Stockholm University
Department of Statistics
(Gebrenegus Ghilagaber)

Home-Re-Examination for Analysis of Survival Data with Demographic Applications (ST303G), Basic-level course, 7.5 Credits, Spring Semester 2023.

The exam is handed-out on **Thursday 20 April 2023 at 13:00** (available, together with any other associated file/s, at the course-site in Athena under the sub-directory Home-Re-Exams). Replies are expected to be handed-in electronically (via http://tenta.stat.su.se/tenta) latest by **Wednesday 26 April 2023 at 18:00**.

The examination consists of 9 questions that add up to 60 points. **Detailed and well-motivated replies are required in order to get full marks on each question**.

For the theoretical/analytical questions, your replies should include all detailed steps and results. For the empirical (data analyses) questions your replies should include the most relevant tables and/or figures of results together with their interpretations, explanations, conclusions, implications, etc. 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 http://tenta.stat.su.se/tenta **no later than 18.00 (6 pm) on Wednesday 26 April 2023**. **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)

Let T_1 and T_2 are two independent random variables (survival times), each exponentially distributed with parameters $\lambda_1 = 1$ and $\lambda_2 = 2$, respectively. These may correspond to survival times to an event of interest that can occur due to two causes (cause 1 and cause 2, respectively). Let T be the time to the event (due to cause 1 or cause 2, whichever comes first).

a. (4p) Find the distribution of T?

b. (2p) Compute P(T > 3).

Question 2 (4 p)

a. (3p) Derive the Log-rank test statistic using the properties of the hyper-geometric distribution (especially its expected value and variance).

b. (1p) What assumptions are made (implicitly or explicitly) in deriving the test statistic in (a) above?

Part 2: Analyses of real-life data set and interpretation of results.
Questions 3-8 are based on the following data set (Question 9 has its own data set):
The uploaded file **Home-Re-Exam2023-Questions3-8.xlsx** contains data on transition to parenthood among a sample of women. The seven columns represent the following variables:

Column 1: Household to which the respondent belongs
Column 2: Months (after marriage) to transition to parenthood or to the survey date
Column 3: Status Indicator (0: not yet parent, 1: parent)
Column 4: Region to which the woman belongs with 11 levels (1: Region 1, …., 11: Region 11)
Column 5: Residence area with 2 levels (1: Urban, 2: Rural)
Column 6: Birth Cohort with 7 levels (1 indicates the youngest cohort and 7 the oldest cohort)
Column 7: Education with 4 levels (0: No Educ, 1: Primary, 2: Secondary, 3: Above secondary)

Question 3 (4p)

Use a simple test of association (independence) to examine the association between the event of interest and each of the four categorical covariates (Region, Residence, Birth Cohort, Education). Note that this doesn't require use of the duration variable. Rather, you are expected to make four simple tests of association (for the association of each covariate with the event) and draw your conclusions based on the simple test statistic.

Question 4 (10p)

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

e. (2p) How does the adding a covariate region in 3(d) affects the effects of the other three covariates on the intensity of transition?

Question 5 (7p)

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

b. (3p) How do your results in 5(a) compare with those in 4(d) with regard to the effects of the covariates?

Question 6 (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) Estimate the survival curves for the different levels of *region* and test if there is a significant difference between them.

Question 7 (10p)

a. (3p) Estimate the overall intensity of experiencing the event as well as the mean and median survival times assuming that time until event (or censoring) is **exponentially** distributed.

b. (4p) Assume now that the time until event (or censoring) is **exponentially** distributed but with different parameters for the four levels of *Education*. Use appropriate procedure to test for the equality of the population-intensities of experiencing the event across the three residences.

c. (2) Are your results in 7(b) in accordance with those in 6(c)? If not, what do you think the reason can be?

Problem 8 (6p)

a. (3p) 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?

b. (3p) Use appropriate method to implement your suggestion in 8 (a) above on question 4(d) and see if/how the estimates change.

Question 9 (5p)

Consider the table of occurrences (Occ) and Exposures (Exp) cross classified by 3 levels of a covariate and 3 time intervals:

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. (2p) 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. (3p) 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)?