

## Exam: Stochastic Modelling and Processing

L. Anesen & R. Brooks

June, 2020

Please state all answers in the "SMP\_exam.ipynb"-file. If you have handwritten answers (that are scanned), please state "In paper" or similar so the examiner knows where to find the answer. Also, please include all ipynb files when you hand in.

### Assignment 1 (20%)

The length of time  $X$  (in hours), needed by students in the SMP course to complete the three hour exam is a continuous random variable with the following density function;

$$f(x) = \begin{cases} q(x^2 + x) & \text{if } 0 \leq x \leq 3 \\ 0 & \text{else} \end{cases}$$

- a. Find the value  $q$  that makes  $f$  a probability density function.
- b. Find the cumulative distribution function.
- c. Find the probability that a student will complete the exam in
  - (i) less than an hour
  - (ii) between one and two hours
  - (iii) more than two hours
  - (iv) during the final ten minutes of the exam.
- d. Find the mean time needed to complete the three hour SMP exam.
- e. Find the variance and standard deviation of  $X$ .

### Assignment 2 (10%)

Disease  $A$  occurs with probability 0.1, and disease  $B$  occurs with probability 0.2. It is not possible to have both diseases. You have a single test. This test reports positive with probability 0.8 for a patient with disease  $A$ , with probability 0.5 for a patient with disease  $B$ , and with probability 0.01 for a patient with no disease - call the latter event  $W$ . If the test comes back positive, what is the probability you have either:

- a. disease  $A$
- b. disease  $B$ , or
- c. neither

Note: You need to calculate three probabilities; one for each of the three events stated in a-c.

### Assignment 3 (15%)

- A survey of the approximately 17,000 VIA students revealed that almost 70% disapprove of daily pot smoking. If 18 of these students are selected at random and asked their opinion, what is the probability that more than 9 but fewer than 14 disapprove of smoking pot daily?
- In an NBA (National Basketball Association) championship series, the team that wins four games out of seven is the winner. Suppose that teams  $A$  and  $B$  face each other in the championship games and that team  $A$  has probability 0.55 of winning a game over team  $B$ . What is the probability that team  $A$  will win the series in 6 games? What is the probability that team  $A$  will win the series?
- Suppose a system contains a certain type of component whose time, in years, to failure is given by  $T$ , and that the mean time to failure is  $\beta = 6$ . If 5 of these components are installed in different systems, what is the probability that at least 2 are still functioning at the end of 8 years? (*Hint: You need to use two different distributions: one to find the probability that a given component functions after 8 years and another to find that at least 2 are functioning!*).

### Assignment 4 (20%)

Hydrocarbon emissions from cars are known to have decreased dramatically during the 2010s. A study was conducted to compare the hydrocarbon emissions at idling speed, in parts per million (ppm), for automobiles from 2010 and 2020. Fifty cars of each model year were randomly selected, and their hydrocarbon emission levels were recorded. The data is displayed in "HydroCarbonEmissions.xlsx".

- Determine estimates for the quartiles, average emission, standard deviation and variance of each model year.
- Setup 95% confidence intervals for the mean of each model year, and accompany the intervals with plots that display the rejection region
- Is it reasonable to assume that the emission of each model year is normally distributed? Explain using plots and discussing skewness and kurtosis.
- Setup a 99% confidence interval for the mean emission difference between the two model years and accompany the intervals with plots that display the rejection region.
- Is there *significant* evidence to support the claim that the mean emission difference between the two model years differ from one another?
- Is there evidence to support the claim that the standard deviations of the two model years differ *significantly*.

### Assignment 5 (15%)

The dataset for this assignment is "Wages\_and\_Work\_Hour.xlsx". This workbook contains data on full-time workers in East North Central United States from the March 1999 CPS. The objective is to determine whether Education, Income, and Gender differ.

#### Variable notes:

- Education Level: Group 1 has less than 13 years of education. Group 2 has between 13 and 15 years of education (both included). Group 3 has 16 years or more of education.
- Income Group: Group 1 has less than or equal to \$20,000 in income. Group 2 has between \$20,000 and \$48,000 in income (both included). Group 3 has more than \$48,000 in income.

- a. Create a contingency table, placing Gender on the vertical axis and Education Level on the horizontal axis, and test whether gender is independent of level of education.
- b. Create a contingency table, placing Gender on the vertical axis and Income Group on the horizontal axis, and test whether gender is independent of income.
- c. Create a contingency table, placing Education Level on the vertical axis and Income Group on the horizontal axis, and test whether Education Level is independent of Income Group.

## Assignment 6 (10%)

State the null and alternative hypotheses to be used in testing the following claims and determine generally where the rejection region is located (i.e. is it a right-, left- or two-tailed test):

- a. The mean snowfall at Bygholm during the month of February is 21.8 centimeters.
- b. No more than 20% of the faculty at VIA are competent teachers.
- c. On the average, children attend schools within 2.62 kilometres of their homes in Denmark.
- d. The proportion of voters favoring the incumbent in the upcoming American election is 0.38.
- e. The average cabbage at the grocery store weighs at least 240 grams (*Source*: our colleague Jakob Knop Rasmussen).

## Assignment 7 (10%)

A professor in the School of Engineering in a university polled a dozen colleagues about the number of professional meetings they attended in the past five years ( $x$ ) and the number of papers they submitted to refereed journals ( $y$ ) during the same period. The summary data are given as follows:

$$n = 12, \quad \bar{x} = 4, \quad \bar{y} = 12$$
$$\sum_{i=1}^n x_i^2 = 232, \quad \sum_{i=1}^n x_i y_i = 318$$

Fit a simple linear regression model between  $x$  and  $y$  by finding out the estimates of intercept and slope.  
*Hint*: Use the Least Squares Estimates formula from the book.