DEDAN KIMATHI UNIVERSITY OF TECHNOLOGY

PRIVATE BAG, DEDAN KIMATHI, 10143

TELEPHONE: (061)-2050000, +254(0) 736-456391, FAX: +254(020) 2417997; Email: actuarialscience@dkut.ac.ke; Website: www.dkut.ac.ke

SCHOOL OF SCIENCE

DEPARTMENT OF STATISTICS & ACTUARIAL SCIENCE

SAS 4101 SURVIVAL ANALYSIS
STUDY GUIDE

DeKUT is ISO 9001:2015 Certified

Better life Through Technology

Copyright page

This material is protected by copyright and has been copied by and solely for the educational purposes of the University under licence. You may not sell, alter or further reproduce or distribute any part of this course pack/material to any other person. Where provided to you in electronic format, you may only print from it for your own private study and research. Failure to comply with the terms of this warning may expose you to legal action for copyright infringement and/or disciplinary action by the University.

Course outline

Course Purpose:
On completion of the course the student should be able to estimate survival rates.
Course Content:

- Distribution and density functions of the random future lifetime, the survival function and the force of hazard
- Estimation procedures for lifetime distributions including censoring, Kaplan-Meier estimate, Nelson-Aalen estimate and the Cox regression model.
- The graduation process. Testing of graduations.
- Use of computer packages.

References

1. Elandt-Johnson & Johnson, *Survival Models and Data Analysis*, John-Wiley and Sons,1999.
2. Irwin Miller and Marylees Miller, Sixth edition. John E. Freund's Mathematical Statistics.
3. Marubini, E; Valsecchi, M G. *Analysing survival data from clinical trials and observational studies*. John Wiley, 1995.
4. Elandt-Johnson , R C; Johnson, Norman L. *Survival Models and data analysis*. John Wiley. 1997
5. Marubini, E & Valsecchi, M. Analysing *Data from clinical trials and Observational studies*. John Wiley. 1995.

Course Lecturer: Mundia 0721302869 simon.maina@dkut.ac.ke

Teaching Schedule

Week 1 Introduction, censoring.

Week 2 Definitions and notations

Week 3 Distributions of failure times **Assignment 1**

Week 4 Parametric Modeling of failure times; estimation. **Cat1**

Week 5 Parametric Modeling of failure times; hypotheses testing.

Week 6 Non-Parametric Modeling of failure times; Kaplan-Meir Method derivation.

Week 7 Non-Parametric Modeling of failure times; Kaplan-Meir Method examples

Week 8 Non-Parametric Modeling of failure times; Nelson-Aalen; Method and properties of the estimator. **Assignment 2**

Week 9 Regression models: Cox proportional hazards; derivation. **Cat2**

Week 10 Regression models: Cox proportional hazards ; estimation of the parameters using partial likelihood.

Week 11 Regression models: Cox proportional hazards derivation; examples

Week 12 Graduation, Statistical tests.

Week 13 Graduation, Statistical tests. **Cat3**

Wk14,15 exams

Note: Cat 1 and 2 will not be taken during class-time.

Contents

1 Introduction

Consider a group of individuals. Associated with each individual is a point event. This event can only occur once to an individual, it is refereed to as *failure*. For example death, failure of a machine, maturity of an insurance policy or any designated experience of interest that may happen to an individual. **Survival analysis** is a collection of statistical procedures for data analysis for which the outcome variable of interest is **time until the event occurs.** The time may be in years, months, weeks, or days from the beginning of follow-up of an individual until **an event** occurs; alternatively, time can refer to the age of an individual when an event occurs. Note that this variable is non-negative.

Although more than one event may be considered in the same analysis, the assumption is that only one event is of designated interest. When more than one event is considered (e.g., death from any of several causes), the statistical problem can be characterized as either a *recurrent event* or a *competing risk* problem. The time variable is referred to as *survival time*, because it gives the time that an individual has survived over some follow-up period.

An example of survival analysis problem, is a follow up of disease-free cohort of individuals over several years to see who develops heart disease. Here the event is "developing heart disease", and the outcome is "time in years until a person develops heart disease". Most survival analyses must consider a key analytical problem called **censoring**. In essence, censoring occurs when one has some information about an individual survival time, but the exact survival time is unknown. Thus there is an incomplete observation of the individual. Considering the above example, if for an individual, the study ends while the the person has not developed the heart disease, then his survival time is considered censored. The survival time is at least as long as the period that the person has been followed, but if the person has a heart disease after the study ends, the complete survival time is unknown. Regardless of the type of censoring, we must assume that it is **non-informative** about the event; that is, the censoring is caused by something other than the impending failure.

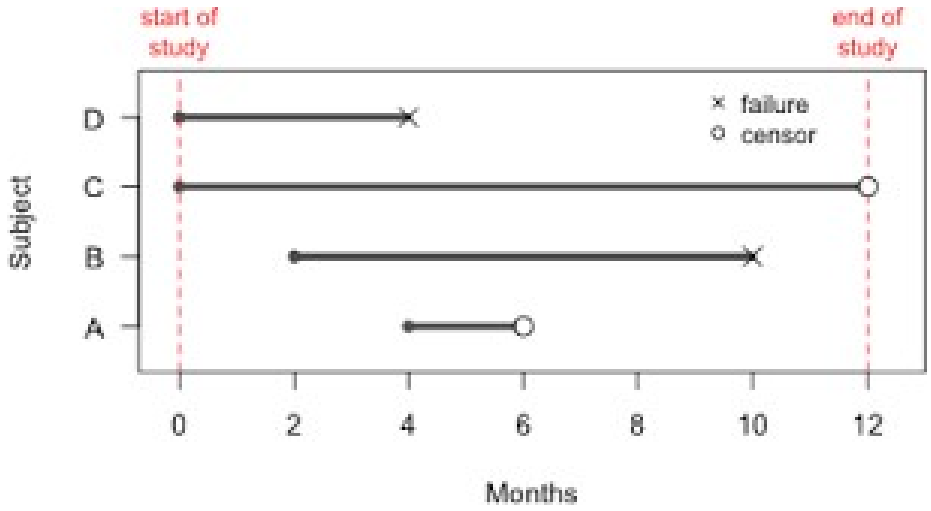

Figure 1: censoring

1.1 Types of censoring

1. **Right censoring**
   Right-censoring occurs when the true unobserved event is to the right of the censored time; i.e., all that is known is that the event has not happened at the end of follow-up. This most common type of censoring dealt with in survival analysis. There are generally three reasons why right censoring may occur:

   - termination the study ;
   - loss of follow-up during the study period;
   - withdrawal from the study.

   There are three major of right censoring

   - **Fixed type I censoring** occurs when a study is designed to end after a fixed time of follow-up. In this case, everyone who does not have an event observed during the course of the study is censored.
   - **random type I censoring**. Here the study is designed to end after a fixed time, but censored subjects do not all have the same censoring time. This is the main type of right-censoring.
   - In **type II** censoring, a study ends when there is a pre-specified number of events.

2. **Left-censoring**: This can occur when a person's true survival time is less than or equal to that person's observed survival time. For example, if we are following persons until they become HIV positive, we may record a failure when a subject first tests positive for the virus. However, we may not know the exact time of first exposure to the virus, and therefore do not know exactly when the failure occurred. Thus, the survival time is censored on the left side since the true survival time, which ends at exposure, is shorter than the follow-up time, which ends when the subject's test is positive.

3. **Interval-censoring**. This can occur if a subject's true (but unobserved) survival time is within a certain known specified time interval. As an example, again considering HIV surveillance, a subject may have had two HIV tests, where he/she was HIV negative at the time (say, t_1) of the first test and HIV positive at the time (t_2) of the second test. In such a case, the subject's true survival time occurred after time t_1 and before time t_2, i.e., the subject is interval censored in the time interval (t_1, t_2).

1.2 Basic objectives of survival analysis.

The basic objectives of survival analysis are to :-

- estimate the survivor and/ or hazard functions from survival data.
- interpret the survivor and/ or hazard functions from survival data.
- compare survivor and/or hazard functions.
- assess the relationship of explanatory variables to survival time.

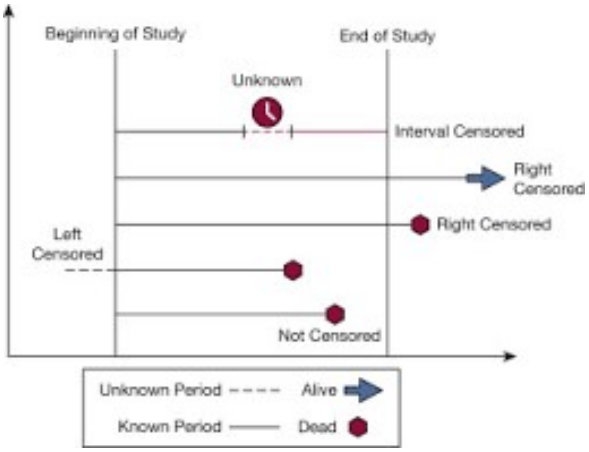

Figure 2: Main types of censoring

1.3 Definitions and Notation

Consider a population of individuals each having a failure time $T \geq 0$. T is a random variable whose origin and scale are assumed to be well defined. The probability that an individual survives upto a time t is

$$\begin{aligned} P(T > t) &= 1 - P(T \leq t) \\ &= 1 - F(t) \\ &= S(t) \end{aligned} \tag{1}$$

$S(t)$ is called the **survivor function** and it gives the probability that an individual will survive past time t. As t ranges from 0 to ∞, the survival function has the following properties

- It is non-increasing
- At time $t = 0$, $S(t) = 1$. In other words, the probability of surviving past time 0 is 1.
- At time t = ∞, $S(t) = S(\infty) = 0$. As time goes to infinity, the survival curve goes to 0.

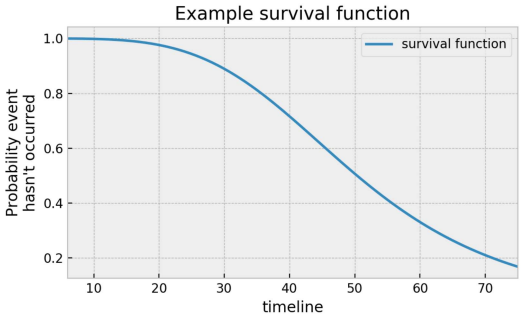

Figure 3: Survivor function

Note: These are theoretical properties of survivor curves. In practice, when using actual data, step functions rather than smooth curves are obtained.
Another function of importance is the instantaneous failure rate also called the hazard function defined as

$$\begin{aligned} h(t) &= \lim_{\partial t \to 0} \frac{P(t \leq T \leq t + \partial t | T > t)}{\partial t} \\ &= \lim_{\partial t \to 0} \frac{P(t \leq T \leq t + \partial t)}{\partial t} \frac{1}{P(T > t)} \\ &= \frac{f(t)}{S(t)} \\ &= \frac{-S'(t)}{S(t)} \\ &= -\frac{d}{dt} \log S(t) \end{aligned} \tag{2}$$

Note:

- $h(t)\partial t$ approximates $P(t \leq T \leq t + \partial t | T > t)$ the conditional probability of failure within the interval $[t,\ t + \partial t]$ given survival to time t.
- Whenever one thinks of a subject at time t he pictures a situation where failure has not occurred and thus does the thinking in terms of the risk of the event occurring using a conditional probability.

The cumulative or integrated hazard describes the accumulated risk up to time t and is given as

$$H(t) = \int_0^t h(u)\ du \tag{3}$$

Note that

$$\begin{aligned} -\int_0^t h(u)\ du &= \int_0^t \frac{d}{du} \log S(u)\ du \\ &= \log S(t) \\ \Rightarrow\ S(t) &= \exp\left(-\int_0^t h(u)\ du \right) \\ &= \exp[-H(t)] \end{aligned} \tag{4}$$

also

$$\begin{aligned} f(t) &= h(t)S(t) \\ &= h(t) \exp[-H(t)] \end{aligned} \tag{5}$$

Example 1.1. Show that $E(t) = \int_0^\infty S(t)\ dt$

Solution.

$$\begin{aligned}
\int_0^\infty S(t)\ dt &= \int_0^\infty \left(\int_t^\infty f(s)\ ds\right)\ dt \\
\text{let } u &= \int_t^\infty f(s)\ ds = 1 - F(t) \quad \Rightarrow\ du = -f(t)\ dt \\
\text{Also let } dt &= dv \quad \Rightarrow v = t
\end{aligned}$$

Thus

$$\begin{aligned}
\int_0^\infty S(t)\ dt &= \left[t\int_t^\infty f(t)\ dt\right]_0^\infty + \int_0^\infty t\ f(t)\ dt \\
&= E(t)
\end{aligned}$$

Question 1.1. Express $P(T > t + k|T > t)$ in terms of $H(t)$.

For a discrete T, let $u_0 < u_1 < u_2 < u_3 < \cdots$ be the points at which the p.m.f. is non zero. Then the pmf of T is

$$f_j = P(T = u_j)$$

The survivor function

$$\begin{aligned}
S_j &= P(T \geq u_j) \\
&= f_j + f_{j+1} + f_{j+2} + \cdots
\end{aligned} \tag{6}$$

and the hazard function is

$$\begin{aligned}
h_j &= P(T = u_j|T \geq u_j) \\
&= \frac{P(T = u_j)}{P(T \geq u_j)} \\
&= \frac{f_j}{S_j} \\
&= \frac{f_j}{f_j + f_{j+1} + f_{j+2} + \cdots} \\
&\text{But} \quad f_j + f_{j+1} + f_{j+2} + \cdots = f_j + S_{j+1} \\
\Rightarrow\ h_j &= \frac{f_j}{f_j + S_{j+1}}
\end{aligned} \tag{7}$$

Now

$$\begin{aligned} S_{j+1} &= \frac{f_j}{h_j} - f_j \\ &= \frac{f_j}{h_j}(1-h_j) \\ &= S_j(1-h_j) \\ &= S_{j-1}(1-h_{j-1})(1-h_j) \\ &= S_{j-2}(1-h_{j-2})(1-h_{j-1})(1-h_j) \\ &\vdots \quad \vdots \\ &= S_0(1-h_0)(1-h_1)(1-h_2)\cdots(1-h_{j-1})(1-h_j) \\ \text{But} \quad & S_0 = P(T \geq 0) = 1 \\ \Rightarrow \; S_{j+1} &= \prod_{k=0}^{j}(1-h_k) \end{aligned} \tag{8}$$

Thus the survivor function at time t is

$$S_t = \prod_{u_j<t}(1-h_j) \tag{9}$$

Also

$$\begin{aligned} S_j &= \frac{f_j}{h_j} \\ \Rightarrow \; f_j &= S_j h_j \\ &= h_j(1-h_0)(1-h_1)(1-h_2)\cdots(1-h_{j-1}) \\ &= h_j \prod_{k=0}^{j-1}(1-h_k) \end{aligned} \tag{10}$$

Example 1.2. Show that if $t = 0, 1, 2, \cdots$ then $E(T) = \sum_{j=1}^{\infty} S_j$

Solution.

$$\begin{aligned} E(T) &= \sum_{j=1}^{\infty} t f_t \\ E(T) &= f_1 \\ &+ f_2 + f_2 \\ &+ f_3 + f_3 + f_3 \\ &\vdots \\ &= S_1 + S_2 + S_3 + \cdots = \sum_{j=1}^{\infty} S_j \end{aligned}$$

Remarks

- In human mortality, the future lifetime of a "life" is considered. The assumption is that future lifetimes are random variables and thus the life of a newborn is uniformly distributed in the interval $[0, \omega]$ where ω is the limiting age.
  Thus $F(t) = P(T \leq t)$ is the distribution function of T and $S(t)$ is the probability of a child surviving to age t.

- In the insurance context the notation needs to be extended so that older people are accommodated.
  Let T_x be the future lifetime after age x, then $F_x(t) = P(T_x \leq t)$ and $S_x(t) = 1 - F_x(t)$. Also ${}_tq_x = F_x(t)$ is the probability of death and ${}_tp_x = 1 - {}_t\, q_x = S_x(t)$ is the probability of a life aged x surviving in the next t years.
  The force of mortality at age x as per the definition in equation(3) is

$$\mu_x = \lim_{h \to 0} \frac{P(T = x + h | T > x)}{h}$$

  For a $h \approx 0$ then $hq_x \approx h\mu_x$

2 Parametric Modeling of Failure Times

The survivor and hazard functions are important tools in the analysis of survival data. We now want to answer the following question. "what underlies failure concepts for a given dataset? i.e. what is the pdf of T ? "
For continuous T the following are some of the pdfs used to model the failure times.

2.1 The Exponential Distribution

Consider a constant hazard function $h(t) = \mu$. The cumulative hazard is $H(t) = \mu t$. The survivor function $S(t) = \exp(-\mu t)$ and the pdf of T is $f(t) = \mu \exp(-\mu t)$.

