Kasarani Campus
Off Thika Road
P. O. Box 49274, 00101
NAIROBI
Westlands Campus
Pamstech House
Woodvale Grove
Tel. 4442212
Fax: 4444175

KIRIRI WOMENS' UNIVERSITY OF SCIENCE AND TECHNOLOGY
UNIVERSITY EXAMINATIONS, 2024/2025 ACADEMIC YEAR
FIRST YEAR, SECOND SEMESTER EXAMINATION
FOR MASTER OF SCIENCE IN APPLIED STATISTICS AND DATA ANALYTICS
KMA 5113: SURVIVAL ANALYSIS

DATE: 28TH JANUARY, 2025
TIME: 9:00 AM – 12:00 PM

INSTRUCTIONS TO CANDIDATES
ANSWER QUESTION ONE (COMPULSORY) AND ANY OTHER THREE QUESTIONS

QUESTION ONE: COMPULSORY (40 MARKS)

a) Differentiate between Left and right censoring and state any two reasons why an observation might be censored. **[4 Marks]**

b) Consider a distribution with a hazard function $h(t) = \beta$ (constant) for $t > 0$. Find

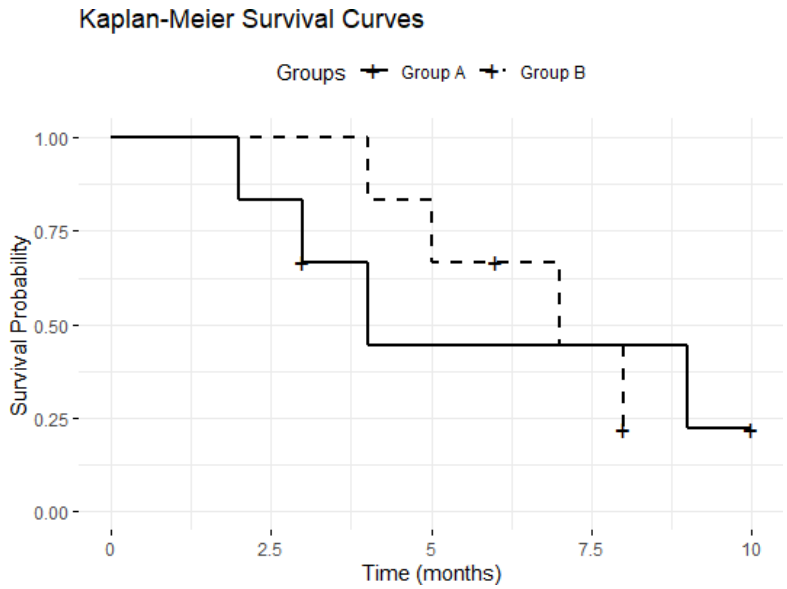
i) The survival function $s(t)$ **[2 Marks]**

ii) The probability density function $f(t)$ **[2 Marks]**

iii) Construct the likelihood and the log likelihood function for a random sample given by
$x_1, x_2, ..., x_d, x_{d+1}, x_{d+2}, ..., x_n$ where $x_1, x_2, ..., x_d$ are uncensored observations and $x_{d+1}, x_{d+2}, ..., x_n$ are censored observations. **[6 Marks]**

iv) Show that $\hat{\beta} = \dfrac{d}{\sum_{i=1}^{n} x_i}$ and $\text{var}(\hat{\beta}) = \dfrac{d}{\left(\sum_{i=1}^{n} x_i\right)^2}$. Then a sample of 12 patients suffering from a minor injury was taken from a hospital. The hazard function of the population is given by $h(t) = \beta$ (constant) and t is the time to recovery in months. The sample times were: 0.06, 0.07,0.10, 0.15, 0.17, 0.18, 0.27, 0.33, 0.45, 0.49, 0.67,1.08. **[6 Marks]**

c) In a clinical trial, 100 patients with hypertension were randomly assigned to two treatment groups: Group A received the new treatment, while Group B received the standard treatment. The trial tracked the time until patients either experienced a significant health event (e.g., stroke, heart attack) or were censored due to dropping out or reaching the end of the study. The survival data was analyzed using Kaplan-Meier estimators, and the survival curves for the two groups are shown below.



i) Compare the survival curves of Group A and Group B. How do the curves differ in terms of steepness and shape? What does this suggest about the effectiveness of the new treatment versus the standard treatment? **[3 Marks]**

ii) Do the survival curves cross at any point? If so, what does this crossing indicate about the survival experience of the two groups at that specific time? **[2 Marks]**

iii) At what time point does the survival curve for Group B show a significant drop compared to Group A? What does this indicate about the relative event rates between the two groups at that time. **[3 Marks]**

Note: Group A: Dotted survival curve and Group B is the continuous survival curve.

d) A Cox proportional hazards model was used to model the survival times of cancer patients. Tumor size (in mm) was included as a covariate, with coefficient β. The maximum likelihood estimate of β was $\hat{\beta} = 0.0176$ with standard error 0.004.

i) Find an estimate of the hazard ratio between two individuals with tumors measuring 46 mm and 37 mm who are identical in other ways. **[2 Marks]**

ii) Construct a 95%CI for the hazard ratio. **[3 Marks]**

e) A researcher fits a Cox proportional hazards model to study the effect of smoking and age on survival time. The following results were obtained from the global test for proportional hazards assumption using Schoenfeld residuals:

Covariate	***p*-value**
Smoking	0.152
Age	0.004
Global	0.010

i) Based on the table, evaluate whether the proportional hazards assumption holds for each covariate and for the model as a whole. **[3 Marks]**

ii) If the assumption is violated for" Age," suggest one method to address this issue. **[1 Mark]**

g) Define an Accelerated Failure Time (AFT) model in the context of survival analysis and list three commonly used distributions for the error term in AFT models. **[3 Marks]**

QUESTION TWO: (20 MARKS)

a) Discuss the role of censored data in the Kaplan-Meier survival analysis and how it might affect the interpretation of the survival curves. **[5 Marks]**

b) A study was conducted to estimate the survival probabilities of cancer patients based on their time to death (in years). The following data were collected for 8 patients:

t_i: Time to death (in years).

