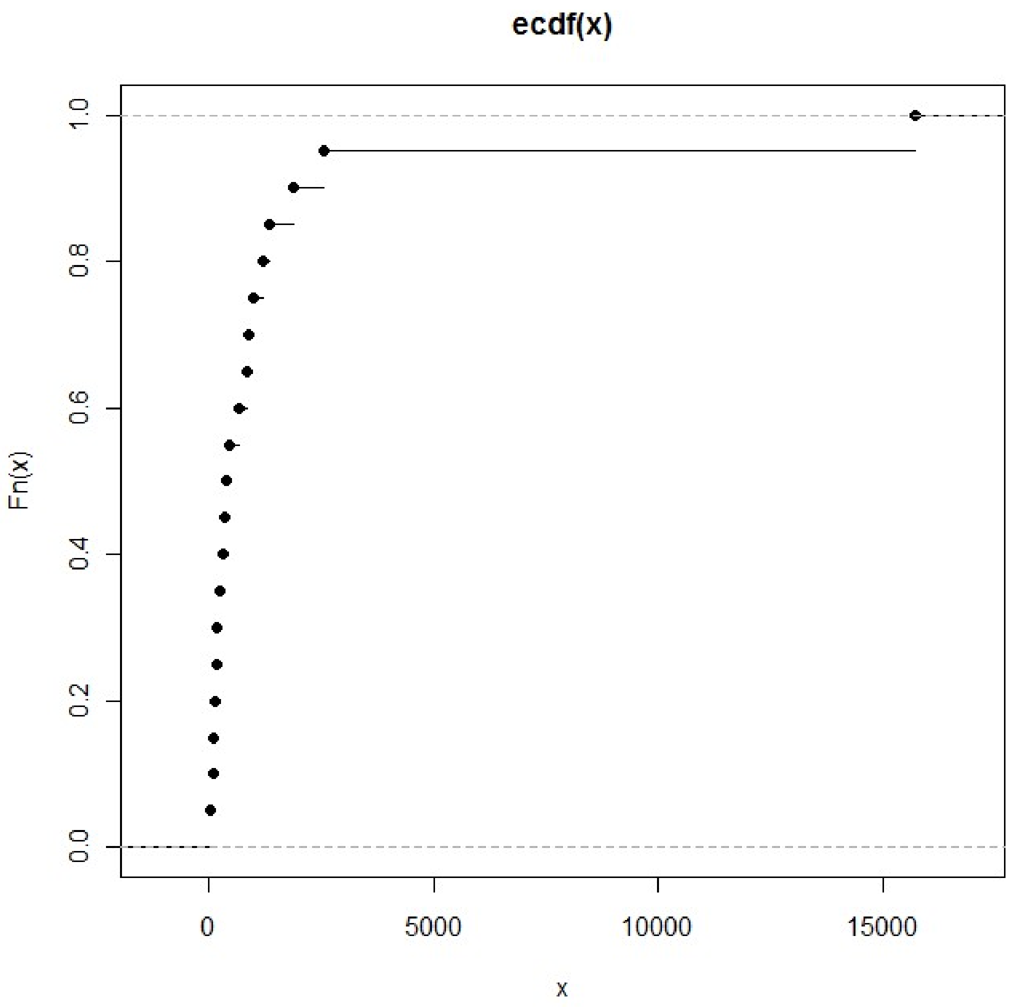
Total number
of observations 94935

Number of accidents per year per policy
1956-1958 – Dropkin paper



Data Set B - Amounts paid on Workers Compensation medical benefits – artificial data

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



Data set B - Empirical distribution function using R

```
> # 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) # Gives the mean and the 5 numbers summary
                 To be used only if all values in x are unique!!!
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 
> quantile(F20,c(0.25,0.5,0.75))
    25%     50%     75% 
 159.50  420.50 1028.75 
>  plot(F20)
```

Data Set A - Empirical distribution function using R

```
> # 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
> # function tapply: apply the function (sum in our case) to each
group of element of z. The groups are defined using the factor x
> 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
```

Empirical distribution for grouped data

- What are grouped data? Grouped data and censoring.
- For grouped data it is not possible to construct the empirical distribution function. The main idea is to approximate it using an intuitive approach:
  - 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). When using the linear interpolation, we are assuming a uniform behavior inside each group.

- 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 $F_n(c_0) = 0$. Then $F_n(x) = F_n(c_{j-1}) + \frac{x-c_{j-1}}{c_j-c_{j-1}}\left(F_n(c_j) - F_n(c_{j-1})\right)$ for $c_{j-1} < x < c_j$.

- 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.

- **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

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

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

- **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. Data set C is a random sample of payments from 227 claims from a general liability insurance. Data is classified by payment range.