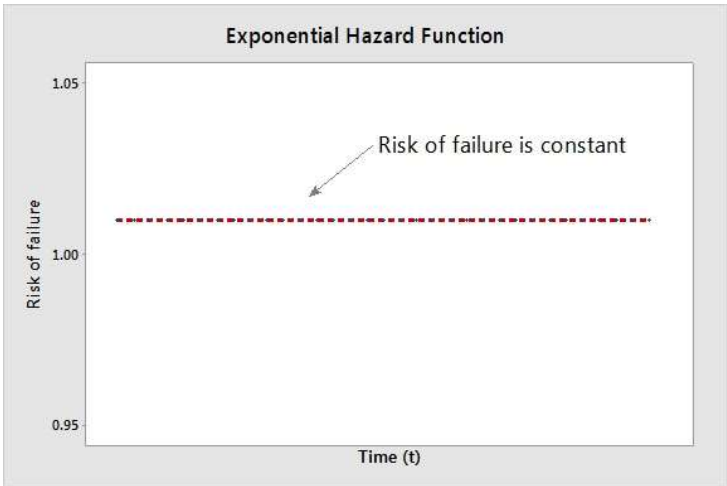

Figure 4: Constant hazard

Note: If T has a constant hazard then it has exponential distribution. The converse of this is also true.
A constant hazard implies that after sometime the probability of failure does not change. Consider

$$\begin{aligned}
P(t < T < t + \partial t | T > t) &= \frac{P(t < T < t + \partial t \cap T > t)}{P(T > t)} \\
&= \frac{P(t < T < t + \partial t)}{P(T > t)} \\
&= \frac{1 - [P(< T) + P(T > t + \partial t)]}{P(T > t)} \\
&= \frac{1 - [1 - P(t > T) + P(T > t + \partial t)]}{P(T > t)} \\
&= \frac{P(t > T) - P(T > t + \partial t)}{P(T > t)} \\
&= \frac{S(t) - S(t + \partial t)}{S(t)} \\
&= 1 - \frac{S(t + \partial t)}{S(t)}
\end{aligned}$$

Since T has exponential distribution then

$$P(t < T < t + \partial t | T > t) = 1 - \exp(-\mu \partial t)$$

which is independent of t. Hence the failure time is independent of the history. It is said to be memoryless.

Question 2.1. If a r.v. T has a continuous distribution show that $R(t) = -\log S(t)$ is distributed as a standard exponential r.v. with a unit hazard function.

2.2 The Wei-Bull Distribution

Here the hazard is of the form $h(t) = \mu\beta(\mu t)^{\beta-1}$ where $\mu,\ \beta > 0$ are the scale and index parameters respectively. If $\beta = 1$ we have a constant hazard hence $h(t)$ is referred as the generalised two parameter exponential hazard. The integrated hazard is $(\mu t)^{\beta}$ so that the survivor function $S(t) = \exp(-(\mu t)^{\beta})$. The pdf of T is $f(t) = \mu\beta(\mu t)^{\beta-1}\exp(-(\mu t)^{\beta})$. The distribution is convenient due to its simple form and in that the hazard has several shapes.

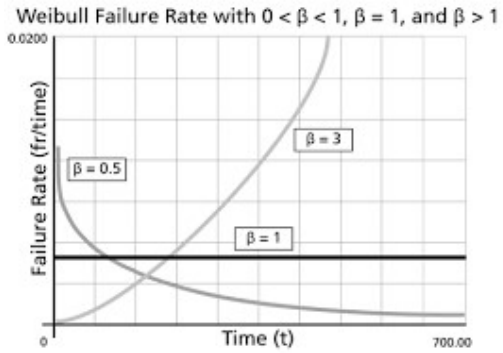

Figure 5: Weibull hazards

Other distributions are Log-Logistic, Lognormal and Gompertz.

Assignment I

1. For each of the distributions mentioned above obtain the possible shape(s) of the hazard give an example of a situation in which the hazard function may be utilised.
2. What are the major drawbacks of using parametric models to analyse survival data.

2.3 Inference on the Parametric Models

Here the assumption made is that the data can be modelled by a specified family of distribution $f(t;\theta)$ whose form is form is known except the parameter θ. Suppose there is a single sample of failure times possibly subject to censoring and inference on θ is required. The inference will be based on the likelihood function.
Let T is continuous. A subject observed to fail at t contributes $f(t;\theta)$, while censored at t contributes $S(t;\theta)$ to the likelihood. The total likelihood is

$$L = \prod_{u} f(t_i;\theta) \prod_{c} S(t_i;\theta) \tag{11}$$

The log likelihood is

$$\begin{aligned}\ell &= \sum_u \log f(t_i;\theta) + \sum_c \log S(t_i;\theta) \\ &= \sum_u \log h(t_i;\theta) + \sum_u \log S(t_i;\theta) + \sum_c \log S(t_i;\theta) \\ &= \sum_u \log h(t_i;\theta) + \sum \log S(t_i;\theta) \\ &= \sum_u \log h(t_i;\theta) - \sum H(t_i;\theta) \end{aligned} \tag{12}$$

For a discrete T define
d_j=number of failures at time u_j.
r_j=number of subject at risk at time u_j including those censored.
$h_j(\theta)$=probability of failure at the j^{th} interval conditional on survival at the start of the interval.
Then subjects failing time u_j contribute $[h_j(\theta)]^{d_j}$ to the likelihood while those at risk contribute $[1-h_j(\theta)]^{r_j-d_j}$. Thus for a sample of size n the total likelihood is

$$L = \prod_{j=1}^{n} [h_j(\theta)]^{d_j} [1-h_j(\theta)]^{r_j-d_j} \tag{13}$$

and the log-likelihood is

$$\ell = \sum_{j=1}^{n} d_j \log h_j(\theta) + \sum_{j=1}^{n} [r_j - d_j] \log(1-h_j(\theta)) \tag{14}$$

Using equations(12) and (14) one can obtain the mle for θ.
On hypotheses testing, we may wish to test $H_0 : \theta = \theta_0$. One may use the likelihood ratio approach. Let

$$Q = 2(\log \hat{\theta} - \log \hat{\theta}_0) \tag{15}$$

Q is the likelihood ratio statistic and $\hat{\theta}$ and $\hat{\theta}_0$ are the mles under the null and the alternative respectively. Under H_0 :, $Q \sim \chi^2$ with $dim(\theta)$ d.f. Thus the $100(1-\alpha)\%$ may be obtained.

Example 2.1. Consider the exponential distribution. Here the hazard function is $h(t) = \mu$, the cumulative hazard is $H(t) = \mu t$ and the survivor function $S(t) = \exp(-\mu t)$. The log-likelihood is

$$\begin{aligned}\ell &= \sum_u \log \mu - \mu \sum t_i \\ &= d \log \mu - \mu \sum t_i \end{aligned} \tag{16}$$

where d is the observed number of failures. $\sum t_i$ is called the total time at risk. Differentiating and equating to zero we obtain the mle of μ as $\dfrac{d}{\sum t_i}$. The variance of this estimate is

$$-\left[\frac{d^2\ell}{d\mu^2}\right]^{-1}_{\mu=\hat{\mu}} = \frac{d}{(\sum t_i)^2} \tag{17}$$

In the absence of censoring the log-likelihood is

$$\ell \quad = \quad n \log \mu - \mu \sum t_i \tag{18}$$

Here the exact inference about μ is possible as $\sum t_i$ has gamma distribution and $\frac{2n\mu}{\hat{\mu}}$ has χ^2 with $2n$ d.f.
Consider the the following numerical example, which is a report on a clinical trial to evaluate the efficacy of maintenance chemotherapy for acute Leukaemia. Patients were randomly allocated to group A; which received the treatment and group B which did not. The time to remission for both groups are

group A 9 13 13* 18 23 28* 31 34 45* 48 167*
group B 5 5 8 8 12 23 27 30 33 43 45

For group A
d=7, $\sum t_i = 429$ and $\hat{\mu} = 7/429$ and thus $var(\hat{\mu}) = 0.017^2/7$.
The 95% CI for the estimate is $\hat{\mu} \pm 1.96\hat{\sigma} = [0.005, 0.029]$.
For group B there is no censoring
n=11, $\sum t_i = 239$ and $\hat{\mu} = 11/239$.
Now

$$\begin{aligned} \frac{2n\mu}{\hat{\mu}} \quad &\sim \quad \chi^2(2n) \\ &\Rightarrow \quad P\left(a \leq \frac{2n\mu}{\hat{\mu}} \leq b\right) = 0.95 \\ &\Rightarrow \quad P\left(\frac{a\hat{\mu}}{2n} \leq \mu \leq \frac{b\hat{\mu}}{2n}\right) = 0.95 \\ &\text{but} \quad a = \chi^2_{0.025}(22) \text{ and } b = \chi^2_{0.975}(22) \end{aligned}$$

but $a = \chi^2_{0.025}(22)$ and $b = \chi^2_{0.975}(22)$ so that the 95% CI can be obtained.

Question 2.2. A random variable, T, has the Weibull distribution with probability density function

$$f(t) = \begin{cases} \lambda\gamma t^{\gamma-1}\exp(-\lambda t^{\gamma}), & t > 0 \quad \lambda > 0 \quad \lambda > 0 \\ 0, & \text{otherwise} \end{cases}$$

(i) Derive the survivor function, $S(t)$, of T.

(ii) Show that $\log\{-\log(S(t))\}$ is a linear function of $\log(t)$. State the intercept and slope of the line if $\log\{-\log(S(t))\}$ on the vertical axis is plotted against log(t) on the horizontal axis.

(iii) 20 refrigerator motors of a particular type were each tested on an accelerated life test, and their times till first failure (hours) were recorded. The results are listed below, where * denotes that the motor was still functioning properly when the test was brought to an end.

2 4* 5 5 5 6* 7 7 7* 8 8 9* 11 11 12 12* 12* 12* 16 18*

Referring to the result from part (i) and (ii), use a suitable graphical method to investigate whether or not these data come from a Weibull distribution.
Draw a straight line through the points on your graph by eye and use it to estimate the parameters, λ and γ, of a Weibull distribution fitted to these data.

Assignment II
Consider a Wei-Bull distribution with hazard function $h(t) = \kappa\rho(\rho t)^{\kappa-1}$. Assuming that there are d failures obtain the mles of the parameters.
The following are remission times in weeks of Leukaemia patients after treatment

6 6 6 6* 7 9* 10 10* 11* 13 16 17* 19* 20*

Obtain the mles for the parameters.

3 Non-Parametric Estimation of the Survivor Function

Here the distributional form of the data is not necessary. Two approaches will be considered.

3.1 Kaplan-Meier

Consider the following where there is no censoring;
1 1 2 2 3 4 4 5 5 8 8 8 8 11 11 12 12 15 17 22 23
and one wishes to estimate $S(t)$.

t	$\widehat{S(t)}$
$t \leq 1$	21/21
$1 < t \leq 2$	19/21
$2 < t \leq 3$	17/21
$\vdots$	$\vdots$

Thus in the estimator is

$$\widehat{S(t)} = \frac{\text{no. of individuals with } T \geq t}{\text{sample size}} \tag{19}$$

Plot the graph of this estimate.
In the presence of censoring the denominator of equation(19) is varies. We can partition the observed timespan into a series of fine intervals so that there is a separate interval for each time of failure or censoring.

		F		C		C	F, C		

Consider the j^{th} interval. There are four possible outcomes;

(i) **No event**. The conditional probability of surviving this interval is 1.

(ii) **Censoring**. Assume that the censored individuals survive to the end of this interval so that the conditional probability of surviving this interval is 1.

(iii) **Failure but no censoring**. The conditional probability of not surviving this interval the number of deaths in this interval d_j divided by the number of those at risk r_j.

(iv) **Failure and censoring**. Assume that censoring lasts to the end of the interval so that the probability of surviving this interval is $1 - \frac{d_j}{r_j}$.

Thus the conditional probability of surviving the j^{th} interval is $1 - \frac{d_j}{r_j}$. Since the outcomes in the intervals are independent then

$$\begin{aligned} P(T \geq t) &= \prod_j P(\text{survive } I_j | \text{survival at start } I_j) \\ \widehat{S(t)} &= \prod_{j:u_j \leq t} \left(1 - \frac{d_j}{r_j}\right) \end{aligned} \tag{20}$$

where $u_1, u_2, \cdots, u_k$ is the set of distinct failure times observed.
Note: The assumption that the censored subject lasts until the end of the interval is not quite accurate so that we obtain crude approximations but as the interval gets finer the estimator converges to $S(t)$.
Equation(20) is called the Kaplan-Meier Estimator for the survivor function.

Example 3.1. Consider the following data
6 6 6 6* 7 9* 10 10* 11* 13 16 17* 19* 20* 22 23 25* 32* 32* 34* 35*
Compute the Kaplan-Meier estimate for the survivor function.
Tabulate the information as follows

u_j	d_j	c_j	r_j	$1-\frac{d_j}{r_j}$	$\widehat{S(t)}$
6	3	1	21	6/7	6/7
7	1	0	17	16/17	6/7×16/17
9	0	1	16	1	6/7×16/17
⋮					

Note: Any term with $d_j = 0$ can be omitted. The estimate then is

$$\widehat{S_{km}(t)} = \begin{cases} 1 & 0 \leq t < 6 \\ 6/7 & 6 \leq t < 7 \\ 6/119 & 7 \leq t < 10 \\ \vdots \end{cases}$$

Plot this estimate.

Note: The estimate is a constant after the last duration a failure is observed and only those at risk and those that have failed contribute to the estimate. Thus it is assumed that the censoring is non-informative. We also assume that T is discrete.

Question 3.1. Obtain the K-M estimator of the survivor function for the data in example 2.1

3.2 Properties of the KM estimator

In the absence of censoring the estimator in equation(19) is just like an estimated probability from a binomial distribution by taking the numerator as $X \sim b(n, S(t))$ so that

$$\widehat{S(t)} \sim N\left(S(t), \frac{S(t)[1-S(t)]}{n}\right) \tag{21}$$

In the presence of censoring $\widehat{S_{km}(t)}$ has approximately normal distribution with mean $S(t)$ but the variance is changed as the denominator is affected by the censored observations. This variance may be estimated using the log-likelihood theory as the variance can be estimated using the negative expectation of the second derivative of the log-likelihood.

Now using equation(14)

$$\begin{aligned}
\frac{dl}{dh_j} &= \frac{d_j}{h_j} - \frac{r_j - d_j}{1-h_j} \\
\frac{d^2l}{dh_j^2} &= -\frac{d_j}{h_j^2} - \frac{r_j - d_j}{(1-h_j)^2} \\
&= -\frac{\hat{h}_j r_j}{h_j^2} - \frac{r_j - \hat{h}_j r_j}{(1-h_j)^2} \quad \text{since} \quad d_j = \hat{h}_j r_j \\
\left.\frac{d^2l}{dh_j^2}\right|_{h_j=\hat{h}_j} &= \frac{-r_j}{\hat{h}_j(1-\hat{h}_j)} \\
\Rightarrow \quad Var(\hat{h}_j) &= \left[-E\left\{\left.\frac{d^2l}{dh_j^2}\right|_{h_j=\hat{h}_j}\right\}\right]^{-1} \\
&= \frac{\hat{h}_j(1-\hat{h}_j)}{r_j}
\end{aligned} \tag{22}$$

Also, the $\hat{h}_j$'s are independent in large enough samples. Since $\widehat{S(t)}$ is a function of the $\hat{h}_j$'s, its variance can be estimated using the delta method.
Delta method: If X is normal with mean μ and variance σ^2, then $g(X)$ is approximately normally distributed with mean $g(\mu)$ and variance $[g'(\mu)]^2\sigma^2$.

Example 3.2. Let $Y = \log X$ then $E(Y) = \log\mu$ and $var(Y) = \dfrac{\sigma^2}{\mu^2}$
Also if $Y = \exp X$ then $E(Y) = \exp\mu$ and $var(Y) = [\exp\mu]^2\sigma^2$

Now

$$\begin{aligned}
\widehat{S(t)} &= \prod_{j:u_j\le t}\left(1-\hat{h}_j\right) \\
\log\widehat{S(t)} &= \sum_{j:u_j\le t}\log\left(1-\hat{h}_j\right) \\
var(\log\widehat{S(t)}) &= \sum_{j:u_j\le t} var(\log(1-\hat{h}_j)) \\
&= \sum_{j:u_j\le t}\left(\frac{1}{1-\hat{h}_j}\right)^2 var(\hat{h}_j) \\
&= \sum_{j:u_j\le t}\left(\frac{1}{1-\hat{h}_j}\right)^2 \frac{\hat{h}_j(1-\hat{h}_j)}{r_j} \\
&= \sum_{j:u_j\le t}\frac{d_j}{r_j(r_j-d_j)}
\end{aligned} \tag{23}$$

Thus, noting that $E(\log\widehat{S(t)}) = \log\widehat{S(t)}$ then

$$\begin{aligned}
var(\widehat{S(t)}) &= [\widehat{S(t)}]^2 var(\log\widehat{S(t)}) \\
&= [\widehat{S(t)}]^2 \sum_{j:u_j\le t}\frac{d_j}{r_j(r_j-d_j)}
\end{aligned} \tag{24}$$

Equation(24) is referred to as the Greenwood's formulae. Note that it is a function of t. The value seems reasonable but it underestimates variance at the tails of the distribution. Thus one can now obtain the piecewise confidence interval for $\widehat{S(t)}$.

Question 3.2. Consider Example 3.1 construct the 95% confidence interval for $\widehat{S(t)}$.

Example 3.3. A study of mortality of 12 insects is undertaken. They are observed from birth upto death or the end of the study, at which point insects still alive are treated as censored. The following is the KM estimate of survivor function based on the 12 insects

$$\widehat{S_{km}(t)} = \begin{cases} 1 & 0 \leq t < 1 \\ 0.9167 & 1 \leq t < 3 \\ 0.7130 & 3 \leq t < 6 \\ .42878 & t \geq 6 \end{cases}$$

Calculate the number of insects dying at week 3 and 6 and those that are censored.

Solution. Let u_j be the duration at which event take place.
At $u_j = 1$ $\widehat{S_{km}(t)}$ falls from 1 to 0.9167. Thus $1 - h_1 = 0.9167 \quad h_1 = 0.083\dot{3}$ and since $h_1 = d_1/r_1$ then $d_1 = 1$ as $r_1 = 12$.
At $u_j = 3$ 07130. $= 0.9167(1 - h_3)$ and thus $h_3 = 0.22\dot{2}$ and since $h_3 = d_3/r_3$ then $d_3 = 2$. As we had at most 11 insects at risk the $r_3 = 9$. Two insects were thus censored at $u_j = 1$.
At $u_j = 6$ $0.42878 = 0.7130(1 - h_6)$ and thus $h_6 = 0.4 = 2/5$. Since there were at most 7 insects at risk at $u_j = 6$ then $d_6 = 2$ and $r_6 = 5$ and thus 2 insects were censored at $u_j = 3$. Also since the total number were 12 then those censored at the end of the study is 3. This may be summarised as

u_j	$S(t)$	h_j	r_j	c_j	d_j
0	1	0	12	0	0
1	0.9167	0.0833	12	2	1
3	0.7130	0.22	9	2	2
6	0.42878	0.4	5	3	2

Question 3.3. A certain profession admits new members at the status of student. The student may qualify as fellows in the profession by the virtue of passing a series of exams. The exams are done whilst working for an employer. There are two sessions of exams each year.
An employer supports student members to the profession. He wishes to assess the cost of providing this study support and therefore wishes to know the average time a student takes to qualify. The employer maintains records for 23 students who first sat for the exams in the first session of 2003 and this record maintained upto the last session of 2009. The following was observed
Qualified: 6,8,8,9,9,9,11,11,13,13,13
Stopped Studying: 4,5,8,11,14
The remaining students were still studying for the exams at the end of 2009.

