1.1 Introduction

In the main text and computer programs I will use the symbols/variables for each variable name as listed in APPENDIX 01, unless otherwise stated.

"In survival analysis, interest centres on a group or groups of individuals for each of whom (or which) there is defined a point event, often called failure, occuring after a length of time called the failure time. Failure can occur at most once on any individual.

To determine failure time precisely the time origin must be defined, the event for measuring the time scale must be agreed and the meaning of failure must be clear." (CO, 1984)

I will assume that these requirements are satisfied in all my computer programs that involve simulation (see section 1.5) and these requirements will be described in section 1.2.

I will be studying single groups, but later on I will be looking at 2 or more groups. In the next chapters, I will write and use Fortran 77 programs to demonstrate the various methods and/or models in survival analysis, eg estimation, product limit estimator, etc.

"Survival analysis is univariate, as there is a single response variable, failure time, even though there are many other explanatory variables." (CO, 1984)

Examples of explanatory variables are censor times and indicator variables for censoring, etc. Sometimes, in many scientific and medical applications, failure time is plotted on the x-axis and physical quantities/statistics (eg heart rate, survivor function, body temperature, etc) are plotted on the y-axis.

1.2 The definition of failure time

"We now comment briefly on the requirements for measuring failure time.

The time origin should be precisely defined. It is also desirable that, subject to any known differences on explanatory variables, all individuals should be as comparable as possible to their time origin.

The time origin need not be and usually is not at the same calendar time for each individual. Most clinical trials have staggered entry, so that the patients enter over a substantial time period. Each patient's failure time is usually measured from his own date of entry." (CO, 1984)

I will assume that this is so in all my computer programs when I do the simulation (see section 1.5). Figure 1.1 illustrates the calculation.

```
-   ........-           -.......o
-.......                -.......*
-     ....              -....*
____________________-____     __________________
A     B       C         0     5      10 time
   (a)                     (b)
```

Figure 1.1 - Experience of 3 individuals with staggered entry and follow up until C (where A=1975, B=1980 and C=1985)
* = death ; o = censoring
(a) Real time
(b) From entry into study (Cox and Oakes, 1984)

CHAPTER 01 - The scope of survival analysis

According to Cox and Oakes (1984), clock time (real time) is usually the 'scale' for measuring time. I will assume that time is measured in arbitary units in all my computer programs when I do the simulation (see section 1.5). "However, other possibilities certainly arise, such as the use of operating time of a system and the age of an individual. The only universal requirement for failure times is that they are nonnegative.

One reason for the choice of a timescale is direct meaningfulness for the individual concerned, justifying the use of real time in investigating survival in a medical context." (CO, 1984)

According to Cox and Oakes (1984), we should define the meaning of failure precisely. Table 1.1 gives meanings of failures in each subject area. Those in italics are taken from Cox and Oakes.

Subject area	Meaning of failure(s)
Medicine/Science	Attenuation of an X-ray beam Capacitor discharging Decaying of radioactive nuclei
Industry *(EG Car manufacturing)*	*Performance measured in some quantitative way falls below an acceptable level.* EG fuel economy in mpg.
Some applications	*Little or no importance*

Table 1.1 - Examples of meanings of failure

"There will be some arbitariness in the definition of failure and it will be for consideration whether to concentrate on the failure time or whether to analyse the whole performance measure as a function of time." (CO, 1984)

1.3 Censoring

"A special source of difficulty in the analysis of survival data is the possibilty that some individuals may NOT be observed for the full time to failure. At the close of a life-testing experiment in industrial reliabilty, NOT all components may have failed. Some patients (many, it is hoped) will survive to the end of a clinical trial. A patient who has died from heart disease cannot go on to die from lung cancer.

An individual who is observed, failure-free, for 10 days and then withdrawn from study has a failure time which must exceed 10 days. Such incomplete observation of the failure time is called censoring. Note that, like failure censoring is a point event and must be recorded.

