SMP Re-exam E-2018/2019

Richard Brooks

Note: Please state your answers in the attached notebook ("SMP_Re-exam.ipynb"). If you hand in any exercise by hand, please state so in the notebook under the respective assignment (e.g. state "# on paper" in the code block). Please make sure that you hand in in .pdf format (you can save the notebook as .html and then convert/print this to .pdf). Before handing in, change the name of the file to your name. Please make sure that your line of reasoning is apparent either by means of explanation or in terms of code or in terms of both.

Assignment 1 (15%)

Compute the expected value, $E(X)$, if X has a density function as follows:

a. $f(x) = \begin{cases} \frac{1}{4}xe^{-\frac{x}{2}} & x > 0 \\ 0 & otherwise \end{cases}$

b. $f(x) = \begin{cases} 5x^{-2} & x > 5 \\ 0 & otherwise \end{cases}$

The density function of X is given by

$$f(x) = \begin{cases} a + bx^2 & 0 \leq x \leq 1 \\ 0 & otherwise \end{cases}$$

c. If $E(X) = \frac{3}{5}$, find a and b.

Assignment 2 (10%):

You arrive at a train station at 10 o'clock, knowing that the train will arrive at some time uniformly distributed between 10 and 10:30.

a. What is the probability that you will have to wait longer than 10 minutes until the train arrives?
b. If at 10:15 the train has no yet arrived, what is the probability that you will have to wait at least an additional 10 minutes?

Assignment 3 (10%):

a. A man claims to have extrasensory perception. As a test, a fair coin is flipped 10 times and the man is asked to predict the outcome in advance. He gets 7 out of 10 correct. What is the probability that he would have done at least this well if he had had no extrasensory perception, i.e. if the probabilities of heads and tails were equally likely?

b. According to information about the consumption of repair parts, it was established that, in the repair of automobile engines, part No.1 was changed in 36%, and part No.2 in 42%, of the cases examined, and that both of these parts were changed simultaneously in 30% of the cases. On the basis of these data, is it possible to deduce that replacement of part No.1 and replacement of part No.2 are statistically dependent? Find the probability that in repairing an engine, part No.2 will be changed, under the condition that part No.1 has been changed.

Assignment 4 (25%)

Two producers of batteries measure the longevity of 30 batteries of the same type, which were randomly chosen from a larger batch of such batteries. The lifetime (in hundreds of hours) is displayed "Batteries.xlsx".

a. Check the dataset for outliers and replace any outliers with the mean lifetime of the producer in question. Use this cleaned dataset in the following questions.
b. Determine estimates for the quartiles, average lifetime, standard deviation and variance of each producer's battery
c. Setup 95% confidence intervals for each mean battery lifetime from the two producers, and accompany the intervals with plots that display the rejection region.
d. Is it reasonable to conclude that the lifetime of the two producer's battery follow a normal distribution? Explain using plots and discussing skewness and kurtosis.
e. Setup a 95% confidence interval for the difference between the two producer's battery, and accompany the intervals with plots that display the rejection region.
f. Is there *significant* evidence to support the claim that the mean lifetime of the batteries from the two producers differ from one another?
g. Setup a test to test whether the standard deviations of the two batteries differ significantly.

Assignment 5 (20%)

Data collected in 1960 from the National Cancer Institute provides the per capita numbers of cigarettes sold along with death rates for various forms of cancer (see "Smoking and Cancer.xlsx").

a. Build regression models with cigarettes sold as the independent variable and each of the four cancer types as the dependent variable. Accompany each model with a scatterplot and a trend line as well as confidence intervals for the regression parameters.
b. For each model, inspect the residuals to confirm that the assumptions about normality and non-patterns are met.
c. Which of the four cancer types exhibit the best correlation with cigarettes sold? Assess using the correlation coefficient.
d. In which data pairs is cigarettes sold a good predictor for the type of cancer? Assess using the correlation of determination and interpret the meaning of this number.
e. For the model that has the best correlation, find the predicted value of deaths per 100k for 40 and 50 cigarettes sold per capita. Feel free to include 95% prediction intervals to your predictions.

Assignment 6 (20%)

The dataset for this assignment is the infamous Titanic data set (please see "Titanic.xlsx"). The objective is to determine whether the survival rates differ between selected variables.

Variable notes:

- pclass: A proxy for socio-economic status
- 1st = Upper
- 2nd = Middle
- 3rd = Lower
- age: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5
- sibsp: The dataset defines family relations in this way...
- Sibling = brother, sister, stepbrother, stepsister
- Spouse = husband, wife (mistresses and fiancés were ignored)
- parch: The dataset defines family relations in this way...
    - Parent = mother, father
    - Child = daughter, son, stepdaughter, stepson

    - Some children travelled only with a nanny, therefore parch=0 for them.
- alone: is a variable that was created from combining sibsp and parch.

a. Create a contingency table, placing survived on the vertical axis and pclass on the horizontal axis.
b. Test whether survival rate is independent of pclass.
c. Create a contingency table, placing survived on the vertical axis and sex (gender) on the horizontal axis.
d. Test whether survival rate is independent of sex (gender).
e. Create a contingency table, placing survived on the vertical axis and alone on the horizontal axis.
f. Test whether survival rate is independent of whether a person travelled alone or not.