Analytic for Observational Data

(IT142IU)

Lab 1-2: Probability distributions

Student Name: Nguyen Duc Nguyen Phuc

Student’s ID: ITDSIU21108

* 1. Objectives

- Apply probability distributions to the provided datasets.

- Apply moment generating functions to find the moments of random variables.

- Dataset sources:

<https://www.kaggle.com/datasets/fedesoriano/wind-speed-prediction-dataset>

<https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store>

<https://archive.ics.uci.edu/dataset/252/climate+model+simulation+crashes>

* 1. Exploring the data

|  |  |
| --- | --- |
| Questions | Answers |
| Data set name | Wind Speed Prediction Data-set |
| Identify data objects, attributes,and attribute types. | Data object:  Wind Speed  Attributes:  DATE, WIND, IND, RAIN, IND.1, T.MAX, IND.2, T.MIN, T.MIN.G  Attributes types: |
| Find and choose the data objects changing over time. | WIND: Average wind speed [knots]  RAIN: Precipitation Amount (mm)  T.MAX: Max temperature(0C)  T.MIN: Min temperature (0C)  T.MIN.G: Grass Minimum temperature (0C) |
| Identify and describe the data attributes that are considered as random variables from the chosen data objects.  Note: at least two data attributes chosen. | WIND: Average wind speed [knots]  RAIN: Precipitation Amount (mm)  T.MAX: Max temperature(0C)  T.MIN: Min temperature (0C)  T.MIN.G: Grass Minimum temperature (0C) |
| Draw box plots for each numeric attribute, present five-number summaries.  Note: recognize appropriate data areas in the data and draw boxplots |  |
| Present the distributions of data regions of random variables using probability  functions.  What are the probability distribution function? | I used Kolmogorov-Smirnov Test (N >5000) to check normal distribution. The null hypothesis is rejected if the p-value is less than the selected significance threshold, such as 0.05, which shows that the data is not regularly distributed:    We use kernel density estimation (KDE) to approximate the probability density function (PDF) because the data does not follow a normal distribution. KDE is a non-parametric technique that uses the data points that are available to estimate the PDF of a random variable:    The PDF for Rain and Wind:  PDF of wind:    PDF of rain: |
| Find the first and second moments and central moments of the random variables. | Code: |
| Remark random variables if they are useful for modeling or learning (classification or clustering) | I used k-means to cluster and PCA to decrease dimension  and group by Clustering to calculate mean and var of Wind, Rain, T.Max, T.Min  we can summary: |