(i) Determine the median number of sessions taken to qualify and comment

(ii) Calculate the KM estimate for the hazard of qualifying.

3.3 Estimating the Cumulative Hazard

Suppose one wishes to estimate $H(t) = \int_0^t h(u)du$ .Divide the observed timespan into a series of fine intervals so that there is only one event per interval. Then

$$\widehat{H(t)} = \sum_j h_j \partial t \tag{25}$$

where the sum is over all possible intervals and ∂t is the width of each interval. Since $\hat{h}_j \partial t$ is the approximate probability of failure in the interval it can be further approximated by d_j/r_j. Thus equation(25) becomes

$$\widehat{H(t)} = \sum_j \frac{d_j}{r_j} \tag{26}$$

Note that the estimator will only change at failures times only. Thus we have

$$\widehat{H(t)}_{NA} = \sum_{u_j \leq t} \frac{d_j}{r_j} \tag{27}$$

This is called the Nelson-Aalen estimator for the cumulative hazard.
Alternatively using the KM estimator note that for small values of x, $e^x \cong 1 + x$. Taking $x = -\frac{d_j}{r_j}$ then we have

$$\begin{aligned} \widehat{S_{km}(t)} &= \prod_{j:u_j \leq t} exp(-\frac{d_j}{r_j}) \\ -\log(\widehat{S_{km}(t)}) &= \sum_{j:u_j \leq t} \frac{d_j}{r_j} \\ &= \widehat{H(t)}_{NA} \end{aligned} \tag{28}$$

Similar to the KM estimator the variance of the NA estimator is

$$\begin{aligned} Var(\widehat{H(t)}) &= Var\left(\sum_{j:u_j \leq t} \frac{d_j}{r_j}\right) \\ &= \sum_{j:u_j \leq t} Var(\hat{h}_j) \\ &= \sum_{j:u_j \leq t} \frac{h_j(1-h_j)}{r_j} \\ &= \sum_{j:u_j \leq t} \frac{d_j(r_j - d_j)}{r_j^3} \end{aligned} \tag{29}$$

Example 3.4. A school offers a one year course in a foreign language as an evening class. This is divided into three terms of 13 weeks each with one lesson per week. At the end of each lesson all the students sit a test and any that pass are awarded a qualification, and no longer attend the course.

Last year 33 students started the course. Of these 13 dropped out before completing the year, and 16 passed the test before the end of the year. The last lesson attended by the students who did not stay for the whole 39 lessons is shown in the table below along with their reason for leaving.

Number of students	Last lesson attended	Reason for leaving
5	1	Dropped out
1	6	Dropped out
2	7	Passed test
2	13	Dropped out
5	14	Passed test
6	27	Passed test
4	28	Dropped out
1	30	Dropped out
1	36	Passed test

Calculate the Nelson-Aalen estimate for the integrated hazard and estimate $Var(\widehat{H(15)})$.

4 Regression Models

Consider a case where for each individual under the study there is a defined $p \times 1$ vector $\mathbf{Z}$ of explanatory variables (covariates). These components of $\mathbf{Z}$ may represent various features thought to affect the failure times. Some of these covariates could be

- Treatments: In a simple situation one may wish to compare two treatments, a "new" and a "control " with a subject receiving treatment having a binary covariate 1 and those in the control having 0. If the treatment is specified by a dose the dosoge or log of the dosage may be another covariate
- Intrinsic properties of the subjects which may include demographic variables. eg age, body weight, etc.
- Exogenous properties of the subjects which may include the qualitative groupings e.g. martial status, lifestyle habits, etc.

Mostly it is convenient to define the vector $\mathbf{Z}$ so that $\mathbf{Z} = 0$ corresponds to a meaningful set of conditions and thus the model can developed in two parts; when $\mathbf{Z} = 0$ and when $\mathbf{Z} \neq 0$.
There are two broad classes of regression models; the **Accelerated Failure Time (AFT)** and **Proportional Hazards (PH)** models. We discuss the later only.

4.1 Proportional Hazards(PH) models

Here the hazard is of the form

$$h(t; Z) = h_0(t)\Psi(Z) \tag{30}$$

The function $\Psi(\mathbf{Z})$ links the failure times and the covariates, $h_0(t)$ is called the baseline hazard and it is required that $\Psi(0) = 1$. The function $\Psi(\mathbf{Z})$ can be parametrised as $\Psi(Z, \beta)$ where $\boldsymbol{\beta}$ is a $1 \times p$. A special case is when

$$\Psi(Z) = e^{\boldsymbol{\beta}'\mathbf{Z}} \tag{31}$$

Then equation(30) becomes

$$\begin{aligned} h(t; Z) &= h_0(t)e^{\boldsymbol{\beta}'\mathbf{Z}} \\ &= h_0 \exp(\beta_1 Z_1 + \beta_2 Z_2 + \cdots + \beta_p Z_p) \end{aligned} \tag{32}$$

This is called the Cox Proportional Hazards (PH) model. This is the most common model used for survival data due to its flexibility in the choice of covariates, fairly easy to fit and standard software exists. It also ensures that the hazard is always positive.
Why is it called proportional hazards?
Consider the example, where Z = 1 for treated and Z = 0 for control. Then taking $h_1(t)$ as the hazard rate for the treated group, and $h_0(t)$ as the hazard for control,then

$$h_1(t) = h(t; Z = 1) = h_0(t) \exp(\beta Z) = h_0(t) \exp(\beta)$$

Thus the ratio of the two hazards is a constant, ϕ, which does not depend on time, t. In other words, the hazards of the two groups remain proportional over time. The constant

$$\phi = \frac{h_1(t)}{h_0(t)} = \exp(\beta) \tag{33}$$

is called the **hazard ratio**
Generally if the i^{th} individual has a set of covariates $Z_i = (Z_{1i}, Z_{2i}, \cdots, Z_{pi})$, then the hazard rate will be some multiple of the baseline hazard rate:

$$h_i(t, Z_i) = h_0(t) \exp(\beta_1 Z_{1i} + \beta_2 Z_{2i} + \cdots + \beta_p Z_{pi}) \tag{34}$$

This means that the log of the hazard ratio for the i^{th} individual to the reference group as:

$$\log\left(\frac{h_i(t)}{h_0(t)}\right) = \beta_1 Z_{1i} + \beta_2 Z_{2i} + \cdots + \beta_p Z_{pi} \tag{35}$$

Thus the Cox Proportional Hazards model is a linear model for the log of the hazard ratio.
The major advantage of the framework of the Cox PH model is that the parameters $\boldsymbol{\beta}$ which reflect the effects the covariates can estimated without having to make any assumptions about the form of $h_0(t)$. That's what makes the model **semi-parametric**.
How do we estimate the model parameters?
The basic idea is that under PH, information about $\boldsymbol{\beta}$ can be obtained from the relative orderings (i.e. ranks) of the survival times, rather than the actual values. Suppose T follows a PH model:

$$h(t; Z) = h_0(t) \exp(\boldsymbol{\beta}'\mathbf{Z})$$

Now consider $T^* = g(T)$, where g is a monotonic increasing function. Then it can be shown that T^* also follows the PH model, with the same multiplier, $\exp(\beta' Z)$. Therefore, when considering likelihood methods for estimating the model parameters, only the ranks of the survival times are considered.
Suppose we observe (T_i, d_i, Z_i) for individual i, where T_i is a censored/ failure time random variable, d_i is the failure/censoring indicator (1=fail, 0=censor) and Z_i represents a set of covariates. Further suppose there are k distinct failure times, and let $u_1, \cdots, u_k$ represent the k ordered, distinct failure times. For now, assume there are no tied failure times and let $R(u_j) = j : u_j \geq t$ denote the set of individuals who are at risk for failure at time t. To estimate the parameters the partial likelihood is used.
Intuitively, it is a product over the set of observed failure times of the conditional probabilities of obtaining the observed failures, given the set of individuals at risk at those times. At each failure time u_j the contribution to the likelihood is:

$$\begin{aligned} L_j(\beta) &= P(individual\ j\ fails \mid one\ failure\ from\ R(u_j)) \\ &= \frac{P(individual\ j\ fails \mid at\ risk\ at\ u_j)}{\sum_{l \in R(u_j)} P(individual\ l\ fails\ at\ risk\ at\ u_j)} \\ &= \frac{h(u_j; Z_j)}{\sum_{l \in R(u_j)} h(u_j; Z_l)} \end{aligned} \tag{36}$$

Under the PH assumption, $h(t; Z) = h_0(t) \exp \boldsymbol{\beta}'\mathbf{Z}$, so we get:

$$\begin{aligned} L(\beta) &= \prod_{j=1}^{k} \left[\frac{h_0(u_j) \exp \beta Z_j}{\sum_{l \in R(u_j)} h_0(u_j) \exp \beta Z_l} \right] \\ &= \prod_{j=1}^{k} \left[\frac{\exp \beta Z_j}{\sum_{l \in R(u_j)} \exp \beta Z_l} \right] \end{aligned} \tag{37}$$

Note: Equation(37) is referred to as the partial likelihood as part of the full likelihood has been omitted. Only failures contributed to the likelihood. The partial likelihood is not a product of independent terms, but of conditional probabilities.

Example 4.1. Consider the following set of data

individual	u_j	d_i	Z_i
1	9	1	4
2	8	0	5
3	6	1	7
4	10	1	3

To obtain the partial likelihood, consider the first failure at $u_j = 6$, the risk set is { 1,2,3,4 } and the contribution to the likelihood is

$$\frac{\exp 7\beta}{\exp 4\beta + \exp 5\beta + \exp 7\beta + \exp 3\beta}$$

at $u_j = 9$, the risk set is { 1,4 } the contribution to the likelihood is

$$\frac{\exp 4\beta}{\exp 4\beta + \exp 3\beta}$$

at $u_j = 10$, the risk set is { 4 } the contribution to the likelihood is

$$\frac{\exp 3\beta}{\exp 3\beta} = 1$$

The partial likelihood is the product of these three terms.

Example 4.2. Consider the data below

Group 0: $4+$, 7 , $8+$, 9, $10+$ $Z_i = 0$

Group 1: 3, 5 , $5+$, 6, $8+$ $Z_i = 1$

Obtain the partial likelihood.

Remarks

1. The mle of β is obtained by maximising the partial likelihood. It is possible to do it manually when only one covariate is considered. Statistical software are used in case of higher number of covariates.

2. In practise the number of failures at time u_j may greater than one, i.e. there are ties. Here the Breslow's approximation used

$$L(\beta) = \prod_{j=1}^{k} \left[\frac{\exp \beta S_j}{\left[\sum_{l \in R(u_j)} \exp \beta Z_l \right]^{d_j}} \right] \tag{38}$$

where S_j is the sum of covariate's vectors Z and d_j is the number of ties observed to fail at u_j.

3. The partial likelihood behaves like the full likelihood i.e. it yields an estimator for β which is asymptotically multivariate normal and unbiased, whose variance can be estimated by the inverse of the information matrix. For the one parameter case when only one covariate is used then the mle of β is obtained by solving $\frac{\partial \ell(\beta)}{\partial \beta} = 0$ and $var(\hat{\beta}) = -\left[E\frac{\partial^2 \ell(\beta)}{\partial \beta^2}\right]^{-1}_{\beta=\hat{\beta}}$. Thus the $100(1-\alpha)\%$ confidence interval for β may be obtained.

4. In most practical situations several covariates may be present and the model fitting process involves selection of the significant ones. Therefore a criteria is required for assessing the effects of each covariates alone or in combination. The likelihood ratio criteria is mostly used. Suppose two models are fitted with p and $p+q$ covariates respectively. Let ℓ_p and ℓ_{p+q} be the maximised loglikelihood. Then the loglikelihood ratio statistic $-2[\ell_p - \ell_{p+q}] \sim \chi^2(q)$ under the hypothesis

$$H_0 : \beta_{p+1} = \beta_{p+2} = \cdots = \beta_{p+q} = 0$$

i.e. the extra q have no effect. Inorder to build a model one may start with a null model and add possible covariates one at a time or start with a full model and then eliminate the insignificant ones. It is also important to check the interaction between covariates.

Example 4.3. An investigation was undertaken into the effect of a new treatment on the survival times of cancer patients. Two groups of patients were identified. One group was given the new treatment and the other an existing treatment.The following model was considered:

$$h(t, Z) = h_0(t) \exp(\beta_1 Z_1 + \beta_2 Z_2)$$

where where $\beta_i \quad i = 1, 2$ are parameters and

$$Z_1 = \begin{cases} 1 & \text{if patient is male} \\ 0 & \text{if patient is female} \end{cases}$$

$$Z_2 = \begin{cases} 1 & \text{if new treatment} \\ 0 & \text{if old treatment} \end{cases}$$

The results of the investigation showed that, if the model is correct:
The risk of death for a male patient is 1.02 times that of a female patient; and
The risk of death for a patient given the existing treatment is 1.05 times that for a patient given the new treatment

(i) Estimate the value of the parameters

(ii) Estimate the ratio by which the risk of death for a male patient who has been given the new treatment is greater or less than that for a female patient given the existing treatment.

(iii) Determine, in terms of the baseline hazard only, the probability that a male patient will die within 3 years of receiving the new treatment.

Solution. The hazard for female patient

$$h_f(t, Z) = h_0(t) \exp(0 + \beta_2 Z_2)$$

The hazard for male patient

$$h_m(t, Z) = h_0(t) \exp(\beta_1 + \beta_2 Z_2)$$

We have

$$\begin{aligned} h_m(t, Z) &= 1.02 h_f(t, Z) \\ \exp(\beta_1 + \beta_2 Z_2) &= 1.02 \exp(\beta_2 Z_2) \\ &\Rightarrow \hat{\beta}_1 = 0.0198 \end{aligned}$$

Also

$$\begin{aligned} \exp(\beta_1 Z_1) &= \exp(\beta_1 Z_1 + \beta_2) \\ &\Rightarrow \hat{\beta}_2 = -0.0488 \end{aligned}$$

The hazard for a male patient who has been given the new treatment is:

$$\begin{aligned} h_{m,\ n}(t, Z) &= h_0(t) \exp(\beta_1 + \beta_2) \\ &= 0.9714 h_0(t) \end{aligned}$$

The hazard for a female patient given the existing treatment is the baseline hazard. Hence, the ratio of the hazard for a male patient who has been given the new treatment to that for a female patient given the existing treatment is:

$$\frac{h_{m,\ n}}{h_0(t)} = 0.9714$$

The probability of death is given by:

$$\begin{aligned} 1 - S_{m,\ n}(3) &= 1 - \exp\left\{ - \int_0^3 h_{m,\ n}(t, Z)\ dt \right\} \\ &= 1 - \exp\left\{ - \int_0^3 0.9714 h_0(t)\ dt \right\} \\ &= 1 - \exp\left\{ - \int_0^3 h_0(t)\ dt \right\}^{0.9714} \end{aligned}$$

Example 4.4. An education authority provides children with musical instrument tuition. The is concerned about the number of children giving up playing their instrument and is testing a new tuition method with a proportion of the children which it hopes will improve persistency rates. Data have been collected and a Cox proportional hazards model has been fitted for the hazard of giving up playing the instrument. Symmetric 95% confidence intervals (based upon standard errors) for the regression parameters are shown below.

(i) Write down the expression for the Cox proportional hazards model, defining all terms that you use.

Covariate	Confidence Interval
Instrument	
Piano	0
Violin	[-0.05,0.19]
Trumpet	[0.07,0.21]
Tuition method	
Traditional	0
New	[-0.15,0.05]
Sex	
Male	[-0.08,0.12]
Female	0

(ii) State the regression parameters for the fitted model.

(iii) Describe the class of children to which the baseline hazard applies.

(iv) Discuss the suggestion that the new tuition method has improved the chances of children continuing to play their instrument.

(v) Calculate, using the results from the model, the probability that a boy will still be playing the piano after 4 years if provided with the new tuition method, given that the probability that a girl will still be playing the trumpet after 4 years following the traditional method is 0.7.

Solution. The cox PH model is

$$h(t, Z) = h_0(t) \exp(\beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 Z_4)$$

where $h_0(t)$ is the baseline, $\beta_i \quad i = 1, 2, 3, 4$ are the model's parameters and

$$Z_1 = \begin{cases} 1 & \text{if student plays violin} \\ 0 & \text{otherwise} \end{cases}$$

$$Z_2 = \begin{cases} 1 & \text{if student plays trumpet} \\ 0 & \text{otherwise} \end{cases}$$

$$Z_3 = \begin{cases} 1 & \text{if student is taught using new tuition method} \\ 0 & \text{otherwise} \end{cases}$$

$$Z_4 = \begin{cases} 1 & \text{if student is male} \\ 0 & \text{if student is female} \end{cases}$$

The parameter values are $\beta_1 = 0.07$, $\beta_2 = 0.14$, $\beta_3 = -0.05$ and $\beta_4 = 0.02$. The baseline hazard refers to a female student,following traditional tuition method and playing the piano.
The parameter associated with the new tuition method is - 0.05. Because the parameter is negative, the hazard of dropping out is reduced by the new tuition method. Therefore the new tuition method does appear to improve the chances of a child continuing with his or her instrument. However the 95% confidence interval for the parameter includes zero. So at the 5% significance level it is not possible to conclude that the new tuition method has improved the chances of children continuing to play their instrument.

