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 2022.

The exam is handed-out on **Monday 14 March 2022 at 15:00** and replies are expected to be handed-in electronically (via tenta.stat.su.se/tenta.html) latest by **Monday 21 March 2022 at 18:00**.

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 18.00 (6 pm) on Monday 21 March 2022. 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 (24 hrs.)** 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)

The time in years until family initiation though cohabitation, T_1, is assumed to follow an exponential distribution with parameter λ_1 while the time in years until family initiation though direct marriage, T_2, is assumed to follow an exponential distribution with parameter λ_2. Further, these two random variables, T_1 and T_2, are assumed to be independent. Let T be the time (in years) to family initiation (due to any of the two ways).

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

b. (2p) Compute the probability that a randomly selected individual is still single (has not yet initiated a family) four years after exposure if $\lambda_1 = 3$, and $\lambda_2 = 1$.

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 3-8 are based on the following data set:
The file **HomeExam2022-Questions3-8.xlsx** contains data on transition to parenthood among a sample of women. The six columns represent the following variables:

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

Question 3 (8p)

a. (4p) Model the probability of experiencing the event (entering into parenthood) 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 transition 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 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.** (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 transition to parenthood.

- **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 (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) Attempt to implement your suggestion in 7 (a) above on question 5 (c) and see if/how the estimates change.

Question 8 (6p)

a. (3p) Group the time-variable into four intervals such that the first interval covers the first 36 months after exposure, the second interval covers months 37 to 60 inclusive, the third interval covers months 61 to 96 inclusive and the last interval covers month 97 and above. Compute the number of events, exposure months, and occurrence/exposure (Occ/Exp) rates in each interval. Give also the total events and exposure.

b. (3p) Fit appropriate model for the rate of occurrence of the event as a function of the time intervals and compare the results with the crude Occ/Exp rates in 8 (a) above.