Payments	0-7500	7500- 17500	17500- 32500	32500- 67500	67500- 12500	125000- 300000	>300000
Nº policies	99	42	29	28	17	9	3

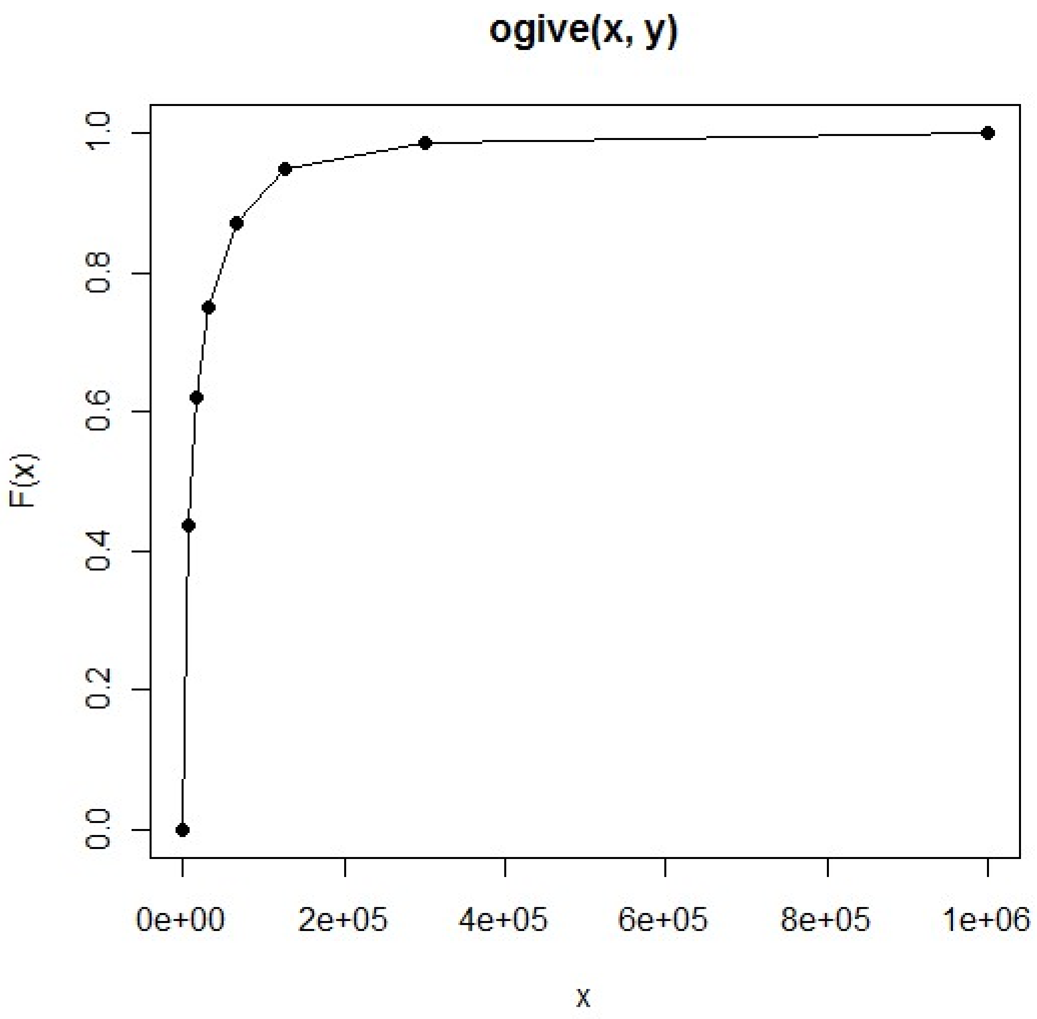
Use R and **actuar** library to define the empirical distribution function and the histogram

Challenging questions:

- Can you do it without using actuar library?
- Are you able to write a function like ogive?

Using library actuar

```
> # reading 1000000 is arbitrarily chosen to replace Inf
> x=c(0,7500,17500,32500,67500,125000,300000,1000000) # breaks
> y=c(99,42,29,28,17,9,3)  # counts
>
> library(actuar)  # should have been installed before
Attaching package: ‘actuar’
…
> # using function ogive
> Fn=ogive(x,y)
> Fn
Ogive for grouped data
Call: ogive(x = x, y)
    x =      0,   7500,  17500,  ...,  3e+05,  1e+06
 F(x) =      0, 0.43612, 0.62115,  ..., 0.98678,      1
> plot(Fn)
> Fn(1000); Fn(7500); Fn(300000); Fn(302000); Fn(1050000)
[1] 0.05814978
[1] 0.4361233
[1] 0.9867841
[1] 0.9868219
[1] 1
```

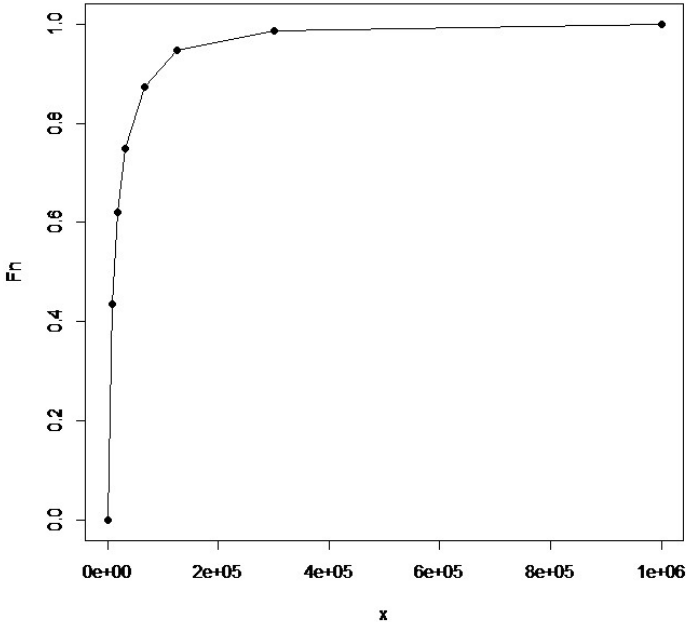

ogive(x, y)
F(x)
0.0
0.2
0.4
0.6
0.8
1.0
0e+00
2e+05
4e+05
6e+05
8e+05
1e+06
x

Using library actuar (cont) - Now repeat the example using
`x=c(0,7500,17500,32500,67500,125000,300000,Inf)` and analyze the main differences.

