Analytics for Observational Data (IT142IU)

Lab 1-2: Probability distributions

1.1. 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/berkeleyearth/climate-change-earth-surface-temperature-data
    - https://www.kaggle.com/mkechinov/ecommerce-behavior-data-from-multi-category-store
    - https://archive.ics.uci.edu/ml/datasets/climate+model+simulation+crashes
- Programming languages: Python/Java

1.2. Exploring the data

Questions	**Answers**
Dataset name	
Identify data objects, attributes, and attribute types.	
Find and choose the data objects changing over time.	
Identify and describe the data attributes that are considered as random variables from the chosen data objects. Note: at least two data attributes chosen.	

Draw boxplots 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.	
Find the first and second moments and central moments of the random variables.	
Remark random variables if they are useful for modeling or learning (classification or clustering)	

1.3. Some references

- Gallery of Distributions: https://www.itl.nist.gov/div898/handbook/eda/section3/eda366.htm
- NumPy: https://numpy.org/doc/stable/reference/random/generated/numpy.random.normal.html
- Seaborn: https://seaborn.pydata.org/generated/seaborn.distplot.html
- SciPy: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.moment.html