The hazard for a girl being taught the trumpet by the traditional method giving up is $h_0(t)\exp(0.14)$ .Therefore the probability of her still playing after 4 years is

$$\begin{aligned} S_f(4) &= \exp\left(-\int_0^4 h_0(t)\ \exp(0.14)\ dt\right) \\ &= \exp\left(-1.150274\int_0^4 h_0(t)\ dt\right) \end{aligned}$$

Since this is equal to 0.7, then

$$\int_0^4 h_0(t)\ dt = 0.310078$$

The hazard of giving up for a boy taught the piano by the new method is

$$h_0(t)\exp(-0.05+0.02) = h_0(t)\exp(-0.03)$$

.

Therefore the probability of him still playing after 4 years is

$$\begin{aligned} S_m(4) &= \exp\left(-\int_0^4 h_0(t)\ \exp(-0.03)\ dt\right) \\ &= \exp\left(0.310078\times 0.970446\right) \\ &= 0.74014. \end{aligned}$$

Example 4.5. A study is made of the impact of regular exercise and gender on the risk of developing heart disease among 50-70 year olds. A sample of people is followed from exact age 50 years until either they develop heart disease or they attain the age of 70 years. The study uses a Cox regression model.

(i) List reasons why the Cox regression model is a suitable model for analyses of this kind.

The investigator defined two covariates $Z_1 = 1$ if male, 0 if female and $Z_2 = 1$ if takes regular exercise, 0 otherwise. The investigator then fitted three models, one with just gender as a covariate, a second with gender and exercise as covariates, and a third with gender, exercise and the interaction between them as covariates. The maximised log-likelihoods of the three models and the maximum likelihood estimates of the parameters in the third model were as follows:

null model	-1,269
gender	-1,256
gender + exercise	-1,250
gender + exercise + interaction	-1,246
Covariate	Parameter
Gender	0.2
Exercise	-0.3
Interaction	-0.35

(ii) Show that the interaction term is required in the model by performing a suitable statistical test.

(iii) Interpret the results of the model.

Solution. Cox's model ensures that the hazard is always positive.
Standard software packages often include Cox's model.
Cox's model allows the general "shape" of the hazard function for all individuals to be determined by the data, giving a high degree of flexibility.
The data in this investigation are censored, and Cox's model can handle censored data.
In Cox's model the hazards of individuals with different values of the covariates are proportional, meaning that they bear the same ratio to one another at all ages.
If we are not primarily concerned with the precise form of the hazard, we can ignore the shape of the baseline hazard and estimate the effects of the covariates from the data directly.
A suitable statistical test is that using the likelihood ratio statistic. Comparing the model with gender + exercise with the model with gender + exercise + interaction. The log-likelihood for these two models are ℓ and $\ell_{interaction}$ respectively, then the test statistic is

$$-2(\ell - \ell_{interaction}) = 8$$

. Under the null hypothesis that the parameter on the interaction term is zero, this statistic has a chi-squared distribution with one degree of freedom (since the interaction term involves one parameter).
Since $8 > 7.879$, the critical value of the chi-squared distribution at the 0.5% level (or $8 > 3.84$ for the 5% level), the null hypothesis is rejected even at the 99.5% level (or 95% level) and conclude that the interaction term is required in the model.

The baseline category is females who do not take regular exercise. The hazards of developing heart disease in the other three categories, relative to the baseline category, are as follows:

Gender	Regular exercise	Hazard ratio
Male	No	$\exp(0.2) = 1.22$
Male	Yes	$\exp(0.2 - 0.3 - 0.35) = 0.64$
Female	Yes	$\exp(0.3) = 0.74$

Males who do not take regular exercise are more likely to develop heart disease than females.
Regular exercise decreases the risk of heart disease for both males and females.
The effect of regular exercise in reducing the risk of heart disease is greater for males than for females, so much so that among those who take regular exercise, males have a lower risk of developing heart disease than females.

Example 4.6. Consider the following set of data

individual(i)	T_i	Z_i
1	9*	0
2	9	0
3	8*	0
4	10*	1
5	13	0
6	12*	0
7	8	1
8	11	1

Construct the 95% confidence interval for the parameter.

5 Graduation and Statistical Tests

The crude mortality rates derived from a mortality investigation will not be the final rates that are published for use in actuarial calculations. The rates will have to pass through a further process called **graduation**.
Graduation refers to the process of using statistical techniques to improve the estimates provided by the crude rates. The aims of graduation are to produce a smooth set of rates that are suitable for a particular purpose, to remove random sampling errors (as far as possible) and to use the information available from adjacent ages to improve the reliability of the estimates. Graduation results in a "smoothing" of the crude rates.
Given the crude estimates , one often wants to know if they are consistent with another known experience to ascertain their consistency. For example, if it is the recent experience of the policyholders of a life insurance company, one might wish to know if the estimates are consistent with the company's own past experience, or is the experience changing. This could be important for pricing life insurance contracts. Also it could be necessary to know if they are consistent with the published(standard) life tables. This is important if the company plans to use published tables for any financial calculations. Here term consistency covers two concepts: the shape of the mortality curve over the range of ages and the level of mortality rates. When comparing with standard tables, the question is whether $\overset{\circ}{q}$ the crude estimates are consistent with the values of the standard table $\overset{s}{q}$. A statistical test for the hypothesis that the underlying mortality rates at each age x for the experience are the rates in the standard tables. Thus one will require

(i) the probabilistic model (multiple-state, Poisson or binomial);

(ii) the data (the observed numbers of deaths, the exposed-to-risks and our crude estimates $\overset{\circ}{q}$ and

(iii) a standard table.

Why graduate?
The crude estimates will progress more or less roughly, i.e. it is unlikely that the crude estimates will progress smoothly. This is largely because they have each been estimated independently and hence suffer independent sampling errors. The smaller the sample size the less smoothly the crude estimates are likely to progress. There are two arguments for graduation

1. **The theoretical argument**.
   The intuitive idea that quantities such as q_x should be smooth functions of age. There is some evidence from large investigations to support this, but it is nevertheless an assumption . It follows that a crude estimate at any age x also carries information about the values of q_{x-1} and q_{x+1}. For example, if q_x is smooth and not changing too rapidly, then $\overset{\circ}{q}_x$ should not be too far away from estimating q_{x-1} and q_{x+1} , as well as being the "best" estimate, in some sense, of q_x . By smoothing, we can make use of the data at adjacent ages to improve the estimate at each age. Smoothing reduces the sampling errors at each age.

2. **The practical argument**
   The mortality data will be used in life tables to compute financial quantities, such as premiums for life insurance contracts. It is very desirable that such quantities progress smoothly with age, since irregularities (jumps or other anomalies) are

hard to justify in practice. These quantities could be calculated using the crude mortality rates, and then smooth the premium rates etc directly, but it is much more convenient to have smoothed mortality rates to begin with. One would never, in any case, apply the results of a mortality experience directly to some financial problem without considering carefully its suitability. This means comparing it with other relevant experiences and tables, not just in aggregate but over age ranges of particular financial significance. It is often the case that a mortality experience must be adjusted in some way before use, in which case there is little point in maintaining the roughness of the crude estimates.

Note: The crude estimates of mortality $\overset{\circ}{q}_x$ provide an estimate of the true underlying mortality for a particular age. However, since the believe is that the underlying rates of mortality will follow a smooth curve as the age varies, the additional information provided by the numbers of deaths at nearby ages can be used to improve the estimate. This process of applying statistical techniques to improve the estimates provided by crude rates over a range of ages is called **graduation**.
Graduation cannot remove any bias in the data arising from faulty data collection or otherwise. It produces results as reliable as the original data. The aims of graduation are to:

- produce a smooth set of rates that are suitable for a particular purpose.
- remove random sampling errors.
- use the information available from adjacent ages.

What are the desirable features of a graduation?
Any graduation should produce smoothed quantities but it should be such that it adheres to the data and is suitable for the purpose at hand. Smoothness and adherence to data are usually conflicting requirements. Perfect smoothness (extreme example: a straight line) pays no attention to the data, while perfect adherence to the data means no smoothing at all. If the graduation process results in rates that are smooth but show little adherence to the data, then the data may be **overgraduated** but if insufficient smoothing is done but there is a better adherence to the data the there is **undergraduation**. Thus one needs to get a satisfactory compromise between the two. To do this we need to test for smoothness and adherence to the data.
The suitability of a graduation for practical work depends very much on what that work is, and can only be assessed in particular cases. For example in life insurance work, losses result from premature deaths (benefits are paid sooner than expected) so mortality must not be underestimated while in pensions or annuity work, losses result from delayed deaths (benefits are paid for longer than expected) so mortality must not be overestimated .

5.1 Testing the smoothness of a graduation

The test for smoothness is also used as a check for undergraduation. It is possible to fit a high-order polynomial to any set of observed data. The fitted polynomial is smooth in the mathematical sense, i.e it is differentiable many times, but it does not progress smoothly from age to age. To test for smoothness, the third differences of the graduated

quantities($\overset{\circ}{q}_x$) is calculated. i.e.

$$\begin{aligned} \text{The first difference } \triangle \overset{\circ}{q}_x &= \overset{\circ}{q}_{x+1} - \overset{\circ}{q}_x \\ \text{The second difference } \triangle^2 \overset{\circ}{q}_x &= \triangle \overset{\circ}{q}_{x+1} - \triangle \overset{\circ}{q}_x \\ \text{The third difference } \triangle^3 \overset{\circ}{q}_x &= \triangle^2 \overset{\circ}{q}_{x+1} - \triangle^2 \overset{\circ}{q}_x \end{aligned} \tag{39}$$

The third differences measure the change in curvature. The criterion of smoothness usually used is that the third differences of the graduated quantities $\overset{\circ}{q}_x$ should: be small in magnitude compared with the quantities themselves; and progress regularly.
Note: How to judge if this criterion is met takes some practice. However, since most methods of graduation now in use automatically give smooth results, this is not of great importance. Only the graphical method presents difficulties in achieving smoothness.

5.2 Testing adherence to data

The tests of adherence to data have much in common with the statistical tests of an experience against a standard table. They rely on the assumption that the true parameters of the underlying probability model are the graduated estimates. The null hypothesis is H_0 : the true underlying mortality rates at each age x for the experience are the graduated rates.
Several tests are available

5.2.1 χ^2 tests

Used to decide whether the observed numbers of individuals who fall into specified categories are consistent with a model that predicts the expected numbers in each category. It is a test for overall *goodness of fit.* It is based on the statistic $\chi^2 = \sum_{all\ i} \frac{(O_i - E_i)^2}{E_i}$
where O_i is number observed in the i^{th} category and E_i is the expected number predicted by the assumed probabilities and the sum is over all possible categories. Each term in the sum represents the square of the discrepancy between the actual and expected values for one group (with an appropriate weighting factor applied). A high value for the total indicates that the overall discrepancy is quite large and would lead to the rejection of the model. A low value indicates that the observed data fit the model well.
Note: This parameter in the χ^2 distribution reflects the amount of "freedom" present in the system. The correct number of degrees of freedom to use in a test depends on the number of constraints that restrict the way individuals can be allocated to the different categories. If the groups form a set of mutually exclusive and exhaustive categories (so that their probabilities must add up to 1) or the expected numbers for each category were determined based on the total number for all groups, then subtract 1 from the total number of categories to the required df. Subtract a further 1 for each parameter that has been estimated. Additionally if the expected number in a group is small (less than 5 say), a difference of just one person can make a big difference to the value of $\triangle^3 q_x$ and the approximation becomes unreliable. This problem can be overcome by combining the expected and actual numbers in small groups.

Example 5.1. The mortality rates for a population for the age range 30-34 were estimated by fitting a straight line $a + bx$ on the crude values of $\log \frac{q_x}{p_x}$. Test whether this

model (with estimated parameter values of $a = -10.9446$ and $b = 0.110404$) can be considered to give a good fit to the data shown in the table below given that the initial exposed to risk in was approximately 700,000 at each age.

Age x	30	31	32	33	34
Number of deaths	335	391	428	436	458

Solution.

$$\begin{aligned} \log \frac{q_x}{p_x} &= a + bx \\ \Rightarrow \overset{\circ}{q}_x &= \frac{1}{1 + \exp -(a + bx)} \end{aligned}$$

Age x	$\overset{\circ}{q}_x$	O_x	E_x	$\frac{(O_x - E_x)^2}{E_x}$
30	0.0004842	335	338.96	

5.2.2 Tests based on standardised deviations

Here the statistical tests will be based on the assumption that:

(a) the numbers of deaths at different ages are independent

(b) When comparing the experience with a standard table, the number of deaths at age x, $D_x \sim N(E_x, \overset{s}{\lambda}_{x+1/2})$ in case the number of deaths follow the Poisson model and $D_x \sim N(E_x, \overset{s}{q}_x)$ for the binomial model.

(c) When testing the adherence to data of a graduation, the number of deaths at age x, $D_x \sim N(E_x, \overset{\circ}{\lambda}_{x+1/2})$ in case the number of deaths follow the Poisson model and $D_x \sim N(E_x, \overset{\circ}{q}_x$ for the binomial model.
In parts (b) and (c) of the assumption, the Normal distribution is used as an approximation to the Poisson distributions with small intensity and large exposure, and as an approximation to the Binomial distribution (which is acceptable if mean is greater than 5).

Definition 5.1. The deviation at age x is given as

$$\begin{aligned} & \text{Actual deaths} - \text{Expected deaths} \\ = \; & d_x - E_x \overset{\circ}{\lambda}_{x+1/2} \quad \text{for poisson model} \\ = \; & d_x - E_x \overset{\circ}{q}_x \qquad \text{for binomial model} \end{aligned} \tag{40}$$

and the standardised deviation,

$$z_x = \begin{cases} \dfrac{d_x - E_x \overset{\circ}{\lambda}_{x+1/2}}{\sqrt{d_x - E_x \overset{\circ}{\lambda}_{x+1/2}}} & \text{for poisson model} \\ \dfrac{d_x - E_x \overset{\circ}{q}_x}{\sqrt{E_x \overset{\circ}{q}_x}} & \text{for binomial model} \end{cases} \tag{41}$$

Note: When comparing with a standard table the superscript $\circ$ in equations(40) and (41) will be replaced by s .The $z'_x s$ are referred to as individual **standardised deviations**. If $q \approx 0$ then $(1-q) \approx 1$
Assuming that there is a sufficient number of (independent) lives at each age x, then the hypotheses, under all the models, may be replaced with the following, by virtue of the Central Limit Theorem:-

- $z_x \sim N(0,1)$ for all x
- The $z'_x s$ at different ages are mutually independent.

The following tests will then be based on the individualised standardised deviations

5.2.3 Chi square test

An important test whether we are comparing an experience with a standard table, or testing the adherence to data of a graduation.
The test is used to assess whether the observed numbers of deaths at each age are consistent with a particular set of graduated mortality rates or a particular graduation formula. The test will indicate overall goodness of fit. A high value of the test statistic indicates that the discrepancies between the observed numbers and those predicted by the graduated table are large, i.e. the fit is not very good. This may be because of overgraduation.
The following assumptions are made

1. The mortality rates is homogeneous within each age group and lives are independent.
2. The expected numbers of deaths are high enough (usually at least 5 in each cell) for the chi square approximation to be valid.

To perform the test combine any small groups by pooling the actual and expected deaths, so that the expected number of deaths is never less than 5.
Compute the $\chi^2 = \sum_{all\ x} z_x^2$. When comparing an experience with a standard table, then χ^2 can be assumed to have degrees of freedom equal to the number of age groups. When testing the adherence to data of a graduation, the statistic has fewer than the no. of age groups degrees of freedom. How many fewer depends on the method of graduation.
If the chi square statistic is large then this indicates a poor fit or overgraduation. The contributions to the statistic from each term in the sum can be used to identify the ages where the fit is worst. Note that if the statistic is very low, this may indicate undergraduation. However, if undergraduation is suspected then one will usually test for it in other ways.
Remarks:
This is a good test for overall goodness of fit but fails to detect several defects that could be of considerable financial importance.

- There could be a few large deviations offset by a lot of very small deviations. The test could be satisfied although the data do not satisfy the distributional assumptions that underlie it as the statistic summarises a lot of information in a single figure.

- The graduation might be biased above or below the data by a small amount. The statistic will fail to detect consistent bias if it is small.
- Even if the graduation is not biased as a whole, there could be significant groups of consecutive ages (called runs or clumps) over which it is biased up or down. The statistic will not detect this.
- As the because the statistic is based on squared deviations, it has no information about the direction of any bias or the nature of any lack of adherence to data of a graduation, even if the bias is large or the lack of adherence manifest.

Example 5.2. The actuary to a large pension scheme has attempted to graduate the scheme's recent mortality experience with reference to a table used for similar sized schemes in a different industry. He has calculated the standardized deviations between the crude and the graduated rates, z_x, at each age and has sent you a printout of the figures over a small range of ages. Unfortunately the dot matrix printer on which he printed the results was very old and the dots which would form the minus sign in front of numbers no longer function, so you cannot tell which of the standardized deviations is positive and which negative. Below are the data which you have.

Age	60	61	62	63	64	65	66	67	68	69	70
z_x	2.40	0.08	0.80	0.76	1.04	0.77	1.30	1.76	0.28	0.68	0.93

Carry out an overall goodness-of-fit test on the data. Comment on your result.