Without using library actuar - challenge 1:

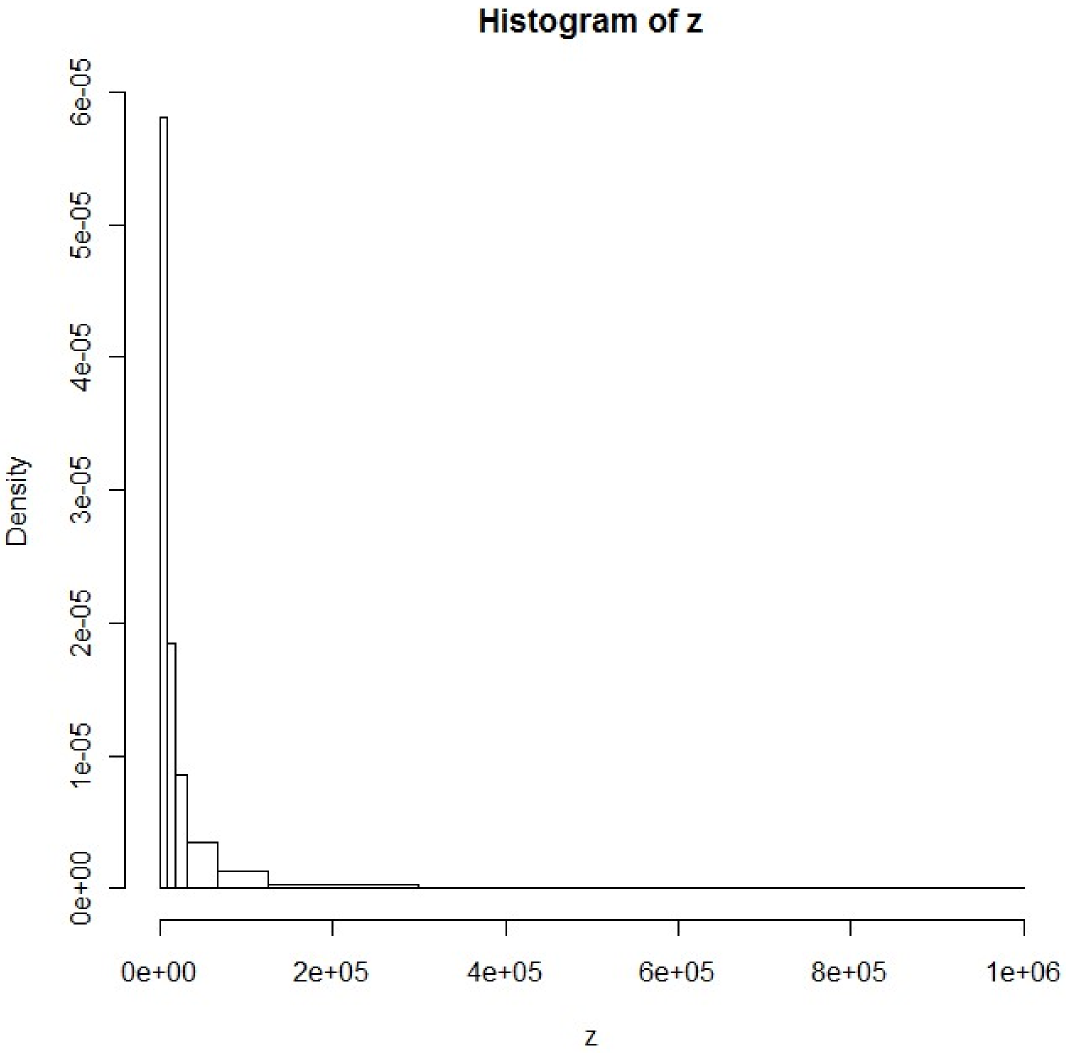
```
> lb=x[-length(x)]; ub=x[-1]
> a=cumsum(y)/sum(y); # ecdf at the boundaries
> lF=c(0,a[-length(a)]); #Fn(c_j-1)
> uF=a                   #Fn(c_j)
> lb; ub; lF; uF
[1]      0   7500  17500  32500  67500 125000 300000
[1]    7500   17500   32500   67500  125000  300000 1000000
[1] 0.0000000 0.4361233 0.6211454 0.7488987 0.8722467 0.9471366
0.9867841
[1] 0.4361233 0.6211454 0.7488987 0.8722467 0.9471366 0.9867841
1.0000000
>
> #see formula (slide 14)
> intercept=(ub*lF-lb*uF)/(ub-lb)
> slope=(uF-lF)/(ub-lb)
>
> ogive_table=data.frame(lower_bound=lb,upper_bound=ub,
+                        intercept=intercept,slope=slope)
```

```
> ogive_table
  lower_bound upper_bound intercept        slope
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     1000000 0.9811202 1.887980e-08
>
> Fn=c(0,uF)
> # plotting the ogive
> plot(x,Fn,type="l")
> points(x,Fn,pch=16)
```