$\delta_i = 1$ if the individual i died at time t_i and $\delta_i = 0$ if individual i was censored at that time for $i = 1, 2, \ldots, 8$.

Patient No.	1	2	3	4	5	6	7	8
Time to death t_i	2	5	8	11	12	15	20	23
Indicator δ_i	1	1	1	0	0	1	1	0

i) Estimate the Kaplan-Meier survival function $S(t)$ and the variance estimates of the survival probabilities using the formulas **[12 Marks]**

$$\hat{S}(t) = \prod_{i:t_i \le t} \left(1 - \frac{d_i}{r_i}\right) \qquad Var(\hat{S}(t)) = \hat{S}(t)^2 \sum_{i:t_i \le t} \frac{d_i}{r_i(r_i - d_i)}$$

ii) Draw the Kaplan-Meier survival curve and describe its key features. **[3 Marks]**

QUESTION THREE: (20 MARKS)

a) Explain the primary purpose of the log-rank test in survival analysis and list three key assumptions of the log-rank test**.** **[5 Marks]**

b) Two groups of patients were given different treatments for a virulent form of malaria. The days to recovery (as checked by a daily blood test) were as follows.

Treatment (X)	2	3	4	6	9
Treatment (Y)	5	7	10	12*	14*

*Indicates that the doctor lost touch as in these cases the patient had still not recovered, but had not reported for further testing after the given number of days. Use the log rank test at the 5% level of significance to determine whether there is a significant difference between the number of days taken for patients to recover from each of the two treatments. The tabular value of the chi - square statistic at 5% is 3.84. **[15 Marks]**

QUESTION FOUR: (20 MARKS)

a) Describe the Cox regression model (provide the model equation and define all terms). List assumptions that must hold in order for a Cox PH model to be valid? **[6 Marks]**

b) It is often of interest to compare the relative risk for two people who have taken out on insurance policy, where the volume of the explanatory variables such as age at entry, gender or whether a smoker, or heavy drinker or HIV/AIDS status may be considered. The relative risk is measured by the ratio of the values of the hazard functions of the two concerned individuals. The proportional hazard model for the i^{th} person insured during time t is given by

$$h_i(t) = h_0(t)\exp[0.01(x_i - 30) + y_i - 0.05z_i]$$

Where $h_0(t)$ denotes the baseline hazard duration t, x_i denotes age of entry of the i^{th} person insured. $y_i = 1$ it the, i^{th} person insured is a smoker, otherwise 0 if nonsmoker. $Z_i = 1$ if the i^{th} person insured is a female, otherwise zero it male)

i) Describe the person to whom the baseline hazard function applies. **[2 Marks]**

ii) Determine the relative risk indicated by the model comparing a male smoker aged

30 at entry with female smoker. **[6 Marks]**

c) Suppose we have survival data on six individuals as follows where * indicates a right censored time.

Individual	1	2	3	4	5	6
Survival time	25	12*	19	28	21*	35*

Also suppose we model these data using a Cox proportional hazards model, so that the hazard function for individual i_i is $h_i(t) = \phi_i h_0(t)$. Write down the Cox partial likelihood in terms of the parameters ϕ_i and simplify the expression. **[6 Marks]**

QUESTION FIVE: (20 MARKS)

a) A clinical study records the following survival data:

Time (Months)	Event (1 = Death, 0 = Censored)	Number at Risk
2	1	10
4	1	8
6	0	7
8	1	5

i) Calculate the Nelson-Aalen estimator for the cumulative hazard at each event time. **[3 Marks]**

ii) Interpret the cumulative hazard at time $t - 8$. **[3 Marks]**

b) A researcher is studying the survival times (in months) of patients with a chronic disease using an Accelerated Failure Time (AFT) model. The dataset includes three covariates:

Age: Patient's age (in years).

Treatment: A binary variable (1 = Received treatment, 0 = No treatment).

Smoking: A binary variable (1 = Smoker, 0 = Non-smoker).

The model assumes a Weibull distribution for survival times. Below is the R output from fitting the AFT model using the survival package:

```
Call:
survreg(formula = Surv(time, status) ~ Age + Treatment +
Smoking, data = dataset, dist = "weibull")

Coefficients:
                Value Std. Error     z       p
(Intercept)     4.500      0.300 15.00 <0.001
Age            -0.015      0.007 -2.14   0.032
Treatment       0.800      0.150  5.33 <0.001
Smoking        -0.600      0.200 -3.00   0.003

Scale= 1.200
Log-likelihood: -120.5
AIC: 251.0
```

i) Write the general form of the AFT model equation for this output. **[2 Marks]**

ii) Interpret the coefficients for Age, Treatment, and Smoking in terms of their impact on survival time. **[3 Marks]**

iii) Explain the meaning of the scale parameter (Scale=1.200) in the Weibull AFT model. **[1 Mark]**

iv) Calculate the acceleration factor (exp(β)) for Age, Treatment, and Smoking. **[2 Marks]**

v) What does the acceleration factor for Treatment indicate about its effect on survival time? **[2 Marks]**

vi) Discuss the quality of the model fit based on the log-likelihood and AIC values. **[1 Mark]**

vii) What assumptions are specific to the Weibull AFT model, and how could you test if this distribution is appropriate for the survival data? **[3 Marks]**