5.2.4 The individualised standardised deviations test

The test looks at the distribution of the values of the standardised deviations. Under the hypothesis, the $z_x's$ comprise independent sample values from a standard normal distribution. Thus test just assess the normality of the deviations.
If the graduated rates are not a good fit, the distribution will not be "centred correctly". If there is heterogeneity within the age groups or deaths are not independent, the variance will be (respectively) smaller or greater than expected if our underlying model were correct. If there is undergraduation, then it is expected that the standardised deviations are tightly bunched. Conversely, if we have overgraduation, it is expected the standardised deviations to be too spread out.
The test is performed as follows

- Calculate the standardised deviation z_x for each age or age group.
- Divide the real (number) line into any convenient intervals (the more age groups, the more intervals it might be reasonable to use) where the intervals at either end are $(-\infty, -3]$ and $[3, \infty)$. Count the number of standardised deviations falling into each of the ranges.
- Compare the observed number of the z_x that fall in each interval and the expected number that should fall in each interval, under the hypothesis that $z_x \sim N(0, 1)$.
- To formalise the comparison, use can form a Chi square test.

If the number of age groups is small, then smaller number of intervals are used to ensure that the expected number of standardised deviations in each interval is not less than five.

If there are only a few age groups, a test must be carried out "by eye", by considering the features of the normal distribution:

Overall shape

The number of values in each of the ranges should conform broadly with the percentages for the normal distribution.

Absolute deviations

Ff the $z'_x s$ comprise independent samples from a standard normal, half of them should lie in the interval (- 2/ 3 , 2 /3). If there are a lot of values in the tails (i.e. the absolute deviations are too big), this indicates overgraduation. In this case a one-tailed test is appropriate, as one usually wishes only to identify instances where the number of absolute deviations exceeding 2/3 is large. The null hypothesis (of no difference between the standard table and the mortality underlying the experience, or of no difference between the graduated rates and the mortality underlying the experience) is rejected for large values of the z statistic.

Outliers

If the $z'_x s$ are N (0,1) ,then individual values with absolute values greater than 1.96 should form at most 1 in 20 of the whole set, and there should be only 1 in 100 with an absolute value greater than 2.57.

Symmetry

There should be roughly equal numbers of positive and negative standardised deviations (since the normal distribution is symmetrical). An excess of positive values indicates that the graduation has introduced a positive bias (ie the graduated rates are too low). An excess of negative values indicates that the graduation has introduced a negative bias (ie the graduated rates are too high).

Example 5.3. Analyse the distribution of the following standardised deviations

1.20, 1.57, -0.57, -1.11, -2.31, 1.46, 0.71, -0.37, 0.44, -0.48, 0.48, -0.43, 0.51, -0.83, 0.64, -0.02, -1.03, 1.07, -0.36, 0.04

Interval	$(-\infty, -3)$	(-3, -2)	(-2, -1)	(-1, 0)	(0,1)	(1, 2)	(2,3)	$(3, \infty)$
Actual	0	2	4	4	3	4	3	0
Expected	0.0	0.4	2.8	6.8	6.8	2.8	0.4	0

Solution. There are only 7 values in the range (-2/3, 2/3) . So, there appear to be too few values in the centre of the distribution and too many in the tails ie the variance is greater than predicted by the binomial model. This might indicate overgraduation (an inappropriate graduation formula).

The values are symmetrical (10 positive and 10 negative). So, this shows no evidence of bias in the graduated rates.

If the small groups are combined by pooling the values in the ranges $(-\infty, -1)$ and $(1, \infty)$, the Chi square test to the resulting 4 groups is 10.24. This exceeds 7.815, the upper 95% point of the chi square distribution with 3 degrees of freedom, which confirms that the deviations do not conform to a standard normal distribution. Note that, strictly speaking, chi square-test should not be used even with this broad grouping since one of the expected value is less than 5.

5.2.5 Signs test

A simple test for overall bias, i.e. whether the graduated rates are too high or too low. It identies the second deficiency of the Chi square test, i.e. failure to detect where there

is an imbalance between positive and negative deviations.
If the graduated rates do not tend to be higher or lower than the crude rates on average, it is expected that roughly half the graduated values to be above the crude rates and half below. So, if there are m age groups, the number above (or below) should have a $b(m, 1/2)$ distribution. An excessively high number of positive or negative deviations will indicate that the rates are biased. It is a two-tailed test, ie we are looking for both positive and negative bias.
The test is performed as follows

- Count how many of the graduated rates lie above/below the crude rates. This by looking at the signs of the individual standardised deviations.
- Calculate the probability value for the test by finding the probability of obtaining a split of positive/negative values as extreme as observed. Define the test statistic:

  $$R = \text{Number of } z_x \text{ that are positive/ negative.}$$

  Under the hypothesis, $R \sim b(m, 1/2)$, so the probability function of R is:

  $$P(R = r) = \left(\begin{array}{c} m \\ r \end{array} \right) 0.5^m, \quad r = 0, 1, 2, \cdots, m \tag{42}$$

  An excess of either negative or positive deviations is a defect, so a two- tailed test is used. Since the binomial distribution is discrete we find k * , defined as the smallest value of k for which

  $$\sum_{j=0}^{m} \left(\begin{array}{c} m \\ j \end{array} \right) 0.5^m \geq 0.025 \tag{43}$$

  The test would be satisfied (at the 5% level) if $k* \leq R \leq m - k*$. Another way to carry out the test is to calculate its p -value. If the number of age groups is large, the normal approximation can be used.

If the test shows that the number of positive values is very high or very low, this indicates that the rates are on average too low or too high (respectively). An examination of the pattern of the signs will indicate the range of ages where the bias is worst.
Note: Just looking at the signs of the deviations provides no indication of the extent of the discrepancy. This test is qualitative rather than quantitative.

Example 5.4. A graduation covers 20 age groups and has resulted in 6 positive and 14 negative deviations. Carry out a signs test on these data.

Solution. Under the null hypothesis, $R \sim b(20, 1/2)$. The p -value of the test is:

$$p = 2P(R \leq 6) = 2 \times 0.0577 = 0.1154$$

Since this is greater than 5%, there is insufficient evidence to reject the null hypothesis at the 5% significance level.

5.2.6 Cumulative deviations

Tests whether the overall number of deaths conforms to the model with the mortality rates assumed in the graduation. It can detect overall goodness of fit. Where the fit is not good this may be due to heterogeneity. It addresses the problem of the inability of the chi square test to detect a large positive or negative cumulative deviation over part (or the whole) of the age range. It detects overall bias or long runs of deviations of the same sign.

Consider the hypothesis, $D_x \sim N(E_x,\ \overset{\circ}{\lambda}_{x+1/2})$ then the deviation has (approximate) distribution:

$$d_x - E_x\ \overset{\circ}{\lambda}_{x+1/2} \sim N(0,\ E_x\ \overset{\circ}{\lambda}_{x+1/2})$$

So the accumulated deviation, over the whole age range, has distribution

$$\sum_{all\ x}(d_x - E_x\ \overset{\circ}{\lambda}_{x+1/2}) \sim N\left(0,\ \sum_{all\ x} E_x\ \overset{\circ}{\lambda}_{x+1/2}\right)$$

and

$$\frac{\sum_{all\ x}(d_x - E_x\ \overset{\circ}{\lambda}_{x+1/2})}{\sqrt{\sum_{all\ x} E_x\ \overset{\circ}{\lambda}_{x+1/2}}} \sim N(0,1)$$

This can be tested in the usual way, using a two-tailed test, since either positive or negative deviations are of concern. If the magnitude (ie the absolute value) of the calculated test statistic is high, this indicates that either: the graduated rates are biased (too low if the test statistic is positive, too high if the test statistic is negative), or the variance is higher than predicted by the binomial or Poisson model for the range of ages considered. Note that the test can only detect features that are present over the whole age range considered. An excess of positive deviations over one age range may "cancel out" an excess of negatives over another range.

5.2.7 Grouping of signs test

The test (also called Stevens' test) detects "clumping" of deviations of the same sign. It looks at the number of groups (or runs) of deviations of the same sign and compares this with the number that would be expected if the positive and negative signs were arranged in random order.

If the graduated rates are overgraduated, the standardised deviations will not swap from positive to negative very often and there will be fewer runs than expected. If the rates are undergraduated, the standardised deviations will swap from positive to negative very often and there will be more runs than expected. However, we do not usually use this test to look for undergraduation. So it is a one-sided test as we are worried about a low number of groups.

Define the test statistic:

$$G = \text{Number of groups of positive } z_x \text{ 's.}$$

Also, suppose that of the m deviations, n_1 are positive and n_2 are negative. The hypothesis is that the given n_1 positive deviations and n_2 negative deviations are in random order. Thus the probability that the number of positive groups will be at least G given n_1 and n_2 is computed .
Let $t \leq G$.

- There are (n_2+1) places in which the t positive groups can be located: before the first negative sign, after the last negative sign or in any of the (n_2-1) gaps between the signs. Thus there are $\binom{n_2+1}{t}$ ways to arrange t positive groups among n_2 negative.
- There are $\binom{n_1-1}{t-1}$ ways to arrange n_1 positive signs into t positive groups.
- There are $\binom{m}{n_1}$ ways to arrange n_1 positive and n_2 negative signs. Hence, the probability of exactly t positive groups is

$$\frac{\binom{n_2+1}{t}\binom{n_1-1}{t-1}}{\binom{m}{n_1}} \tag{44}$$

The procedure for the test is as follows

- Determine the sign of the deviation at each age and count the number of groups of positive signs $(= G)$.
- Calculate the probability value for the test by finding the probability of obtaining a number of groups as extreme as observed. Since every pair of positive groups must be separated by a negative group, the numbers of positive and negative groups will be small or large alike, so a one- tailed test is appropriate. Hence one should find the smallest k such that

$$\sum_{t=1}^{k}\left(\frac{\binom{n_2+1}{t}\binom{n_1-1}{t-1}}{\binom{m}{n_1}}\right) \geq \alpha \tag{45}$$

  and reject the hypothesis (at the $\alpha\%$ level) if $G < k$.

Alternatively, one could obtain the critical values from the Tables. If the number of groups of positive deviations is less than or equal to the critical value given in the Tables, the the null hypothesis. rejected.

Example 5.5. A graduation covers 20 age groups. The number of positive deviations is 6, and the number of groups of positive deviations is 2. Carry out a grouping of signs test using these data.

Solution. From the tables, the critical value is 2 when $n_1 = 6$ and $n_2 = 14$. Since there are 2 groups of positive deviations, the null hypothesis is reject at the 5% significance level and conclude that there is evidence of grouping of deviations of the same sign.

Note: For large values of groups the normal approximation may be used, with

$$G \sim N\left(\frac{n_1(n_2+1)}{n_1+n_2}\ , \frac{(n_1\ n_2)^2}{(n_1+n_2)^3}\right) \tag{46}$$

If there are too few runs, this indicates that the rates are overgraduated. The rates do not adhere closely enough to the crude data and may be consistently too high or too low over certain parts of the table.
When applying this test, the choice to count the positive group over negative one is arbitrarily. This test can, in some cases, lead to different conclusions depending on whether positive or negative groups are considered.

5.2.8 Serial correlations test

The serial correlations test detects grouping of signs of deviations by analysing the relationship between the deviations at nearby ages, taking into account the magnitude of the values. The test will address the problem of the inability of the chi square test to detect excessive clumping of deviations of the same sign.
If the graduated rates are neither overgraduated nor undergraduated, one would expect the individual standardised deviations at consecutive ages to behave as if they were independent. However, if the graduated rates are overgraduated, the graduated mortality curve will tend to stay the same side of the crude rates for relatively long periods and, although there will be random variations in the numbers of deaths, one would expect the values of consecutive deviations to have similar values, i.e. they will be positively correlated.
(Conversely, if the rates are undergraduated, the graduated curve will cross the crude rates quite frequently and the values of consecutive deviations will tend to oscillate, i.e. they will be negatively correlated. However this test is one sided to test for overgraduation since undergraduation would be tested by means of the smoothness test.) If correlations are present, then the effect are expected to be strong at adjacent ages or at ages separated by 2 or 3 years.
Under the null hypothesis, the two sequences of lag 1

$$z_1, z_2, \cdots, z_{m-2}, z_{m-1}$$
$$z_2, z_3, \cdots, z_{m-1}, z_m$$

should be uncorrelated. Also the lag 2 sequence

$$z_1, z_2, \cdots, z_{m-3}, z_{m-2}$$
$$z_3, z_4, \cdots, z_{m-1}, z_m$$

The correlation coefficient of the j th lagged sequences is

$$r_j = \frac{\sum_{i=1}^{m-j}(z_i - \bar{z}^{(1)})(z_{i+j} - \bar{z}^{(2)})}{\sqrt{\sum_{i=1}^{m-j}(z_i - \bar{z}^{(1)})^2 \sum_{i=1}^{m-j}(z_{i+j} - \bar{z}^{(2)})^2}} \tag{47}$$

where $\bar{z}^{(1)} = \dfrac{1}{m-j}\sum_{i=1}^{m-j} z_i$ and $\bar{z}^{(2)} = \dfrac{1}{m-j}\sum_{i=1}^{m-j} z_{i+j}$

A positive value of r_j indicates that nearby values of z_x tend to have similar values, whereas a negative value indicates that they tend to have opposite values.

If m is large enough, then $\bar{z}^{(1)}$ and $\bar{z}^{(2)}$ can be approximated by $\bar{z} = \dfrac{1}{m}\sum_{i=1}^{m} z_i$ so that equation(47) becomes

$$\begin{aligned} r_j &\approx \frac{\sum_{i=1}^{m-j}(z_i-\bar{z})(z_{i+j}-\bar{z})}{\frac{m-j}{m}\sum_{i=1}^{m}(z_i-\bar{z})^2} \\ &\approx \frac{\frac{1}{m-j}\sum_{i=1}^{m-j}(z_i-\bar{z})(z_{i+j}-\bar{z})}{\frac{1}{m}\sum_{i=1}^{m}(z_i-\bar{z})^2} \end{aligned} \tag{48}$$

Note that equation(48) is just the ratio of two averages.

Under the null hypothesis $r_j \sim N(0, 1/m)$. Hence, $r_j\sqrt{m}$ (the T ratio) can be tested against the standard Normal distribution. If the T ratio is "too positive", this indicates that the rates are overgraduated. The rates do not adhere closely enough to the crude data and may be consistently too high or too low over certain parts of the table.

Note: The serial correlation test is a parametric test, ie it takes into account actual numerical values, whereas the sign test and grouping of signs test are nonparametric, since they only look at how many. Because the serial correlation test takes into account the numerical values of the deviations, it is possible for correlations in one part of the age range to be cancelled out by opposite correlations in another part. This means that the signs test and grouping of signs test are usually more powerful, ie they are more likely to detect overgraduation if this is present.

Example 5.6. Calculate the serial correlation coefficients and the T ratio for Example 5.3 with lag 1 and interpret your result.

Solution. The mean $\bar{z}$ of the individual standardised deviations is 0.19. The average of the squared deviation is 2.13 and that of the cross products is 0.94 so that the $r_1 = \dfrac{0.94}{2.13} = 0.44$. The $Tratio = \sqrt{20}r_1 = 1.97$.

The T ratio is positive, which might suggest the rates are overgraduated. Its value of1 is just above the upper 2.5% point of the normal distribution.

6 Sample Past exams papers

6.1 Paper I

QUESTION ONE (30 marks) (COMPULSORY)

(a) Define the following as used in survival analysis

(i) Survivor function. **[1 mark]**

(ii) Hazard function. **[1 mark]**

(b) An investigation was undertaken into the length of post-operative stay in hospital after a particular type of surgery. All patients undergoing this surgery between 1^{st} and 31^{st} January 2013 were observed until either they left the hospital, died, or underwent a second operation. The event of interest was leaving the hospital. Patients who died or underwent a second operation during the period of investigation were treated as censored at the date of death or second operation respectively. The investigation ended on 28 February 2013, and patients who were still in the hospital at that time were treated as censored.
State, with reasons, whether the following types of censoring are present in this investigation:

- Type I
- Type II
- Random

Comment on whether censoring in this investigation is likely to be informative.

[4 marks]

(c) For a particular investigation the hazard of mortality is assumed to take the form:

$$\lambda(t) = a + bt$$

where a and b are constants and t represents time.
For each life i in the investigation ($i = 1, \ldots, n$) information was collected on the length of time the life was observed t_i and whether the life exited due to death ($d_i = 1$ if the life died, 0 otherwise).

(i) Show that the likelihood of the data is given by:

$$L = \prod_{i=1}^{n} (a + bt_i)^{d_i} \exp(-at - \frac{1}{2} bt_i^2)$$

[3 marks]

(ii) Derive two simultaneous equations from which the maximum likelihood estimates of the parameters a and b can be calculated. **[4 marks]**

(d) An Institute conducts tuition classes as part of their preparation for the professional exams. The management of the institute is concerned with the withdrawal rates of the students and hence it is testing a new tuition method to improve the persistency rates. Data have been collected and a Cox proportional hazards model has been fitted for the hazard of students leaving the course. Symmetric 95% confidence intervals (based upon standard errors) for the regression parameters are as shown below.

		Parameter Confidence Interval
Course	Engineering	0
	Medical	[0.08, 0.25]
Tuition method	Traditional	0
	New	[-0.05, 0.05]
Sex	Boys	[0.02, 0.12]
	Girls	0

(i) Write down the expression for the Cox proportional hazards model used, defining all terms that you use. **[4 marks]**

(ii) Describe the class of students to which the baseline hazard applies. **[1 mark]**

(iii) Discuss the suggestion that the new tuition method has improved the chances of students continuing with the tuition classes. **[2 marks]**

(e) The following are remission times(in weeks) of Leukemia patients.

6 9 13 13* 18 23 28 31 34 45* 48 56

where a right-censored observation is denoted by *. Assuming that the observation have exponential distribution with parameter μ, obtain the 95% confidence interval for the estimate of μ. **[6 marks]**

(f) If a right-censored observation is denoted by *, derive the Nelson-Aalen estimate of the cumulative hazard for the following data.