We suppose that, in the absence of censoring, the ith individual in a sample of n has failure time T_i, a random variable. We suppose also that there is a period of observation c_i such that observation on that individual ceases at c_i if failure has not occured by then.
Then the observations consist of $x_i=\min(T_i,c_i)$." (CO, 1984)
Cox and Oakes (1984) suggests two main types of censoring that could be useful in this project, called Type I and Type II.
In Type I censoring, according to Cox and Oakes (1984), c_i=constant for all values of i, where i=1,n.
It may look something like this as shown in Figure 1.2.

```
-.......*    +
-...*        +          + = censor point
-...........+..o        o = censored
-...........+.....o     * = death
-......*     +
_____________________
0          Time
```

Figure 1.2 - Typical example of Type I censoring

In Type II censoring, according to Cox and Oakes (1984), the censor time is allowed to be a random variable.
It may look something like this as shown in Figure 1.3.

```
-...+...o
-..........*  +         + = Point at which censoring occurs
-.......+...o           o = censoring
-.+...o                 * = death
-........*   +
____________________
0          Time
```

Figure 1.3 - Typical example of Type II censoring

1.4 How is the data obtained and why?

In practice, it is difficult to obtain real data and so simulation should be used. I used it because

(1) It is quicker to obtain the data
(2) It is cheaper

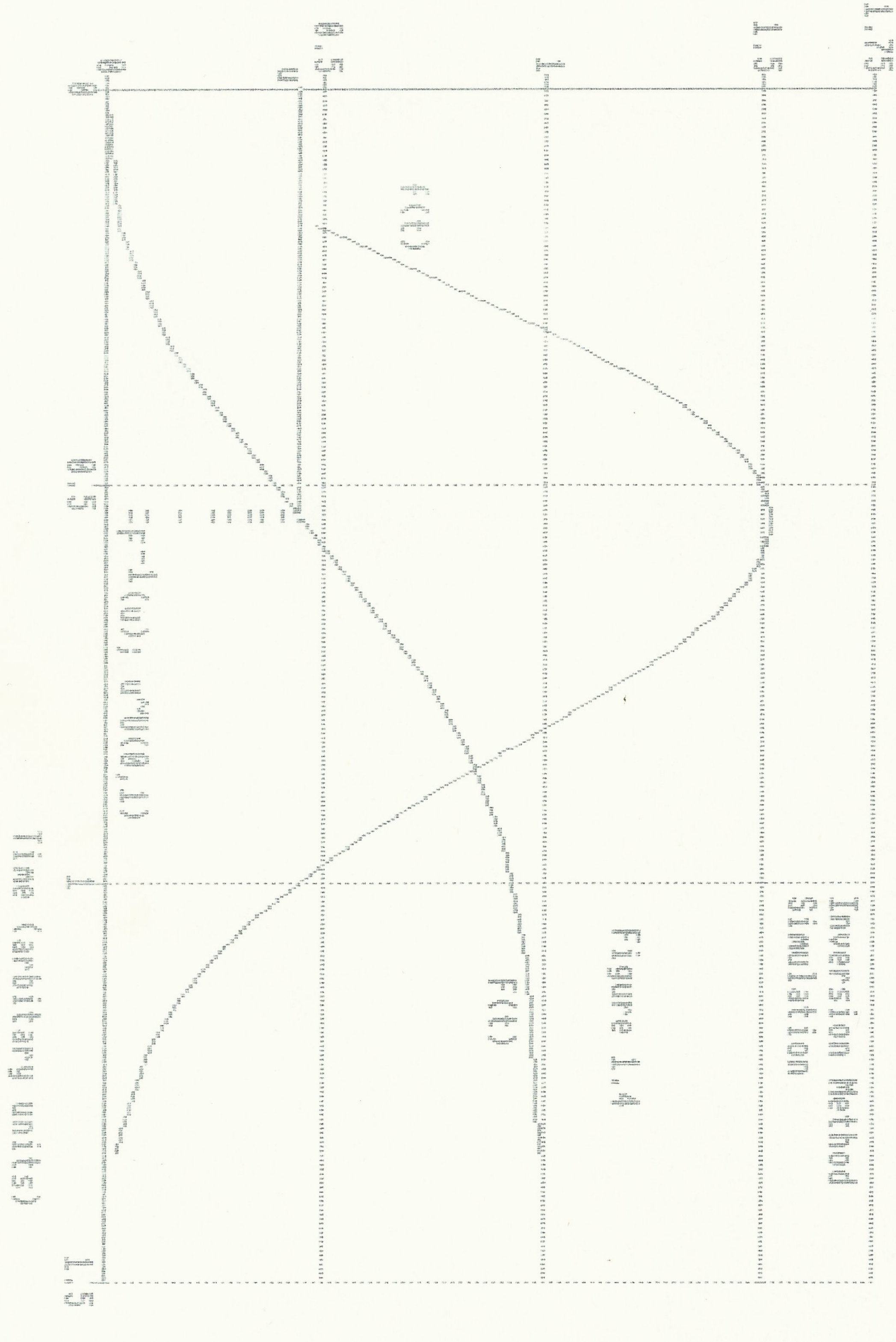
In my computer programs, I have simulated the Weibull distribution with index k and rate p for the remission time T. For the censor times (in realistic case II, details of this and other models in section 1.6), I have simulated an exponential distribution with rate p_1.