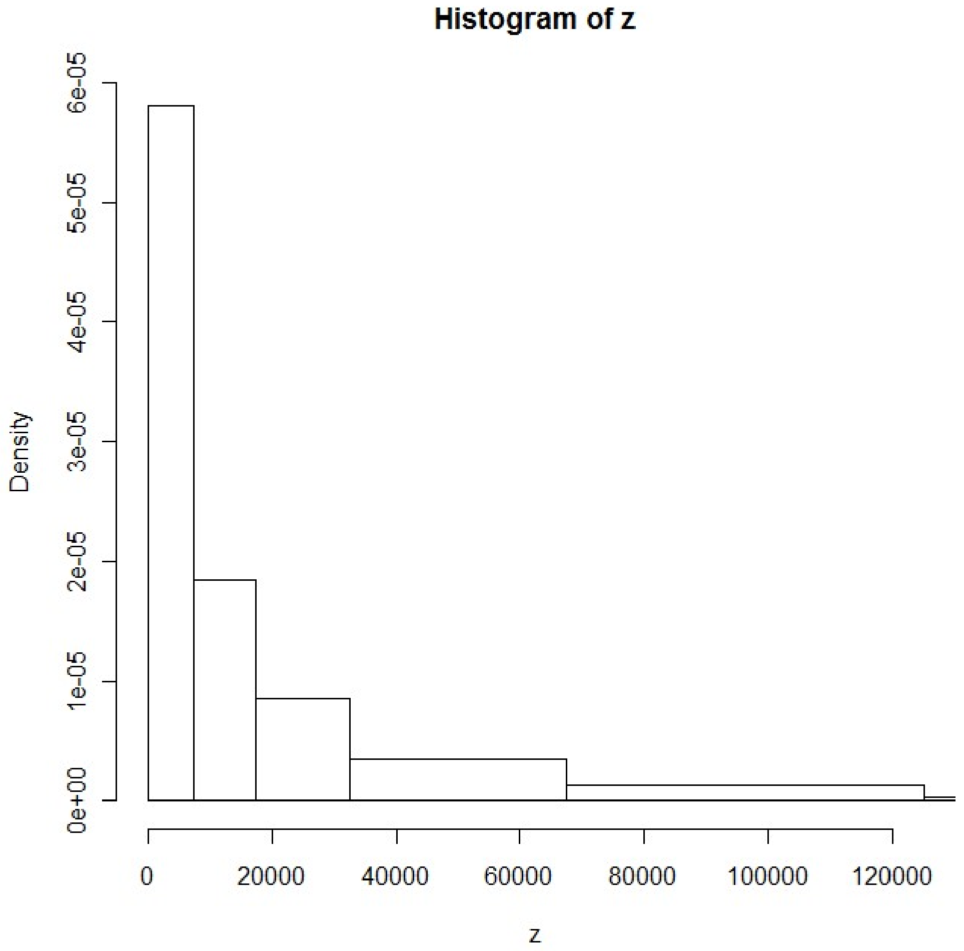


```
# The empirical density is given by the slope(column 4 of ogive table)
# To plot a histogram we need to "simulate" the observations
# As Only the number of observations in each interval matters,
# we choose an arbitrarily value inside each interval and define array z
z=c(rep(5000,99),rep(10000,42),rep(20000,29),rep(50000,28),
    rep(70000,17),rep(150000,9),rep(400000,3))

hist(z,breaks=x)   # Be careful, you need a finite limit for 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 survival function as

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

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 unconditional probabilities 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

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

  where $N_{(a,b]}$ is the number of observations 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.

  - Estimate: $\overset{\wedge}{\Pr}(a < X \leq b) = S_n(a) - S_n(b) = \dfrac{n_a - n_b}{n} = \dfrac{n_{(a,b]}}{n}$
  - Expected value of the estimator:

    $$E\left(\overset{\wedge}{\Pr}(a < X \leq b)\right) = E\left(\frac{N_{(a,b]}}{n}\right) = \frac{n\left(S(a) - S(b)\right)}{n} = \Pr(a < X \leq b)$$ Unbiased

  - Variance of the estimator:

    $$\operatorname{var}\left(\overset{\wedge}{\Pr}(a < X \leq b)\right) = \operatorname{var}\left(\frac{N_{(a,b]}}{n}\right) = \frac{n\left(S(a) - S(b)\right)\left(1 - \left(S(a) - S(b)\right)\right)}{n^2}$$

    $$= \frac{\left(S(a) - S(b)\right)\left(1 - \left(S(a) - S(b)\right)\right)}{n}$$

- Example 1 – Consider data set B and estimate Pr $(X > 1000)$. Assuming that 20 is a large sample (which is not), define a 95% confidence interval for this probability.

Estimator → $\widehat{Pr}(X > 1000) = \frac{N_{(1000,\infty)}}{20}$

Estimate → $\widehat{Pr}(X > 1000) = \frac{n_{(1000,\infty)}}{20} = \frac{5}{20} = 0.25$

Variance of the estimator → $var\left(\widehat{Pr}(X > 1000)\right) = var\left(\frac{N_{(1000,\infty)}}{20}\right) = \frac{S(1000)\,(1-S(1000))}{20}$

Estimator of the variance of the estimator:

$$\widehat{var}\left(\widehat{Pr}(X > 1000)\right) = \widehat{var}\left(\frac{N_{(1000,\infty)}}{20}\right) = \frac{\left(\frac{N_{(1000,\infty)}}{20}\right)\left(1-\frac{N_{(1000,\infty)}}{20}\right)}{20}$$

Estimate of the variance of the estimator:

$$\widehat{var}\left(\widehat{Pr}(X > 1000)\right) = \frac{\left(\frac{n_{(1000,\infty)}}{20}\right)\left(1-\frac{n_{(1000,\infty)}}{20}\right)}{20} = \frac{0.25\times 0.75}{20} = \frac{3}{320} = 0.009375$$

95% CI: (0.0602, 0.4398) $0.25 \pm 1.96\,\sqrt{0.009375}$. The width is quite large due to the sample size.

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

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

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 ${}_{y-x}\hat{q}_x = \dfrac{N_x - N_y}{N_x}$. **This estimator does 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

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

The estimate of the variance is $\mathrm{v\hat{a}r}\left({}_{y-x}\hat{q}_x \mid S(x) = S_n(x)\right) = \dfrac{1}{n_x^3}\, n_y \left(n_x - n_y\right)$

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.

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

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

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

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

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