1* 3 4* 5 5 6* 7* 7 7

[4 marks]

QUESTION TWO (20 marks) (Optional)

(a) A random variable, T, has the Weibull distribution with probability density function

$$f(t) = \begin{cases} \lambda\gamma t^{\gamma-1}\exp(-\lambda t^{\gamma}), & t > 0 \quad \lambda > 0 \quad \lambda > 0 \\ 0, & \text{otherwise} \end{cases}$$

(i) Derive the survivor function, $S(t)$, of T. **[3 marks]**

(ii) Show that $\log\{-\log(S(t))\}$ is a linear function of $\log(t)$. State the intercept and slope of the line if $\log\{-\log(S(t))\}$ on the vertical axis is plotted against log(t) on the horizontal axis. **[4 marks]**

(b) 20 refrigerator motors of a particular type were each tested on an accelerated life test, and their times till first failure (hours) were recorded. The results are listed below, where * denotes that the motor was still functioning properly when the test was brought to an end.

2 4* 5 5 5 6* 7 7 7* 8 8 9* 11 11 12 12* 12* 12* 16 18*

(i) Use the Kaplan-Meier method with these data to estimate the survivor function, $S(t)$, for the time to first failure of a motor of this type. **[5 marks]**

(ii) Referring to the result from part (a)(ii) and (b)(i), use a suitable graphical method to investigate whether or not these data come from a Weibull distribution. **[5 marks]**

(iii) Draw a straight line through the points on your graph by eye and use it to estimate the parameters, λ and γ, of a Weibull distribution fitted to these data. **[3 marks]**

QUESTION THREE (20 marks) (Optional)

154 subjects with burns were monitored in a study of a new treatment to prevent burn wounds becoming infected. 70 of the subjects were given standard care (Treat = 0) while the other 84 had additional care thought to make infection less likely (Treat = 1). The time (in days) until the wound became infected was recorded; for 106 of the subjects, no infection was discovered during the period of follow up and for them total time in the study was treated as a censored survival time. Further information recorded about each subject included their Sex (Male = 0, Female = 1), Race (Non White = 0, White = 1) and Severity (the initial severity of their burns, in percent of body surface area).

(a) It was decided to fit a Cox proportional hazards model to the data, with Treat, Sex, Race and Severity as explanatory variables. Write down the form of this model, interpreting clearly each of the terms in it. **[5 marks]**

(b) The model was fitted and the results shown below were obtained.

	Coefficient	Standard Error
Treat	0.606	0.296
Sex	0.631	0.390
Race	2.12	1.01
Severity	0.00404	0.00703

(i) What can be concluded about the effects of the four explanatory variables? **[8 marks]**

(ii) Obtain and interpret a 95% confidence interval for the hazard ratio of a female subject given additional care compared to a female subject given standard care, assuming that the two subjects are white and have the same initial severity of burns. **[5 marks]**

(iii) Information was also recorded about the cause of the burns, which was characterised as either chemical (9 cases), scald (18), electric (11) or flame (116). Describe briefly how you would extend the model in order to make full use of this new information. **[2 marks]**

QUESTION FOUR (20 marks) (Optional)

(a) State and explain any two features which are desirable when a graduation is performed. **[4 marks]**

(b) The actuary to a large pension scheme has attempted to graduate the schemes recent mortality experience with reference to a table used for similar sized schemes in a different industry. He has calculated the standardized deviations between the crude and the graduated rates, z_x, at each age and has sent you a printout of the figures over a small range of ages. Unfortunately the dot matrix printer on which he printed the results was very old and the dots which would form the minus sign in front of numbers no longer function, so you cannot tell which of the standardized deviations is positive and which negative. Below are the data which you have.

Age	60	61	62	63	64	65	66	67	68	69	70
z_x	2.40	0.08	0.80	0.76	1.04	0.77	1.30	1.76	0.28	0.68	0.93

(i) Carry out an overall goodness-of-fit test on the data. Comment on your result. **[5 marks]**

(ii) List four defects of a graduation which the test you have carried out would fail to detect. For each of the defects, suggest a test which could be used to detect it. **[8 marks]**

(iii) Carry out one of the tests suggested in part(ii). **[3 marks]**

QUESTION FIVE (20 marks)(Optional)

A study was made of a group of people seeking jobs. 700 people who were just starting to look for work were followed for a period of eight months in a series of interviews after exactly one month, two months, etc. If the job seeker found a job during a month, the job was assumed to have started at the end of the month. Unfortunately, the study was unable to maintain contact with all the job seekers.
The data from the study are shown in the table below:

Months since start of study	Found employment	Contact lost
1	100	50
2	70	0
3	50	20
4	40	20
5	20	30
6	20	60
7	12	38
8	6	0

(i) Describe two types of censoring present in the investigation and an example of a person to whom each type applies. **[4 marks]**

(ii) Calculate the Nelson-Aalen estimate of the integrated hazard for the job seekers and hence give estimate of the survival function. **[7 marks]**

(iii) Estimate the variance of Nelson-Aalen estimate at the fourth month and hence its 95% confidence interval **[4 marks]**

(iv) Sketch a graph of the estimated survival function. **[3 marks]**

(v) Estimate the probability that a person will be employed in the fifth month. **[2 marks]**

6.2 Paper II

QUESTION ONE (30 marks) (COMPULSORY)

(a) An investigation was undertaken into the length of post-operative stay in hospital after a particular type of surgery. All patients undergoing this surgery between 1^{st} January and 31^{st} January 2013 were observed until either they left the hospital, died, or underwent a second operation. The event of interest was leaving the hospital. Patients who died or underwent a second operation during the period of investigation were treated as censored at the date of death or second operation respectively. The investigation ended on 28 February 2013, and patients who were still in the hospital at that time were treated as censored. Comment on whether censoring in this investigation is likely to be informative. **[2 marks]**

(b) Consider a continuous random variable T with survivor function $S(t)$. Show that the mean of T is given by $\int_0^\infty S(t)dt$. **[3 marks]**

(c) (i) Suppose that T is a continuous, positive random variable with cumulative distribution function $F(t)$ and probability density function $f(t)$. Let $h(t)$ the hazard function. Derive an expression of $f(t)$ in terms of the hazard and the cumulative hazard. **[6 marks]**

(ii) Derive the survivor function for the Weibull distribution with probability density function

$$f(t;\theta,\beta) = \frac{\beta}{\theta^\beta}\, t^{\beta-1} \exp\left(\frac{-t}{\theta}\right)^\beta \quad t > 0 \quad \beta,\ \theta > 0$$

[3 marks]

(iii) Show that if time to failure follows a Weibull distribution, a scatter plot of a suitable function of the survivor function plotted against $\log(time)$ can be used to estimate the parameters θ and β. **[3 marks]**

(d) The survivor function of a random variable is $S(t) = \exp\{-[\exp(\alpha t)^\beta - 1]\}$. Find an approximation to the probability that a subject with this survivor function fails within the interval (1.5,2.1) given survival to point 1.5 . **[4 marks]**

(e) (i) Show that the Nelson-Aalen estimator for the cumulative hazard equal the negative logarithm of the Kaplan-Meier survivor function estimator . **[2 marks]**

(ii) The following data shows the period in complete months from the initial ill health retirement to the end of observation for those members who died or withdraw with a special permission from the observation before the end of the investigation of two years.
12, 5*, 6, 15, 1*, 18*, 20, 6, 3*, 20, 10*, 23, 8*.
A censored observation is denoted by *.
Compute the Nelson-Aalen estimate of the survivor function and plot the results. **[7 marks]**

QUESTION TWO (20 marks)(Optional)

A pharmaceutical company is undertaking trials on a new drug which, it claims, cures a particularly uncomfortable but not life threatening condition. It has conducted extensive testing of the drug on a large group of people suffering from the condition and has noticed that the drug is much more effective in some groups of patients than others. It has fitted a Cox regression for the hazard of symptoms disappearing $h(t)$ with three parameters

$$h(t) = h_0(t)\exp(\beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3)$$

where $\beta_i \quad i = 1, 2, 3$ are parameters and

$$Z_1 = \begin{cases} 1 & \text{if patient is male} \\ 0 & \text{if patient is female} \end{cases}$$

Z_2 represents the age, in years minus 20, of the patient when the drug was administered.

$$Z_3 = \begin{cases} 1 & \text{if the patient attended a gym} \\ 0 & \text{otherwise} \end{cases}$$

The company has discovered the following, where the age given is the age when the drug was administered:

- a 25 year old female who attended a gym had a hazard of symptoms disappearing equal to that of a male of the same age who did not attend a gym;
- a 45 year old male who did not attend a gym had a hazard of symptoms disappearing half that of a 43 year old male who attended a gym; and
- a 32 year old female who attended a gym had a hazard of symptoms disappearing 60% greater than that of a 45 year old female who did not attend a gym.

(a) Calculate the value of each of the parameters **[11 marks]**

(b) Identify the group of people to whom the baseline hazard applies. **[1 mark]**

(c) Determine for which group of people the drug is most effective. **[1 mark]**

(d) The probability that a woman who attended a gym and was aged 38 years when she was given the drug still had symptoms of the condition after 28 days was found to be 0.75. Calculate the probability of still having symptoms after 28 days for a male aged 26 years when given the drug who did not attend a gym. **[7 marks]**

QUESTION THREE (20 marks) (Optional)

(a) List TWO different methods of graduating crude mortality data. **[2 marks]**

(b) State, for each method, TWO advantages and ONE disadvantage. **[3 marks]**

(c) A large pension scheme is examining its most recent experience and has graduated its data over a range of ages using $\mu_x = 0.0005 + 0.00005(1.1^x)$. The table below gives

(i) Perform an overall goodness of fit test on the data. **[9 marks]**

(ii) Discuss three situations which the test in part (i) is not able to detect and each case identify a suitable test for the detection . **[6 marks]**

Age	Exposed to risk	Observed deaths	Graduated rates
60	7,966	127	0.015724
61	7,728	139	0.017246
62	7,870	162	0.018921
63	7,622	167	0.020763
64	7,097	205	0.022790
65	7,208	179	0.025019
66	6,833	185	0.027470
67	6,474	212	0.030167
68	6,208	209	0.033134
69	5,914	195	0.036398

QUESTION FOUR (20 marks) (Optional)

WeedKill is a new treatment for farms which, according to the manufacturer, kills weeds within days. A study was done in a small town in which 14 residents with weedy farms participated. Seven, chosen at random, were given WeedKill in a plain bottle and treated their farms with it one morning. The other seven were given the most popular herbicide already on the market, also in a plain bottle, and treated their farms with it on the same morning. All 14 were asked to assess their farms each morning following treatment until all the weeds had gone.
Unfortunately children in the neighbourhood of the town on days 5 and 8 played in five of the farms in the study and made such a mess of these farms such that they had to be withdrawn. These five farms are also considered to be censored.The study ended after 16 days, at which time any lawn which still had weed was considered to be censored.
The table below shows how many days elapsed before all the weeds disappeared for each of the 14 farms, or until censoring. Censoring is denoted by an *.

WeedKill group:	5,	6,	8,	8*,	8*,	11,	16*
Alternative treatment group:	3,	4,	5,	5*,	5*	5*,	10

(a) Describe THREE types of right censoring present in this study, giving examples of how they occur. **[3 marks]**

(b) Calculate the Kaplan-Meier estimate of the survival functions of still having moss for each of the two groups separately. **[6 marks]**

(c) Sketch the two estimated survival functions on the same graph. **[3 marks]**

(d) Comment on your results. **[2 marks]**

A local actuarial science student suggests that it would be a good idea to use Cox regression to compare the effectiveness of WeedKill with the alternative treatment. She thinks that using a dummy variable, Z, with the value 1 for WeedKill and 0 for the alternative treatment would be suitable.

(e) Determine the equations for the hazard function for the farms treated with Weedkill and those treated with the alternative treatment, defining all the terms you use. **[2 marks]**

(f) Write the contributions to the partial likelihood of the data in this Cox model at times 3,4, 5 and 6. **[4 marks]**

QUESTION FIVE (20 marks) (Optional)

A random variable, T, has the exponential distribution with hazard function

$$h(t) = \lambda \quad t \geq 0 \quad \lambda > 0$$

(a) Show that for $\partial t > 0$, $P(t \leq T \leq t + \partial t | T > t)$ is independent of t. **[5 marks]**

Suppose there d failures in a single sample of failure times possibly subject to censoring and inference on λ is required.

(b) Obtain the maximum likelihood estimator for λ. **[4 marks]**

(c) Explain how the $100(1-\alpha)\%$ confidence interval for the parameter may be constructed. How can this interval be used to test the hypothesis $H_0 : \lambda = \lambda_0$ against $H_1 : \lambda \neq \lambda_0$ where λ_0 is a known constant. **[7 marks]**

(d) You observe the following sample where * denotes right censoring.

1.2, 1.8, 2.0, 2.1*, 2.9, 3.0*, 3.4, 4.0*, 4.1, 14.2.

Suppose that this random sample comes from an exponential distribution. Find the maximum likelihood estimate of the parameter of the exponential distribution. Give a 95% confidence interval for that parameter.
Test the hypotheses $H_0 : \lambda = 0.1$ against $H_1 : \lambda \neq 0.1$ at 5% level of significance.

[4 marks]

6.3 Paper III

QUESTION ONE (30 marks) (COMPULSORY)

(a) Define the survivor function $S(t)$ and the hazard function $h(t)$ for a continuous random variable T measuring lifetime. Write down an expression for the survivor function in terms of the hazard function. **[3 marks]**

(b) The exponential distribution has constant hazard function $h(t) = \lambda$. Write down expressions for survivor function $S(t)$ and the mean of this distribution in terms of λ. **[2 marks]**

(c) (i) Explain what is meant by a *right-censored* observation. **[1 mark]**

(ii) Give two different examples of ways in which a *right-censored* observation might arise. **[2 marks]**

(d) After a radical mastectomy for breast cancer, ten female patients were randomly assigned to one of two groups, an experimental group who received chemotherapy, and a control group who received no drugs. At the end of two years, survival times in months were recorded and are given in the table below. A right-censored observation is denoted by *, so 16* denotes a right-censored observation at 16 months.

Experimental group	23	16*	18*	20*	24*
Control group	15	18	19	19	20

(i) Compute the Kaplan-Meier estimate of the survivor function for each group and plot the results on one graph. **[10 marks]**

(ii) If survival times have an exponential distribution, estimate the 95% confidence interval for the parameter. **[5 marks]**

(e) A pharmaceutical company is interested in testing a new treatment for a debilitating but non-fatal condition in cows. A randomized trial was carried out in which a sample of cows with the condition was assigned to either the new treatment or the previous treatment. The event of interest was the recovery of a cow from the condition. The results were analyzed using a Cox regression model.
The final model estimated the hazard, $h(t, x)$ as:

$$h(t, x) = h_0(t, x) \exp(\beta_0 z + \beta_1 x + \beta_2 xz)$$

where:
z is a covariate taking the value 1 if the cow was assigned the new treatment and 0 if the cow was assigned the previous;
x is a covariate denoting the length of time (in days) for which the cow had been suffering from the condition when treatment was started;
and t is the number of days since treatment started.
The parameter estimates were $\beta_0 = 0.8$, $\beta_1 = 0.4$ and $\beta_2 = -0.1$.

(i) For a particular cow, the new treatment's hazard is half the previous treatment's hazard.
Calculate the number of days for which that cow had the condition before the initiation of treatment. **[3 marks]**

(ii) Under the previous treatment, cows whose treatment began after they had been suffering from the condition for two days had a median recovery time of 10 days once treatment had started.
Calculate the proportion of these cows which would still have had the condition after 10 days if they had been given the new treatment **[4 marks]**

QUESTION TWO (20 marks) (Optional)

(a) A random variable, T, has the Weibull distribution with hazard function

$$h(t) = \lambda\gamma t^{\gamma-1} \quad t \geq 0 \quad \lambda > 0 \quad \gamma > 0$$

(i) Derive the survivor function, $S(t)$, of T. **[2 marks]**

(ii) Show that the parameters are given by the intercept and slope of the theoretical relationship of the logarithm of the cumulative hazard function plotted against the logarithm of time. **[3 marks]**

(b) 20 refrigerator motors of a particular type were each tested on an accelerated life test, and their times till first failure (hours) were recorded. The results are listed below, where * denotes that the motor was still functioning properly when the test was brought to an end.

2 4* 5 5 5 6* 7 7 7* 8 8 9* 11 11 12 12* 12* 12* 16 18*

(i) Use the Nelson-Aalen method with these data to estimate the survivor function, $S(t)$, for the time to first failure of a motor of this type. **[5 marks]**

(ii) Referring to the result from part (a)(ii) and (b)(i), use a suitable graphical method to investigate whether or not these data come from a Weibull distribution. **[6 marks]**

(iii) Draw a straight line through the points on your graph by eye and use it to estimate the parameters, λ and γ, of a Weibull distribution fitted to these data. **[4 marks]**

QUESTION THREE (20 marks) (Optional)

(a) Outline three reasons why the Cox proportional hazards model is widely used in empirical work. **[3 marks]**

(b) An energy provider is worried about the number of its customers who transfer to other companies within the first two years of their contract and is trying to direct its advertising towards the most loyal section of the population.
The company has looked at its records over recent years and has fitted a Cox proportional hazards model to those who have transferred within the first two years using the factors which appear to have the most impact on early transfer rates.
The following figures have been derived from the data:

(i) Give the hazard function for this Cox proportional hazard model defining all the terms and covariates. **[6 marks]**

	Factor	Parameter Estimate	Variance
Gender	Male	-0.25	0.015
	Female	0	0
Volume of energy consumed	High	0.32	0.008
	Low	0	0
Area of Residence	CityCentre	0.19	0.012
	City (not centre)	0	0
	Rural	-0.35	0.005

(ii) State the features of the person to whom the baseline hazard applies. **[1 mark]**

(iii) Calculate symmetric 90% confidence intervals for the parameters based on the standard errors. **[2 marks]**

(iv) Test the suggestion that men change energy providers more frequently than women. **[3 marks]**

(v) There is a 65% probability that a male customer who is a low consumer of energy and lives in a rural area has transferred providers before the end of three years.
Calculate the probability that a male customer who is a high consumer of energy and lives in a city centre remains with the company for at least three years. **[5 marks]**

QUESTION FOUR (20 marks) (Optional)

(a) List two different methods of graduating crude mortality rates. **[2 marks]**

(b) A life insurance company has graduated its own mortality experience for term assurance business over the past 15 years against a standard table using the following equation

$$q_x = 0.94q_x^s - 0.0001$$

where q_x^s is the mortality rate from the standard table.
The following is an extract from the data.

