Stockholm University
Department of Statistics
(Gebrenegus Ghilagaber)

Home-Examination for Analysis of Survival Data with Demographic Applications (ST303G), Basic-level course, 7.5 Credits, Spring 2021.

The exam is handed-out on **Monday 15 March 2021 at 13:00** and replies are expected to be handed-in electronically (according to the separate instructions) latest by **Monday 22 March 2021 at 13:00**.

The examination consists of 8 questions that add up to 60 points. **Detailed and well-motivated replies are required in order to get full marks on each question**.

Your replies should include all detailed steps and results (for the theoretical/analytical questions), and explanations as well as most relevant tables and/or figures of results together with their interpretations, conclusions, and implications (for the empirical questions). Input codes should be included as an appendix.

The examination will be graded according to the 7-scales that are described in the course description distributed during the start of the course.

For questions about the **content** of the exam, contact the course coordinator via email Gebre@stat.su.se. Incoming questions will be answered continuously during the exam period.

If the course coordinator needs to send out information to all students during the exam, this is done to your registered email address. Therefore, check your email during the exam period.

NOTE! The exam shall be submitted electronically via the department's web site **no later than 13.00 (1 pm) on Monday 22 March 2021**. **The system does not allow submission after the deadline** which is a new setup for this semester. Therefore, start the submission well in advance. The last hour of the exam time is intended for arranging the electronic submission.

Please note that practical help is only available during the **first day** of the exam by email to expedition@stat.su.se. Read carefully the enclosed instructions for exam submission. There, you find all the necessary information about submission, anonymous code, extended writing time etc. If you, despite the instructions have problems submitting the exam, email the exam to tenta@stat.su.se. However, this is only done in exceptional cases. Exams sent in by email after deadline will not be corrected.

NOTE! All forms of cooperation and plagiarism are prohibited. We go over all exams carefully to detect cheating. Suspected cheating is reported to the Disciplinary Board and can lead to suspension.

Part 1: Theoretical/Analytical Questions:

Question 1 (6 p)

An event can occur due to one of three causes. The times to event due to these three causes (say T_1, T_2, T_3) are assumed to be independent and exponentially distributed with parameters λ_1, λ_2 and λ_3, respectively. Let T be the time to the event (due to any of the causes).

a. (4p) Derive the distribution of T.

b. (2p) Compute the probability that T is less than 5 if $\lambda_1 = 0.5$, $\lambda_2 = 1$, and $\lambda_3 = 1.5$.

Question 2 (6 p)

The log-rank test for comparing survival experiences in two groups may be viewed as originating from the hyper-geometric distribution where, at each event time, the observations (in each of two groups) may be considered as belonging to events or non-events.

a. (4p) Use the properties of the hyper-geometric distribution (especially its mean and variance) to derive the log-rank test statistic.

b. (2p) What assumptions are made (implicitly or explicitly) in deriving the test statistic in (a) above?

Part 2: Analyses and interpretation of real-life data set.

Questions 3-7 are based on the following data set (while question 8 has 'its own' data set):

The attached file **HomeExam-Questions3-7.xlsx** contains data on entry into marriage among a sample of women. The six columns represent the following variables:

Column 1: Household to which the respondent belongs
Column 2: Years (after age 15) to entry into parenthood or to the survey date
Column 3: Indicator of transition to parenthood (0: not yet married, 1: married)
Column 4: Birth Cohort with 7 levels (1 indicates the youngest cohort and 7 the oldest cohort)
Column 5: Residence area with 2 levels (1: Urban, 2: Rural)
Column 6: Education with 4 levels (0: No Educ, 1: Primary, 2: Secondary, 3: Above secondary)

Question 3 (8p)

a. (4p) Model the probability of experiencing the event (entering into marriage) as a function of the three covariates (birth cohort, residence area and education) and interpret your results.

b. (4p) Use the estimates in (a) above to estimate the probabilities of marriage for each level of the three covariates.

Question 4 (8p)

a. (2p) Estimate the survival curves for the different levels of *birth cohort* and test if there is a significant difference between them.

b. (2p) Estimate the survival curves for the different levels of *residence area* and test if there is a significant difference between them.

c. (2p) Estimate the survival curves for the different levels of *education* and test if there is a significant difference between them.

d. (2p) How do your conclusions in 4(a) – 4(c) above compare with your results in question 3?

Question 5 (10p)

a. (2p) Model the intensity of marriage as a function of birth cohort. Use the first level of the covariate as baseline (reference) level. Interpret the results and draw your conclusions.

b. (2p) Model the intensity of marriage as a function of birth cohort and residence area (using the first levels of each covariate as baseline/reference levels). Interpret the results and draw your conclusions.

c. (2p) Model the intensity of marriage as a function of the three covariates (birth cohort, residence area and education). Use the first levels of each covariate as baseline (reference) levels. Interpret the results and draw your conclusions.

d. (4p) Test, separately, if adding residence area in 5(b) and education in 5(c) improve the fit of the respective models.

Question 6 (10p)

Suppose now we are interested in modeling the effect of the three covariates on the time to marriage.

a. (6p) Fit all possible models using all three covariates in the models (using, again, the first levels as baseline) and interpret your results.

b. (4p) Which model fits the data 'best'? Justify your answer.

Problem 7 (5p)

In what way do you think the information in *column 1* can be used to improve the analysis of the data set and get better insights into the effect of the covariates on the event of interest?

Question 8 (7p)

Consider the following table of occurrences (Occ) and Exposures (Exp) cross classified by 3 levels of a covariate and 3 time intervals (also attached as HomeExam-Question8.xlsx):

Covariate	Time-Interval	OCC	EXP
1	1	1279	87390.5
	2	377	50092
	3	272	35933
2	1	746	55201
	2	167	34857
	3	167	49845
3	1	212	21032.5
	2	45	17245
	3	14	23809

a. (3p) Fit appropriate model for the rate of occurrence as a function of the time interval and compare the results with the crude Occ/Exp rates from the table.

b. (4p) Fit appropriate model for the rate of occurrence as a function of the covariate and the time interval (with both included in the model). How do the estimates of the effects of time interval change between (a) and (b)?