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

- **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 claim amount. 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

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

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

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.

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

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

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

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

□ Confidence interval: $$\frac{2\,n\,p_n(2) + 1.96^2 \pm 1.96\sqrt{1.96^2 + 4\,n\,p_n(2) - 4\,n\,p_n(2)^2}}{2\left(n + 1.96^2\right)}$$, 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

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

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

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

- 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 for the variance of the estimators

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

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

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

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

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

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

KERNEL DENSITY MODELS

- Although the empirical distribution converges to the distribution of the random variable, as $n \to \infty$, a main point remains: for finite samples the empirical distribution is always discrete, even if the underlying variable is **continuous**. This problem is more annoying when the sample size is moderate.
- Our aim is to smooth, using nonparametric methods (i.e. ignoring the functional form of the density), the empirical distribution to obtain an estimate of the continuous density (or distribution) function.

- **Definition 12.2 (14.2)** – A kernel density estimator of a distribution function is

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

And the estimator of the density function is

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

The function $k_y(x)$ is called the **kernel**.

- **Comments**
  - The kernel is a non-negative real-valued integrable function satisfying $\int_{-\infty}^{+\infty} k_y(x)\,dx = 1$ to guarantee that the kernel method originates a density function. We will also have, $K_y(x) = \int_{-\infty}^{x} k_y(u)\,du$.

    □Question: How can we guarantee that $\hat{f}(x)$ is a density function?
  - In much cases we impose that $\int_{-\infty}^{+\infty} x\,k_y(x)\,dx = y$, that is the expected value is unchanged by the kernel.
  - $p(y_j)$ is the probability assigned to the value y_j, $j = 1,2,\cdots,k$, by the empirical distribution. : If all the sample values are unique we get $p(y_j) = 1/n$ and then $\hat{F}(x) = \sum_{i=1}^{n} (1/n) K_{x_i}(x)$ and $\hat{f}(x) = \sum_{i=1}^{n} (1/n) k_{x_i}(x)$ respectively.

- **Definition 12.3** **(14.3)** (using a different notation)
  - Uniform kernel:

$$k_y(x)=(2b)^{-1}I\left(|x-y|\le b\right)=(2b)^{-1}I\left(y-b\le x\le y+b\right)=\begin{cases}0 & x<y-b\\ 1/(2b) & y-b\le x\le y+b\\ 0 & x>y+b\end{cases}$$

  - Triangular kernel: $k_y(x)=\dfrac{b-|y-x|}{b^2}I(|y-x|/b\le 1)=\begin{cases}0 & x<y-b\\ (x-y+b)/b^2 & y-b\le x\le y\\ (y+b-x)/b^2 & y\le x\le y+b\\ 0 & x>y+b\end{cases}$

  - Gamma kernel: $k_y(x)=\dfrac{x^{\alpha-1}e^{-x\alpha/y}}{(y/\alpha)^{\alpha}\,\Gamma(\alpha)}I_{(0;+\infty)}(x)$