Age x	Exposed to Risk	Deaths	Graduated Rate
40	24,584	14	0.000625
41	32,587	32	0.000683
42	15,784	16	0.000748
43	21,336	22	0.000823
44	25,874	24	0.000908
45	21,544	22	0.001005
46	23,967	25	0.001114
47	25,811	30	0.001239
48	26,911	28	0.001378
49	28,445	38	0.001536
50	30,205	45	0.001713

(i) Carry out a test for overall goodness of fit of the data, using a 95% significance level. **[10 marks]**

(ii) Carry out Cumulative Deviations Test at 95% to check the validity of the graduation. **[8 marks]**

QUESTION FIVE (20 marks)(Optional)

(a) A school offers a one year course in a foreign language as an evening class. This is divided into three terms of 13 weeks each with one lesson per week. At the end of each lesson all the students sit a test and any that pass are awarded a qualification, and no longer attend the course.
Last year 33 students started the course. Of these 13 dropped out before completing the year and considered as censored, and 16 passed the test before the end of the year. The last lesson attended by the students who did not stay for the whole 39 lessons is shown in the table below along with their reason for leaving.

Number of students	Last lesson attended	Reason for leaving
5	1	Dropped out
1	6	Dropped out
2	7	Passed test
2	13	Dropped out
5	14	Passed test
6	27	Passed test
4	28	Dropped out
1	30	Dropped out
3	36	Passed test

(i) Describe two types of censoring present in the investigation and an example of a student to whom each type applies. **[4 marks]**

(ii) Calculate the Kaplan-Meier estimate of the survival function. **[5 marks]**

(iii) Estimate the variance of estimate at the 30^{th} lesson and hence its 95% confidence interval **[4 marks]**

(iv) Determine the probability that a student who starts the course passes by the end of the year. **[1 mark]**

(v) Since only four students had not passed by the end of the year and a total of 16 had passed, the school claims in its publicity that 80% of students are awarded the qualification by the end of the year.Comment on the schools claim in light of your answer to part (iv). **[2 marks]**

(b) A random variable, T, has the exponential distribution with hazard function

$$h(t) = \lambda \quad t \geq 0 \quad \lambda > 0$$

Show that for $\partial t > 0$, $Pr(t \leq T \leq t + \partial t | T > t)$ is independent of t. **[4 marks]**

6.4 Paper IV

QUESTION ONE (30 marks) (COMPULSORY)

(a) (i) Define the following types of censoring in the context of a mortality investigation:

- random censoring;
- right censoring;
- informative censoring.

[3 marks]

(ii) Jane received a bunch of 17 fresh red roses on the evening of her birthday from her boyfriend. She arranged them in a vase and placed them on the table in the garden for all to admire. She needed to do a project for school so decided to use them to conduct an experiment as to how long roses live before they start to wilt. She checked them very often, and noted down the date when any was showing signs of wilting, and immediately removed the wilting rose from the vase. The following shows what she discovered.
Day 2. Very disappointing, already two roses wilting.
Day 3. A neighbour passed with his goat which took a nibble at the bunch, so three damaged, but otherwise fresh, roses had to be removed.
Day 5. One more wilting.
Day 7. Three more wilting.
Day 8. The boy down the road stole a fresh rose to give to his sweetheart.
Day 9. Another one wilting and it is hard to make the remaining ones look good in the vase, so the project is terminated.
For each of the three types of censoring listed in part (i) state and explain which roses (if any) experience that censoring. **[6 marks]**

(b) Consider a continuous non-negative random variable T whose hazard is assumed to take the form:

$$h(t) = \frac{2}{1+t} \quad t > 0$$

Obtain the mean of T. **[5 marks]**

(d) After a radical mastectomy for breast cancer, ten female patients were randomly assigned to one of two groups, an experimental group who received chemotherapy, and a control group who received no drugs. At the end of two years, survival times in months were recorded and are given in the table below. A right-censored observation is denoted by *, so 16* denotes a censored observation at 16 months.

Experimental group	23	16*	18*	20*	24*
Control group	15	18	19	19	20

(i) Compute the Kaplan-Meier estimate of the survivor function for each group and plot the results on one graph. **[6 marks]**

(ii) If survival times have an exponential distribution, estimate the 95% confidence interval for the parameter for each group. **[6 marks]**

(e) (i) Write down the equation of the Cox proportional hazards model in which the hazard function depends on duration t and a vector of covariates z. You should define all the other terms that you use. **[2 marks]**

(ii) Why is the Cox model sometimes described as semi-parametric.

[2 marks]

QUESTION TWO (20 marks)(Optional)

(a) The lifetime of a product can often be modelled by a Weibull distribution. The probability density function of a Weibull distribution is

$$f(t) = \lambda\gamma t^{\gamma-1}\exp[-(\lambda t^{\gamma})] \qquad t > 0;\ \gamma > 0,\ \lambda > 0$$

(i) Derive the hazard function for a Weibull distribution. **[3 marks]**

(ii) In order to show the versatility of the Weibull distribution, sketch the form of the hazard function for each of a suitably chosen set of values for γ .

[4 marks]

(iii) Briefly describe scenarios in which each of your chosen hazard functions in part (ii) could be useful in lifetime analysis. **[3 marks]**

(v) Show that $\log\{-\log(S(t))\}$ is a linear function of $\log(t)$. State the intercept and slope of the line if $\log\{-\log(S(t))\}$ on the vertical axis is plotted against log(t) on the horizontal axis. **[3 marks]**

(b) A quality manager is investigating the lifetimes of springs produced in his factory. He has randomly selected 8 springs and subjected them to a stress level of 800 N/mm^2. The manager wants to know whether the lifetimes of the springs can be modelled by Weibull distributions. He obtains the Kaplan-Meir estimate for the survivor curve function for stress level 800 N/mm^2 as

$$\widehat{S(t)} = \begin{cases} 1 & 0 \le t < 365 \\ 0.875 & 365 \le t < 400 \\ 0.750 & 400 \le t < 462 \\ 0.625 & 462 \le t < 523 \\ 0.500 & 523 \le t < 625 \\ 0.375 & 625 \le t < 1053 \\ 0.250 & 1053 \le t < 1432 \\ 0.125 & 1432 \le t < 2024 \\ 0 & t \ge 1024 \end{cases}$$

By sketching a suitable graph, explain whether you think that the lifetime of the springs at this stress can be modelled with a Weibull distribution.

[7 marks]

QUESTION THREE (20 marks) (Optional)

(a) Describe why a mortality experience would need to be graduated. **[3 marks]**

(b) List three different methods of graduating crude mortality data. **[3 marks]**

(c) An insurance company conducts an investigation into the mortality rates of policyholders who choose to retire at a relatively young age. The following table shows data from the investigation, together with graduated rates $\overset{o}{q}_x$ which were fitted with reference to standard table rates, $\overset{s}{q}_x$ using a link function $\overset{o}{q}_x = \overset{s}{q}_x + constant$.

Age x	Exposed to risk	Deaths	$\overset{o}{q}_x$
55	1,550	15	0.00673
56	2,100	18	0.00689
57	2,300	15	0.00709
58	2,450	21	0.00736
59	2,700	18	0.00770
60	3,250	29	0.00820
61	3,100	25	0.00891
62	3,450	30	0.00978
63	3,600	45	0.01084
64	3,750	41	0.01210

(i) Perform an overall goodness of fit test on the data using a chi-square test. **[11 marks]**

(ii) Discuss three situations which the test in part (i) is not able to detect and each case identify a suitable test for the detection . **[3 marks]**

QUESTION FOUR (20 marks) (Optional)

An energy provider is worried about the number of its customers who transfer to other companies within the first two years of their contract and is trying to direct its advertising towards the most loyal section of the population. The company has looked at its records over recent years and has fitted a Cox proportional hazards model to those who have transferred within the first two yearsusing the factors which appear to have the most impact on early transfer rates. The following figures have been derived from the data:

	Factor	Parameter Estimate	Variance
Gender	Male	-0.25	0.015
	Female	0	0
Volume of energy consumed	High	0.32	0.008
	Low	0	0
Area of Residence	City Centre	0.19	0.012
	City (not centre)	0	0
	Rural	-0.35	0.005

(i) Give the hazard function for this Cox proportional hazard model defining all the terms and covariates. **[6 marks]**

(ii) State the features of the person to whom the baseline hazard applies. **[1 mark]**

(iii) Calculate symmetric 95% confidence intervals for the parameters based on the standard errors. **[4 marks]**

(iv) Comment on the effect of each covariate on the hazard of transfer to other companies. . **[4 marks]**

(v) There is a 70% probability that a male customer who is a low consumer of energy and lives in a rural area has transferred providers before the end of two years. Calculate the probability that a male customer who is a high consumer of energy and lives in a city centre remains with the company for at least two years. **[5 marks]**

QUESTION FIVE (20 marks) (Optional)

An investigation was undertaken into the time spent waiting in check-out queues at a supermarket. A random sample of customers was surveyed, and the times at which they joined the check-out queue and completed their purchases were recorded. If they left the check-out queue without completing a purchase, the time at which they left was also recorded. Below are the data for 12 customers.

(i) Calculate the Nelson-Aalen estimate of the cumulative hazard of the duration between joining the queue and completing a purchase and plot it. **[8 marks]**

(ii) Use the estimate in part(i) to obtain the survivor function. **[2 marks]**

(iii) Obtain the 95% confidence interval of $\widehat{H(7)}$. **[5 marks]**

Customer number	Time joined	Time purchase completed	Time left without making purchase
1	10.00 a.m.	10.08 a.m.	
2	10.07 a.m.	10.09 a.m.	
3	10.10 a.m.	10.16 a.m.	
4	10.25 a.m.	10.31 a.m.	
5	10.30 a.m.	10.32 a.m.	
6	10.45 a.m.	10.49 a.m.	
7	11.10 a.m.		11.20 a.m.
8	11.15 a.m.	11.21 a.m.	
9	11.35 a.m.		11.40 a.m.
10	11.58 a.m.	12.09 p.m.	
11	12.10 p.m.	12.14 p.m.	
12	12.15 p.m		12.22 p.m.

(iv) The supermarket decides to introduce a scheme under which any customer who has to wait at a check-out for more than 10 minutes receives a 20sh refund on the cost of their shopping. The supermarket has 20,000 customers per day.

(a) Give an estimate of the daily cost of the new scheme. **[2 marks]**

(b) Comment on the assumptions that you have made in obtaining the estimate in (a).

[3 marks]

6.5 Paper V

QUESTION ONE (30 marks) (COMPULSORY)

(a) In the context of a survival model:-

(i) Define Type I and Type II Censoring. **[2 marks]**

(ii) Give an example where the censoring is informative. **[1 mark]**

(iii) Give an example for interval censoring. **[1 mark]**

(iv) Suppose that students enrolling for actuarial exams complete all the exams and qualify as actuaries on an average duration of three years. The enrolment for student membership starts exactly on 1^{st} January of every year. A group of 100 students joined the Institute on 01/01/2017. The institute wanted to study the completion rate of unmarried male students. They started the study on 01/01/2017 and completed the study on 31/12/2020. Describe the types of censoring present in the data collected for the above study.

[3 marks]

(b) A college student decided to sell sausages at *Bomas* to earn money. On the first day, he buys 50 sausages from a local vendor and started selling at morning 8 am. He wanted to sell at least 30 sausages so that he can recoup the expenses. He recorded the events of the day as follows.
At **8.20 am**, five sausages are bought by a family.
At **8.40 am** , seven sausages are bought by a group of college students.
At **9.00 am**, the student found that five sausages are not in good condition and he throws them out
At **9.20 am**, ten sausages are bought by a group of workers.
At **9.40 am**, one student bought three sausages but forgot to pay the money.
At **10 am**, seven sausages are bought by a passenger.
At **10.20 am**, one passenger bought eight sausages.
At **10.30 am** , the student left *Bomas* with the remaining sausages.

(i) Estimate the time taken by the student to sell at least 30 sausages using Nelson-Aalen estimator. **[8 marks]**

(ii) Comment whether the above estimate would be a good basis for the student to predict his future sales. **[3 marks]**

(c) An investigation took place into the effect of a new treatment on the survival of cancer patients. Two groups of patients were identified. One group was given the new treatment and the other an existing treatment. The following model was considered:

$$h(t, z) = h_0(t) \exp(\beta_1 Z_1 + \beta_2 Z_2)$$

where:
t is the time since the start of treatment, $h(t, z)$ is the hazard of death and $h_0(t)$ is the baseline hazard at time t.
Z_1= gender (a categorical variable with 0 = female, 1 = male)
Z_2= treatment (a categorical variable with 0 = existing treatment, 1 = new

treatment)
β_1 and β_2 are the parameters of the covariates.
The results of the investigation showed that, if the model is correct the risk of death for a male patient is 5% more than that of a female patient while that for a patient given the existing treatment is 10% more than that for a patient given the new treatment

(i) State the group of patients to which the baseline hazard applies? **[1 mark]**

(ii) Show that the risk of death for a male patient who has been given the new treatment is less than that for a female patient given the existing treatment. **[6 marks]**

(iii) Determine the probability that a male patient will die within 3 years of receiving the new treatment, if the median survival time for female patient whom existing treatment given is 3 years. **[5 marks]**

QUESTION TWO (20 marks)(Optional)

(a) A Government actuary is charged with producing a national life table for general use, based on data from death registrations and censuses.

(i) Explain the role of graduation in the production of this table, and suggest a suitable procedure to use for graduating the crude rates. **[4 marks]**

(ii) A life office is analysing the graduated rates derived from its recent experience. These graduated rates must be consistent with a published standard table. What is meant by consistency here? **[1 mark]**

(iii) What is the purpose of serial correlation test and how it works? **[2 marks]**

(b) The following table gives an extract of data from a mortality investigation conducted by a life insurance company. The raw data have been graduated by reference to a standard table of assured lives.

Age	Expected Deaths	Actual Deaths
60	38.3	36
61	40.45	34
62	42.52	38
63	44.56	40
64	47.63	52
65	51.25	48
66	55.35	57
67	60.45	65
68	63.22	69
69	56.78	74
70	71.56	77

Test whether the graduated rates are reflecting the standard table of assured lives using serial correlation test. **[13 marks]**

QUESTION THREE (20 marks) (Optional)

(a) Information regarding the survival times (in weeks) of 20 patients suffering from leukaemia is provided below:
6, 6, 6, 6*, 7, 9*, 10, 10*, 11*, 13, 16, 17*, 19*, 20*, 22, 23, 25*, 32*, 32*, 34
where * represents the censored time.

(i) Calculate and sketch the Kaplan-Meier estimate of the survival function assuming censoring occurs just after death at the respective duration. **[8 marks]**

(ii) Obtain an approximate 95% confidence interval for the probability that a patient survives for at least 8 weeks after the start of the drug treatment. **[4 marks]**

(b) An investigation is undertaken into the mortality of men aged between exact ages 50 and 55 years. A sample of n men is followed from their 50th birthdays until either they die or they reach their 55th birthdays. The hazard of death (or force of mortality) between these ages, $h(t)$, is assumed to have the following form:

$$h(t) = a + bt$$

where a and b are parameters to be estimated and t is measured in years since the 50th birthday.

(i) Derive an expression for the survival function between ages 50 and 55 years. **[3 marks]**

(ii) Sketch this on a graph if both parameters are positive. **[2 marks]**

(iii) Comment on the appropriateness of the assumed form of the hazard for modelling mortality over this age range. **[3 marks]**

QUESTION FOUR (20 marks) (Optional)

(a) Give three reasons why the Cox PH model is mostly preferred in the analysis of survival data. **[3 marks]**

(b) The table below gives the data for a small sample of heart patients in a hospital. It shows the time in months until death. Observations marked * show that the patient either left the hospital or died due to a cause not relatedto heart condition.

Males	5*	10	12*	14	15*	18*	19	
Females	1*	3	6	7*	9*	11*	16	20*

A Cox proportional hazard model $h(t, z) = h_0(t)\exp(\beta Z)$ is to be fitted to these data where t is time till death $h_0(t)$ is the baseline hazard and $Z = 0$ for males , $Z = 1$ for females

(i) Write down the general expression for the partial likelihood for such investigation. **[1 mark]**

(ii) Derive an expression for the partial likelihood for the above data. **[7 marks]**

(iii) Calculate the maximum partial likelihood estimate of β. **[5 marks]**

Males	19
Females	8*

(iv) The following additional data was generated.
Write down the partial likelihood after including the additional data provided.

[4 marks]

QUESTION FIVE (20 marks) (Optional)

(a) A religious organisation maintains two lists of members:-
The Sick List, so that members may pray for them .
The Dead List, a list of recently deceased members.
Each list is published in a bulletin given to those attending the regular weekly meetings of religious worship. The lists are updated each week half way between the religious worship meetings. A study was made of the mortality of sick members. A sample of members joining the Sick List in the first quarter of 2016 was followed until they left the list. Those who left the list but who did not move to the Dead List were assumed to have recovered. The study terminated on 31 March 2017. Below are given some data from the study. "Week first appeared on Sick List" and "Week last appeared on Sick List" are measured in weeks from the first week of 2016.