1.5 Simulation by what method?

I will be using the inversion method and formulae for the simulations described above. Figure 1.4 illustrates the inversion method simulation. In my FORTRAN 77 programs, I use a random number generator called URAND (described later) to obtain the cdf values and they will be used to obtain the times with its associated parameters of the probability distribution. For Weibull, the parameters are k(index) and p(rate).

Logic of simulation:

(1) Use a random number generator (URAND), that has its value $u \sim U(0,1)$ and use it to obtain the cdf value.
(2) Use this cdf value to obtain the times, using the parameters of the probability distribution.
(3) If any more individuals left to simulate, GOTO (1), else STOP.



1.5.1 The theory of 'URAND'

"Urand is a uniform random number generator, based on theory and suggestions given by KNUTH (1969). The integer iy should be initialised to an arbitary integer prior to the first call to urand. The calling program should not alter the value of iy between subsequent calls to urand. Values of urand will be returned in the interval (0,1). It has the following algorithm:

(1) If first entry, compute machine integer word length
(2) Compute multiplier and increment for linear congruential method
(3) Obtain scale factor for converting to floating point
(4) Compute next random number
(5) If computers allow integer overflow on addition iy=(iy-m2)-m2
(6) If the word length is greater than for multiplication iy=(iy-m2)-m2
(7) If integer overflow affects sign bit iy=(iy+m2)+m2" (Hahn, 1987)

1.5.2 Weibull distribution

The Weibull distribution has cdf $F(t)=1-\exp[-(pt)^k]$

$1-F(t)=\exp[-(pt)^k]$

$\ln[1-F(t)]=-p^k t^k$

$-\ln[1-F(t)]=p^k t^k$

$$t=[(-\ln(1-F(t)))/p^k]^{1/k} \quad \ldots(1.1)$$

Putting k=1, we have the exponential distribution.
In later work, especially accelerated life models, the rate parameter will no longer be constant and will be replaced by p_i, which is dependent on the explanatory variables. That is, $p_i=\exp(b_0+b_1z_1+\ldots)$

However, I can still use (1.1), although in my programs I have to declare the rate parameter as an ARRAY.

1.6 Models

I will be considering the following models :-

No censoring
Constant censoring
Realistic case I
Realistic case II

1.6.1 No censoring

$$x_i=T_i \text{ for all } i, \text{ where } i=1,n$$

1.6.2 Constant censoring

$$x_i=T_i \text{ if } T<\text{constant}$$
$$x_i=\text{Constant if } T>=\text{constant}$$

(Constant can be anything, eg mean of data, etc. The larger the constant, the more likely I get the No censoring model as described in section 1.6.1..)
This is an example of Type I censoring (see Figure 1.2).

1.6.3 Realistic case I

```
  $       .                 o
  $      .                  o $=Start of trial
  $     +------------x      o
  $+----------x-------------o
****************************** time
i 1 3 5 7 9             Stop

n -      *   -
  -     *    -
  -    *ooo>-
i -  *       -
-------------
-    A  B      AB is recruit

  ci  - *
      -   *
   c  -     *
      ---------
       1 2 n  i
```

At the start of trial, each individual is allowed to fail for some time, until the 'cut off' point is reached.

c_i=(constant)+((n-i)/(n-1))

(where constant is the value obtained when n=i.)
This is an example of Type II censoring (see Figure 1.3).

1.6.4 Realistic case II

```
      Poisson - random process
         -----*-----*-----*-----*-----*-----*
              A                       B
```

If there is a random process of withdrawal (censoring) such as emmigration, etc then we might consider a uniform distribution of the infinite time line (ie Poisson rate p_i) hence in RC(1) if and only if censor time from recruitment ~ exp(rate p_i)
ci=min(exp(p_i),(constant)+((n-i)/(n-1)))
xi=min(T_i,c_i)
This is an example of Type II censoring (see Figure 1.3).