    Gamma density with mean y and variance y^2/α. The lesser α the smoother the kernel.

    How to choose α? One can use $\alpha=\sqrt{n}\sqrt{(\hat{\mu}'_4/\hat{\mu}'^2_2)-1}$ (Typo in the book)

    Remember that $\hat{\mu}'_k=\sum y_j^k\,p(y_j)$

- Comments:
  - *b* is called the bandwidth . The higher is *b* the smoother will be the kernel density.
  - The first and second kernels are symmetric around *y.* In symmetric kernels the bandwidth is usually much more important than the choice of a particular kernel.
  - The third kernel is asymmetric and α plays a role similar to the bandwidth. Note that the gamma kernel can be used only with positive valued random variables.

- How to get $K_y(x)$?
  - $K_y(x) = \int_{-\infty}^{x} k_y(u)\,du$
  - For example in the uniform case,

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

- In the remaining of the course we will follow Definition 12.2 (14.2). However, this is not the standard definition of a kernel density estimator. For a standard presentation, see Wasserman (2004).

  A kernel is any smooth function K such that $K(x) \geq 0$, $\int_{-\infty}^{+\infty} K(x)\,dx = 1$, $\int_{-\infty}^{+\infty} x\,K(x)\,dx = 0$ and

  $\sigma_K^2 = \int_{-\infty}^{+\infty} x^2 \; K(x)\,dx < \infty$.

  Given a kernel K and a positive number h, called the bandwidth, the kernel density estimator is defined to be $\hat{f}_n(x) = \sum_{i=1}^{n} \frac{1}{n}\frac{1}{h} K\left(\frac{x - X_i}{h}\right)$.

  Examples of kernels are:

  - The Gaussian kernel: $K(u) = (2\pi)^{-1/2}\, e^{-u^2/2}$
  - The Epanechnikov kernel: $K(u) = \frac{3}{4 \times \sqrt{5}}\left(1 - \frac{u^2}{5}\right) I\left(|u| < \sqrt{5}\right)$
  - The uniform kernel: $K(u) = \frac{1}{2} I\left(|u| \leq 1\right)$
  - The triangular kernel: $K(u) = \left(1 - |u|\right)\; I\left(|u| \leq 1\right)$

All these kernels act symmetrically around each sample point. In this setup the choice of a particular kernel is generally much less important than the choice of the bandwidth. They are methods to approximate the “best” choice of the bandwidth (see Wasserman (2004)).

- **Example 12.13 (14.16)** – Determine the kernel density estimate for Example 11.2 (13.2) using each of the three kernels.

  We will only discuss the uniform kernel with *b*=0.5 (results are presented for *b*=1.0 and *b*=0.1).

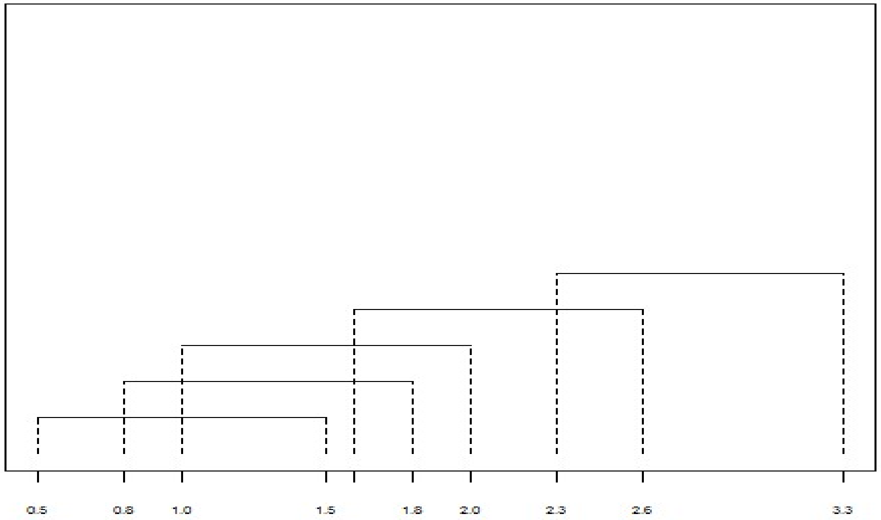
  Sample $(1.0;1.3;1.5;1.5;2.1;2.1;2.1;2.8)$

y_j	1.0	1.3	1.5	2.1	2.8
$p(y_j)$	1/8	1/8	2/8	3/8	1/8

Bandwith *b*=0.5 then $1/(2b)=1$

1.0	→	0.5	1.5
1.3		0.8	1.8
1.5		1.0	2.0
2.1		1.6	2.6
2.8		2.3	3.3

$$\hat{f}(x)=\begin{cases} 1/8 & 0.5<x<0.8 \\ 2/8 & 0.8<x<1.0 \\ 4/8 & 1.0<x<1.5 \\ 3/8 & 1.5<x<1.6 \\ 6/8 & 1.6<x<1.8 \\ 5/8 & 1.8<x<2.0 \\ 3/8 & 2.0<x<2.3 \\ 4/8 & 2.3<x<2.6 \\ 1/8 & 2.6<x<3.3 \\ 0 & \text{otherwise} \end{cases}$$



Bandwith *b*=0.1 then $1/(2b)=5$

1.0	→	0.9	1.1
1.3		1.2	1.4
1.5		1.4	1.6
2.1		2.0	2.2
2.8		2.7	2.9

$$\hat{f}(x)=\begin{cases} 5/8 & 0.9<x<1.1 \\ 5/8 & 1.2<x<1.4 \\ 10/8 & 1.4<x<1.6 \\ 15/8 & 2.0<x<2.2 \\ 5/8 & 2.7<x<2.9 \\ 0 & \text{otherwise} \end{cases}$$

Discuss the problem related to the intervals limit

Bandwith *b*=1.0 then $1/(2b)=0.5$

1.0	→	0.0	2.0
1.3		0.3	2.3
1.5		0.5	2.5
2.1		1.1	3.1
2.8		1.8	3.8

$$\hat{f}(x)=\begin{cases} 1/16 & 0<x<0.3 \\ 2/16 & 0.3<x<0.5 \\ 4/16 & 0.5<x<1.1 \\ 7/16 & 1.1<x<1.8 \\ 8/16 & 1.8<x<2.0 \\ 7/16 & 2.0<x<2.3 \\ 6/16 & 2.3<x<2.5 \\ 4/16 & 2.5<x<3.1 \\ 1/16 & 3.1<x<3.8 \\ 0 & \text{otherwise} \end{cases}$$

Using R