Member	Week first appeared on Sick List	Week last appeared on Sick List	Outcome
1	1	1	Assumed recovered
2	1	3	Moved to Dead List
3	3	4	Moved to Dead List
4	3	65	Still on Sick List 31 March 2017
5	6	17	Moved to Dead List
6	7	14	Assumed recovered
7	9	11	Assumed recovered
8	10	60	Moved to Dead List
9	11	11	Moved to Dead List
10	12	65	Still on Sick List 31 March 2017

(i) Calculate the Nelson-Aalen estimate of the cumulative hazard of the mortality of sick members and plot it. **[8 marks]**

(ii) Use the estimate in part(i) to obtain the survivor function and hence Nelson-Aalen estimate of $S(52)$. **[3 marks]**

(iii) Obtain the 95% confidence interval of $\widehat{H(52)}$. **[3 marks]**

(b) An energy provider is worried about the number of its customers who transfer to other companies within the first two years of their contract and is trying to direct its advertising towards the most loyal section of the population.
The company has looked at its records over recent years and has fitted a Cox proportional hazards model to those who have transferred within the first two years using the factors which appear to have the most impact on early transfer rates.
The following figures have been derived from the data:

	Factor	Parameter Estimate	Variance
Gender	Male	-0.25	0.015
	Female	0	0
Volume of energy consumed	High	0.32	0.008
	Low	0	0
Area of Residence	CityCentre	0.19	0.012
	City (not centre)	0	0
	Rural	-0.35	0.005

(i) Calculate symmetric 95% confidence intervals for the parameters based on the standard errors. **[4 marks]**

(ii) Test the suggestion that men change energy providers more frequently than women. **[2 marks]**

6.6 Paper VI

QUESTION ONE (30 marks) (COMPULSORY)

(a) A new weedkiller is designed to kill weeds growing in grass. To test its effectiveness it was administered via a single application to 20 test areas of grass. Within hours of applying the weedkiller, the leaves of all the weeds went black and died, but after a time some of the weeds re-grew as the weedkiller did not always kill the roots.
The test lasted for 12 months, but after six months five of the test areas were accidentally ploughed up and so the trial on these areas had to be discontinued. None of these five areas had shown any weed re-growth at the time they were ploughed up.
Ten of the remaining 15 areas experienced a re-growth of weeds at the following
durations (in months): 1, 2, 2, 2, 5, 5, 8, 8, 8, 8.
Five areas still had no weed re-growth when the trial ended after 12 months.

(i) Describe, giving reasons, the types of censoring present in the data. **[4 marks]**

(ii) Giving the reason state whether the censoring is informative or not? **[2 marks]**

(iii) Estimate the probability that there is no re-growth of weeds nine months after application of the weedkiller using Nelson- Aalen estimator. **[7 marks]**

(c) (i) Let $t_1, t_2 \ldots, t_n$ be n non-negative observations of a random variable with p.d.f. $f(t;\theta)$ and a hazard function $h(t;\theta)$. Some of the observations are censored values. Obtain the loglikelihood of the observations in terms of the hazard. **[4 marks]**

(ii) The following are the number of days taken by a group of people suffering from malaria to recover after treatment with some drug.

2 4* 5 5 5 6* 7 7 7* 8 8 9* 11 11 12 12* 12* 12* 16 18*

where * denotes a censored observation. Assuming this data is from a Weibull distribution with index 1, find the maximum likelihood estimate of the scale parameter. **[4 marks]**

(d) A researcher is investigating the contributing factors to the speed at which patients recover from a common minor surgical procedure undertaken in hospitals across the country. He has the questionnaires which each patient completed before the surgery and the length of time the patient remained in hospital after surgery and is attempting to fit a Cox proportional hazards model to the data. He has fitted a model with what he assumes are the most common contributing factors and has calculated the parameters as shown in the table below:

Covariate	Category	Parameter
Gender	Male	0
	Female	0.065
Smoker	Non Smoker	-0.035
	Smoker	0
Drinker	Non Drinker	-0.06
	Moderate Drinker	0
	Heavy Drinker	0.085

(i) Give the hazard function for this Cox proportional hazards model, defining all the terms and covariates. **[6 marks]**

(ii) A male moderate drinker who does not smoke has a hazard of leaving hospital after three days of 0.6. Calculate the probability that a female heavy drinker who smokes and who is still in hospital after three days is **NOT** discharged at that point. **[3 marks]**

QUESTION TWO (20 marks)(Optional)

A life insurance company is investigating the mortality of its policyholders over the past year. It wishes to compare the current mortality rates with those obtained from a similar investigation ten years ago. The following is an extract of the data:

Current investigation			Previous investigation
Age x last birthday	Exposed to risk	Observed deaths	Mortality rate
50	5,368	25	0.00479
51	4,986	26	0.00538
52	4,832	30	0.00603
53	5,298	37	0.00675
54	5,741	45	0.00756
55	4,866	46	0.00844
56	4,901	52	0.00942
57	5,003	63	0.01050
58	3,952	45	0.01169
59	2,786	45	0.01299

(i) Sketch a graph, showing clearly both the current and the previous mortality rates. **[6 marks]**

(ii) Carry out a goodness-of-fit test on the data. **[7 marks]**

(iii) Carry out the signs test. **[4 marks]**

(iv) Comment on your answers to part (iii) in the light of your sketch in part (i). **[3 marks]**

QUESTION THREE (20 marks) (Optional)

(a) The lifetime of a product can often be modelled by a Weibull distribution. The hazard function of a Weibull distribution is

$$h(t) = \lambda\gamma t^{\gamma-1}$$

.

(i) Derive the probability density function for a Weibull distribution. **[3 marks]**

(ii) In order to show the versatility of the Weibull distribution, sketch the form of the hazard function for each of a suitably chosen set of values for γ. **[4 marks]**

(iii) Briefly describe scenarios in which each of your chosen hazard functions in part (ii) could be useful in lifetime analysis. **[3 marks]**

(b) A quality manager is investigating the lifetimes of springs produced in his factory. He has randomly selected 24 springs and subjected them to 3 different stress levels, each stress level being applied to 8 of the springs. The table below shows the results of his experiment. The results are the numbers of cycles to failure (in units of 1000 cycles). * denotes data that are right-censored

Stress (N/mm^2)	Failure times
800	625, 1053, 1432, 2024, 523, 400, 462, 365
750	3400, 9413, 1806, 4327, 11524*, 7154, 2969, 3014
700	12513, 12507*, 3028, 12508*, 6253, 11607, 12475*, 7798*

(i) Derive the survival function for stress level $800N/mm^2$. **[3 marks]**

(ii) The manager wants to know whether the lifetimes of the springs can be modelled by Weibull distributions. Use your result in part (i) to derive a function which can be used to investigate this request. Plot a suitable graph and then explain whether you think that the lifetime of the springs at stress level $800N/mm^2$ can be modelled with a Weibull distribution. **[6 marks]**

QUESTION FOUR (20 marks) (Optional)

(a) Give three reasons why the Cox PH model is mostly preferred in the analysis of survival data. **[3 marks]**

(b) An exercise company is developing a computer program to investigate the effect of certain factors on the incidence of a common medical condition which affects millions of people in early middle age. It has identified three factors which appear to have a large impact on the onset of the disease and has set up a Cox regression model for the hazard as follows:

$$h(t) = h_0(t)\exp(\beta_1 A + \beta_2 E + \beta_3 D)$$

where:

A is the age of the individual minus 40 years.

E is an exercise indicator and takes the value of 1 if the person exercises, which in this case means they follow a set regime for 30 minutes each day, and 0 otherwise.

D is a diet indicator and takes the value of 1 if the person diets, which in this case means they consume fewer than 2,000 calories per day, and 0 otherwise.

β_i $i = 1, 2, 3.$ are the parameters to be estimated.

From the data the company has managed to acquire, it has established that:

a 53 year old who exercises but does not diet has a hazard of contracting the condition half that of a 48 year old who does not exercise but diets.

a 55 year old who does not exercise but diets has a hazard of contracting the condition 1.5 times that of a 55 year old who neither exercises nor diets.

a 58 year old who diets but does not exercise has a hazard of contracting the condition double that of a 43 year old who neither diets nor exercises.

(i) Calculate the values of the estimated parameters. **[8 mark]**

(ii) Write down the model and give the baseline the group. **[2 mark]**

(iii) Explain what the values you have calculated in part (i) say about the relative impact of age, diet and exercise on contracting this affliction. **[3 marks]**

(iv) The company has created an advertisement based on the above findings, but the Advertising Regulator has contacted them on the grounds that their model was not sufficiently complex to take into account all the relevant factors. They have suggested four additional factors which might materially impact the hazard of contracting the condition.
Explain how the model could extend to see if any one of the suggested additional factors materially impacts the hazard of contracting the condition. **[4 marks]**

QUESTION FIVE (20 marks) (Optional)

(a) Zebras in a large wildlife conservancy have recently become susceptible to a particular disease called Zebra rabies. Zebras often die as a direct result of contracting this disease. An investigation to monitor deaths due to this disease was carried out between 1 January 2018 and 1 January 2019.
A researcher was interested in the rate at which zebras die once contracting this disease and decided to monitor the health of each zebra on the first day of each month. All zebras in the wildlife park were tagged to ensure that they were identifiable.
14 zebras were diagnosed with rabies during 2018. The data recorded on these 14 zebras are set out below:
Two zebras (reference tags 9 and 21) escaped from the wildlife park on 1 December 2018 having been diagnosed with rabies on 1 July 2018 and 1 November 2018, respectively.
In addition, the following two zebras that contracted the disease were still alive at the end of the investigation:

(i) Determine the Kaplan–Meier estimate of the survival function, where the decrement of interest is death due to rabies. **[8 marks]**

(iii) Plot this survival function. **[3 marks]**

(iv) Obtain the 95% confidence interval for the survival function at time 5. **[5 marks]**

Reference tag	Date of diagnosis	Date of death	Reason for death
1	1 Jan 2018	1 Jun 2018	Rabies
3	1 Jan 2018	1 Dec 2018	Rabies
4	1 Apr 2018	1 Jul 2018	Killed by lion
7	1 Apr 2018	1 Jun 2018	Rabies
8	1 Apr 2018	1 Dec 2018	Rabies
10	1 Jul 2018	1 Sep 2018	Rabies
11	1 Aug 2018	1 Oct 2018	Rabies
12	1 Aug 2018	1 Nov 2018	Rabies
19	1 Sep 2018	1 Oct 2018	Rabies
20	1 Oct 2018	1 Nov 2018	Rabies

Reference tag	Date of diagnosis
13	1 Aug 2018
25	1 Dec 2018

(b) Consider a continuous non-negative random variable T whose hazard is assumed to take the form:

$$h(t) = \frac{2}{1+t} \quad t > 0$$

Obtain the mean of T.

[4 marks]

6.7 Paper VII

QUESTION ONE (30 marks) (COMPULSORY)

(a) A study was undertaken into survival rates following major heart surgery. Patients who underwent this surgery were monitored from the date of surgery until they either died, or they left the hospital where the surgery was carried out, or a period of 30 days had elapsed.

(i) State, with reasons, two forms of right censoring that are present in this study and one form of right censoring that is not present. **[3 marks]**

The analyst collating the results calculated the Nelson–Aalen estimate of the survival function, $S(t)$, as follows:

t(days)	$S(t)$
$0 \leq t < 5$	1.00
$5 \leq t < 17$	0.90
$17 \leq t < 25$	0.86
$25 \leq t$	0.72

(ii) State, using the Nelson–Aalen estimate, the probability of survival for 20 days after the surgery. **[1 mark]**

The analyst also wishes to calculate the Kaplan–Meier estimate of the survival function.

(iii) Determine the Kaplan–Meier estimate of the survival function. **[6 marks]**

(b) Suppose the following proportional hazards regression model is fitted to the mortality data for a sample of life-assurance policyholders.

$$h_i(t) = h_0(t) \exp\{0.01(x_i - 30) + 0.2y_i - 0.05z_i\}$$

where:
$h_i(t)$ denotes the hazard function for life i at duration t ;
$h_0(t)$ denotes the baseline hazard function at duration t ;
x_i denotes the age at entry of life i ;
$y_i = 1$ if life i is a smoker, otherwise zero; and
$z_i = 1$ if life i is female, zero if male.

(i) Describe the class of lives to which the baseline hazard function applies. **[1 mark]**

(ii) What does the model (if correct) tell you about the survival function of a male smoker aged 30 at entry, relative to that of a female smoker aged 40 at entry? **[3 marks]**

(c) A clinician studying survival times of patients with colorectal cancer wishes to use a survival model with the hazard rate $h(t) = \dfrac{\alpha t}{t+1}$, where $\alpha > 0$ is an unknown parameter. This model is fitted to the data comprising n observed survival times $t_1, \cdots, t_n$, some of which may be right-censored.

(i) Obtain the survival function $S(t)$ and the probability density $f(t)$ in this model. **[4 marks]**

(ii) Suggest a suitable coordinate transformation $y = F(S(t))$, $x = G(t)$, which can be used for a graphical assessment of suitability of this model for a given survival data set. How could one obtain a crude graphical estimate of α?

[3 marks]

(iii) Write down the loglikelihood function $\ell(\alpha)$, defining any notation that you use, and derive the maximum likelihood estimator $\hat{\alpha}$ of the parameter α, explaining clearly why this is a maximum. **[6 marks]**

(iv) Explain how the hypotheses $H_0 : \alpha = \alpha_0$, where α_0 is a known constant may be tested against $H_1 : \alpha \neq \alpha_0$. **[4 marks]**

QUESTION TWO (20 marks)(Optional)

(a) State and explain any two features which are desirable when a graduation is performed. **[4 marks]**

(b) The actuary to a large pension scheme has attempted to graduate the schemes recent mortality experience with reference to a table used for similar sized schemes in a different industry. He has calculated the standardized deviations between the crude and the graduated rates, z_x, at each age and has sent you a printout of the figures over a small range of ages. Unfortunately the dot matrix printer on which he printed the results was very old and the dots which would form the minus sign in front of numbers no longer function, so you cannot tell which of the standardized deviations is positive and which negative. Below are the data which you have.

Age	60	61	62	63	64	65	66	67	68	69	70
z_x	2.40	0.08	0.80	0.76	1.04	0.77	1.30	1.76	0.28	0.68	0.93

(i) Carry out an overall goodness-of-fit test on the data. Comment on your result.

[5 marks]

(ii) List four defects of a graduation which the test you have carried out would fail to detect. For each of the defects, suggest a test which could be used to detect it.

[8 marks]

(iii) Carry out one of the tests suggested in part(ii). **[3 marks]**

QUESTION THREE (20 marks) (Optional)

(a) The lifetime of a product can often be modelled by a Weibull distribution. The hazard function of a Weibull distribution is

$$h(t) = \lambda\gamma t^{\gamma-1}$$

(i) Derive the probability density function for a Weibull distribution. **[3 marks]**

(ii) In order to show the versatility of the Weibull distribution, sketch the form of the hazard function for each of a suitably chosen set of values for γ. **[4 marks]**

(iii) Briefly describe scenarios in which each of your chosen hazard functions in part (ii) could be useful in lifetime analysis. **[3 marks]**

(b) In a clinical study of acute myeloma leukemia, patients were classified into two groups according to the presence (Group 1) or absence (Group 2) of a certain morphologic characteristic of white cells termed AG. The following death times t_i(in months) were recorded in group 1, where $\delta_i = 0$ if t is censored and $\delta_i = 1$ otherwise.

Group 1 (AG-positive)

t_i	5	7	8	8	8	10	10	12	13	14	14	14	15
δ_i	1	1	1	0	0	0	0	1	1	0	0	0	0

Values of the Kaplan–Meier estimate $\hat{S}(t)$ for the the group are given below:

Group 1 (AG-positive)

t	5	7	8	12	13
$\hat{S}_1(t)$	0.923	0.846	0.769	0.641	0.513

Apply a suitable graphical method to assess the validity of Weibull's model for Groups 1 with a shape parameter γ and scale parameter λ_1. Using fitted line, obtain an estimate for λ_1. **[10 marks]**

QUESTION FOUR (20 marks) (Optional)

(a) Give four reasons why the Cox PH model is mostly preferred in the analysis of survival data. **[2 marks]**

(b) The table gives the data for a small sample of employees in a factory. It shows the time in months until the first absence from work. Observations marked * show the time of leaving for those employees who left employment without being absent from work.

Male employees	6*	11	13*	15	16*	19*	20	
Female employees	2*	4	7	8*	10*	12*	17	21*

A Cox Proportional Hazards Model

$$h(t; z) = h_0(t)\exp(\beta z)$$

is to be fitted to these data where t is the time until the first absence from work, $h_0(t)$ is the baseline hazard and $z = 0$ for males, $= 1$ for females.

(i) Show that the partial log-likelihood for these data can be written

$$\ell(\beta) = 3\beta - 4\log(1 + \exp(\beta)) - 2\log(2 + \exp(\beta)) + c$$

where c is a constant that does not depend on β. **[8 marks]**

(ii) Calculate the maximum partial likelihood estimate of β. **[5 marks]**

(iii) Calculate the asymptotic standard error of this estimate. **[3 marks]**

(iv) Test the hypothesis that female employees experience higher first absence rates than male employees. Explain the steps in your argument and state your conclusions. **[2 marks]**

QUESTION FIVE (20 marks) (Optional)

(a) In an experiment, individuals had to solve a difficult task. Some of the individuals did not want to complete the task and stopped early. Others did not complete the task by the end of the experiment. The experiment lasted for two hours, and the times(in minutes) until finishing were recorded:

50, 51, 66*, 82, 92, 120*, 120*, 120*,

where the * indicates that the individual did not complete the task. The interest is to model the time that an individual takes to complete the task.

(i) Calculate the Kaplan-Meier estimate $\hat{S}(t)$ of the survival function $S(t)$ and sketch its plot. **[9 marks]**

(ii) Compute the symmetric 95% confidence interval for $\hat{S}(t)$ for $t = 100$ minutes. **[4 marks]**

(b) Consider a continuous random variable T with survivor function $S(t)$. Show that the mean of T is given by

$$\int_0^\infty S(t)dt$$

Hence obtain the mean of a random variable whose hazard is assumed to take the form:

$$h(t) = \lambda$$

[7 marks]