Introduction to Probability Theory and Statistics
Exam, June 2020

Probability

Problem 1: Simplify the expressions for the events ABC and $A \cup B \cup C$ in the following cases (Hint: use Venn diagrams)

(a) $A \subset B$ and $A \subset C$

(b) $B \subset C$ and $A \subset B$

(c) $B \subset C$ and $A \subset C$

Problem 2: A fair coin is flipped twice. Let X be a number of observed heads in these two trials and Y be a number of observed tails.

(a) find joint pmf of X and Y

(b) find marginal pmfs

(c) Are X and Y independent?

(d) Find the mean and variance of X and Y

(e) Write the covariance matrix of X and Y

(f) Draw a pmf and a cdf of X

(g) Let a new random variable be defined $Z = 2(X + Y)$. Find the mean and variance of Z.

Problem 3: According to some studies, the duration (in minutes) of a phone call to an IT Support has an exponential distribution with parameter $\lambda = 0.2$.

(a) What is the probability that a phone call lasts less than 5 minutes?

(b) If a person is already speaking with IT Support for 3 minutes, what is the probability that he will be speaking for additional 2 minutes?

(c) Write down a cdf for the phone call duration. Using it, please show how to calculate the probability that the duration of a call exceeds 3 minutes, but is less than 5 minutes?

(d) Write down a pdf for the phone call duration. Answer the previous question (c), but now make your calculations using a pdf instead. That is, using a pdf, please show how to calculate the probability that the duration of a call exceeds 3 minutes, but is less than 5 minutes?

(e) What is the probability that a phone call lasts exactly 3 minutes?

(f) What is the average duration of a call?

(g) Imagine that a new inexperienced team for IT Support has been hired. This resulted in increase of the average call duration twice. How parameter λ of the distribution has been affected, assuming that the call duration can be still modelled as an exponential distribution?

Statistics

Problem 4: In an outbreak like COVID-19, the case fatality rate is one of the key indicators for epidemiology. It is defined as the proportion of deaths from a certain disease compared to the total number of people diagnosed with the disease for a certain period of time. To calculate this value, the table below shows the confirmed cases and mortality for the first ten days since the report of first death from COVID-19 in a European country.

Day	D (number of deaths)	T (total number of confirmed cases)
1	2	222
2	3	259
3	5	400
4	10	500
5	17	673
6	28	1073
7	35	1695
8	54	2277
9	55	2277
10	133	5232

(a) Use a linear regression to model the dependence of the number of deaths with the total number of confirmed cases. Determine the case fatality rate, calculated as the slope of the regression model.

(b) Calculate the coefficient of determination.

(c) In some countries the exponential growth is a better model for the initial stages of the epidemic, such that: $D = A \cdot e^{B \cdot T}$ (D is the number of deaths, T is the total number of confirmed cases and A and B are unknown constants). Repeat the regression with this model and calculate the coefficient of determination.

(d) Which model, linear or exponential, describes better the data available for this country? Why?

Problem 5

Let $(X_1, X_2, ..., X_n)$ be the sample of size n from a population distributed according to $f(x, \theta)$, where $\theta > -1$ is unknown.

$$f(x, \theta) = (\theta + 1) \cdot x^{\theta} \tag{1}$$

(a) Find the maximum likelihood estimator of θ

(b) Estimate the value of θ when the random sample is: 0.7 0.9 0.6 0.8 0.9 0.7 0.9

Problem 6

According to the World Bank Open Data, 26 percent of the male population in Denmark aged 15 and over smoked in 2010. A scientist has recently claimed

that this percentage has since increased, and to prove his claim he randomly sampled 1000 individuals from this population.
(a) If 268 of them were smokers, is his claim proved? Use the 5 percent level of significance.
(b) Repeat the test using a normal approximation