```
y=c(1.0,1.3,1.5,2.1,2.8); s=c(1,1,2,3,1); n=sum(s)
p_y=s/n
x=seq(0,4,by=0.025); fx=rep(NA,length(x))

# Uniform kernel
b=0.5; LU=y-b; UU=y+b
for(i in 1:length(x)) fx[i]=sum(p_y*dunif(x[i],LU,UU))
label.plot=paste("example 12.13 - Uniform kernel with b=",toString(b),sep="")
plot(x,fx,type="l",main=label.plot)

# Gamma kernel
alpha=50
for(i in 1:length(x)) fx[i]=sum(p_y*dgamma(x[i],shape=alpha,scale=y/alpha))
label.plot=paste("example 12.13 - Gamma kernel with alpha=",toString(alpha),sep="")
plot(x,fx,type="l",main=label.plot)
```

- **Example 1 (New)** – Using the same data as before, estimate $f(2)$ and $F(2)$ using a normal kernel with $\sigma = 0.3$.

  Sample $(1.0;1.3;1.5;1.5;2.1;2.1;2.1;2.8)$

y_j	1.0	1.3	1.5	2.1	2.8
$p(y_j)$	1/8	1/8	2/8	3/8	1/8

$k_{y_j}(x) = \frac{1}{\sigma\sqrt{2\pi}}\, exp\left(-\frac{(x-y_j)^2}{2\,\sigma^2}\right)$ and then $k_{y_j}(2) = \frac{1}{0.3\sqrt{2\pi}}\, exp\left(-\frac{(2-y_j)^2}{0.18}\right)$

$$\hat{f}(2) = \sum_{j=1}^{5} p(y_j)\, k_{y_j}(2) = \frac{1}{8}\,\frac{1}{0.3\sqrt{2\pi}}\, exp\left(-\frac{(2-1)^2}{0.18}\right) + \cdots + \frac{1}{8}\,\frac{1}{0.3\sqrt{2\pi}}\, exp\left(-\frac{(2-2.8)^2}{0.18}\right)$$