1.7 Examples

I have obtained the following examples for each model as described in section 1.6. I am going to use them again to use various methods and models. Note in section 1.7.2 I took constant = 1.25, section 1.7.3 I took the constant when n=i, as 0.5 and in section 1.7.4, I took the constant when n=i, as 0.5 and the rate parameter of the exponential distribution as 0.8. In later work, I will have at least 2 treatments and other examples will be given in the relevant appendices. (* denotes censored)

1.7.1 Uncensored

1.339	.412	.461	.670	.509	.365	.329	.706	.628	.415	.495	.571
.233	.547	.636	1.207	.612	1.043	.665	.679	.244	.524	.674	.531
.852	1.147	.946	.882	.494	.742	.249	.172	.552	.489	.903	.689
.673	.830	.878	.194	.663	.488	.476	1.312	.598	.397	.287	.630
.341	.328	.964	.831	.467	.770	.569	.285	.502	.533	.184	.599
1.101	.537	.939	.847	.390	.926	.527	.413	.491	.580	.669	.984
.900	.616	.865									

1.7.2 Simplest (constant) censoring

1.250*	.412	.461	.670	.509	.365	.329	.706	.628	.415	.495	.571
.233	.547	.636	1.207	.612	1.043	.665	.679	.244	.524	.674	.531
.852	1.147	.946	.882	.494	.742	.249	.172	.552	.489	.903	.689
.673	.830	.878	.194	.663	.488	.476	1.250*	.598	.397	.287	.630
.341	.328	.964	.831	.467	.770	.569	.285	.502	.533	.184	.599
1.101	.537	.939	.847	.390	.926	.527	.413	.491	.580	.669	.984
.900	.616	.865									

1.7.3 Realistic case I

1.339	.412	.461	.670	.509	.365	.329	.706	.628	.415	.495
.571	.233	.547	.636	1.207	.612	1.043	.665	.679	.244	.524
.674	.531	.852	1.147	.946	.882	.494	.742	.249	.172	.552
.489	.903	.689	.673	.830	.878	.194	.663	.488	.476	
.919*	.598	.397	.287	.630	.341	.328	.824*	.811*	.467	.770
.569	.285	.502	.533	.184	.599*	.689	.537*	.662*	.649*	.390
.622*	.527	.413	.491	.568*	.554*	.541*	.527*	.514*	.500*	

1.7.4 Realistic case II

.694*	.363*	.461	.031*	.509	.074*	.329	.706	.247*	.415	.495
.571	.233	.547	.636	.808*	.612	.817*	.184*	.679	.244	.524
.674	.268*	.852	1.147	.317*	.130*	.424*	.742	.249	.172	
.097*	.489	.903	.240*	.673	.191*	.878	.194	.663	.488	
.289*	.699*	.598	.177*	.266*	.630	.341	.328	.824*	.811*	.467
.770	.569	.086*	.502	.531*	.184	.224*	.689*	.537	.662*	
.649*	.390	.622*	.127*	.413	.211*	.568*	.554*	.465*	.527*	
.514*	.500*									

1.8 Computing

It is essential to write computer programs in order to use various methods and models in Survival Analysis. I write programs in Fortran 77 (on the Norsk machines, using VI Editor on UNIX) and their listings and output can be found in APPENDIX 02. I also write GLIM 3.77 programs/macros and their listings and output can be found in APPENDIX 03.

1.9 Summary

According to Cox and Oakes (1984), survival analysis is univariate, as there is a single response variable, failure time, even though there are many other explanatory variables, such as censor times. Sometimes, in many scientific and medical applications, failure time is plotted on the x-axis and physical quantities/statistics (eg heart rate, survivor function, body temperature, etc) are plotted on the y-axis.

"To determine failure time precisely, the time origin must be defined, the event for measuring the time scale must be defined and the meaning of failure must be clear." (CO, 1984)

"Incomplete observation of the failure time is called CENSORING. Note that like failure, censoring is a point event and must be recorded." (CO, 1984)

It is difficult to obtain real data. The only way to get data is to use simulation. Simulation is cheaper and quicker. Computers are increasingly important, nowadays.
I used the inversion method and formula for the simulation.

Type I censoring is where the censor times are constant.
Type II censoring is where the censor times are allowed to be random variables.

I used the following models:
No censoring, Constant censoring (Type I), Realistic case I and Realistic case II (Type II)

No censoring

$x_i=T_i$ for all i, where i=1,n

Constant censoring:

$x_i=T_i$ if $T_i<\text{Constant}$

$x_i=\text{Constant}$ if $T_i>=\text{Constant}$

Realistic case I:

$c_i=(\text{constant})+((n-i)/(n-1))$

Realistic case II:

$c_i=\min(\exp(p_i,(\text{constant})+((n-i)/(n-1)))$

$x_i=\min(T_i,c_i)$