=0.570943

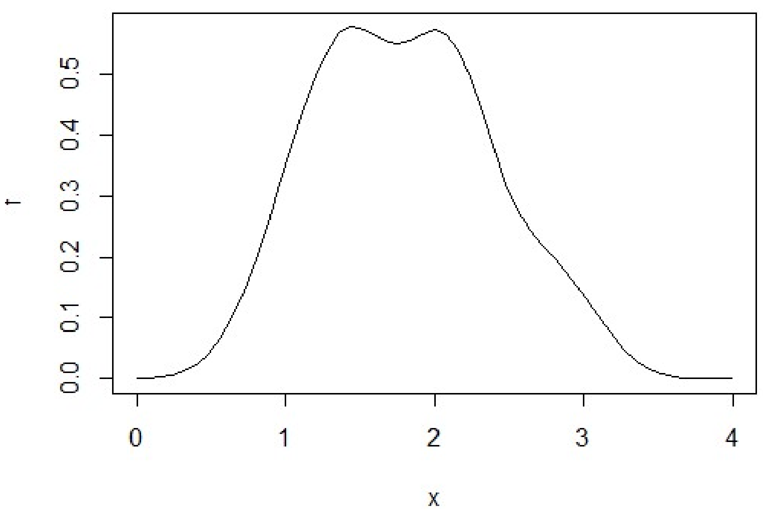
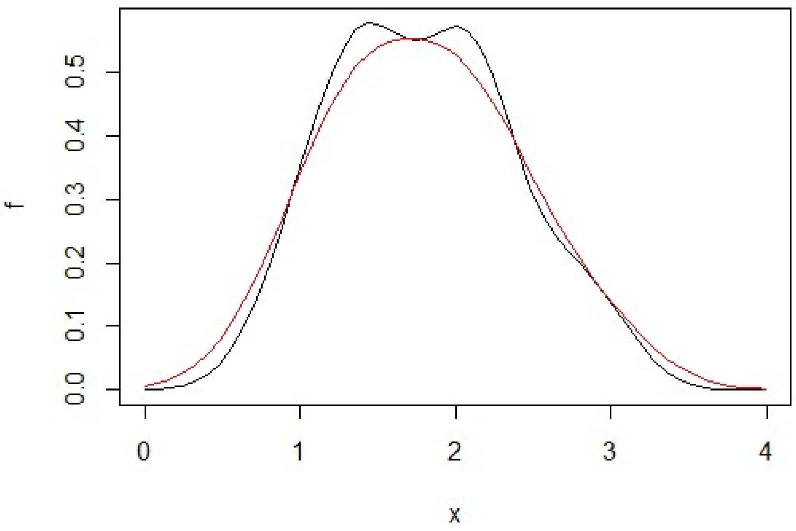
$K_{y_j}(x) = F(2|\mu = y_j, \sigma = 0.3)$ where F is the distribution function of a normal rv with mean μ and standard deviation 0.3.

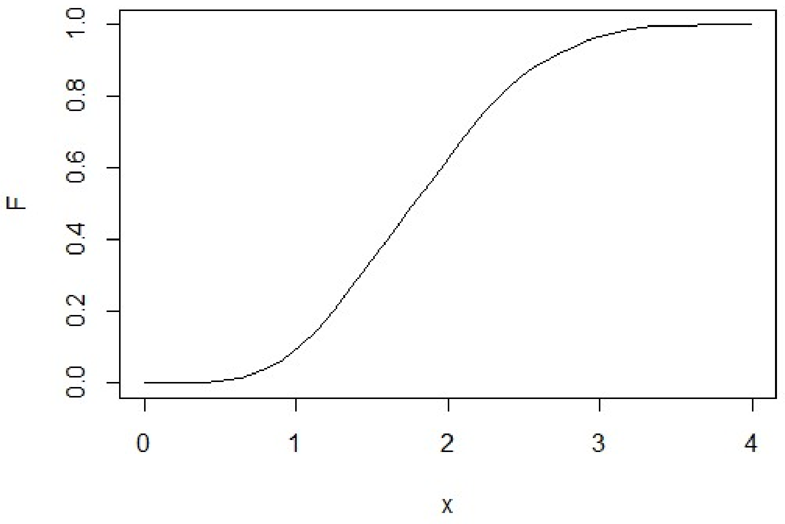
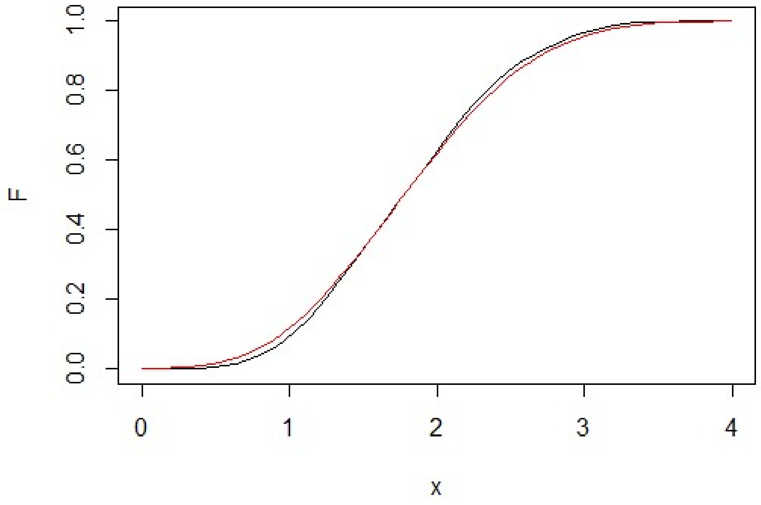
$$\hat{F}(2) = \sum_{j=1}^{5} p(y_j)\, K_{y_j}(2) = \frac{1}{8}\, F(2|\mu = 1.0, \sigma = 0.3) + \cdots + \frac{1}{8} F(2|\mu = 1.0, \sigma = 0.3)$$

R code

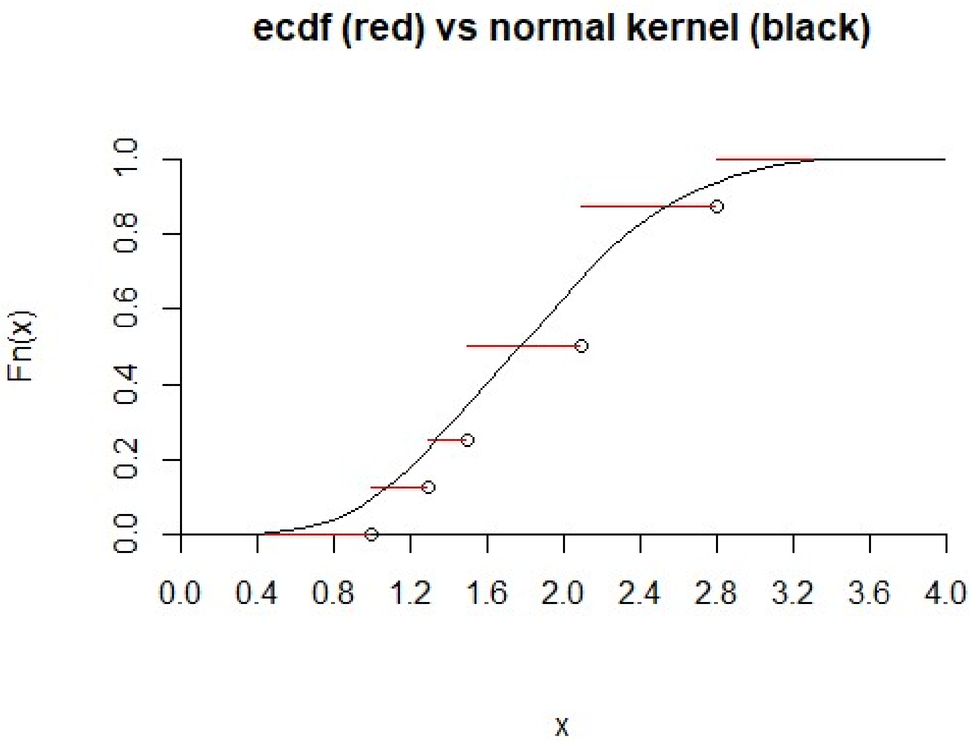
```
> y=c(1,1.3,1.5,2.1,2.8); p.y=c(1/8,1/8,2/8,3/8,1/8)
>
> k_y=dnorm(2,y,0.3)
> f_=sum(p.y*k_y)
> f_
[1] 0.5709434
> K_y=pnorm(2,y,0.3)
> F_=sum(p.y*K_y)
> F_
[1] 0.6257912
>
```

Using R, we can do a few plots

	
Estimated density function Normal kernel with $\sigma = 0.3$	Comparing estimated densities $\sigma = 0.3$ (black) versus $\sigma = 0.4$ (red)

	
Estimated distribution function Normal kernel with $\sigma = 0.3$	Comparing estimated distribution functions $\sigma = 0.3$ (black) versus $\sigma = 0.4$ (red)

Comparing the kernel with the ecdf


ecdf (red) vs normal kernel (black)
Fn(x)
x