CSDS 313/413: Introduction to Data Analysis Fall 2025

Homework 1: Data & Distributions

©Vo Linh Chi Dao, Mehmet Koyutürk, Khoa Tran, Rui Yang, Crystal Zhu @ CWRU

Task 1: Comparing Visualizations

In data science, the choice of visualization plays a critical role in shaping how insights are derived and communicated. Different visual encodings can highlight or obscure structure in data — including spread, skew, outliers, modality, or differences between categories. In this task, you will explore how three different types of visualizations can be used to compare distributions across categories, and evaluate which is most appropriate depending on the dataset context.

For this task, you are provided with three datasets, each containing categorical grouping variables and a numerical measurement. Each dataset comes with a research scenario/question. Your task is to clean the data, visualize the distribution across categories using multiple plotting techniques, and discuss which visualization is the most appropriate in addressing the research question for each dataset.

Datasets and Research Question Associated

- **Dataset 1: Best-Selling Albums**
  - *Attributes:* Year, Ranking, Artist, Album, Genre, Worldwide Sales, Tracks, Album Length
  - *Scenario:* A media analytics firm is interested in understanding whether certain genres consistently produce top-selling albums or if success is more scattered across genres.
  - *Research Question:* How does the distribution of album sales vary across music genres for albums in the previous decade (released after 2015), and are high-sales outliers concentrated in certain genres?
- **Dataset 2: Anime**
  - *Attributes:* Rank, Name, Japanese_name, Type, Episodes, Studio, Release_season, Tags, Rating, Release_year, End_year, Description, Content_Warning, Related_Mange, Related_anime, Voice_actors, staff
  - *Scenario:* A streaming service is considering expanding its short anime series catalog (< 25 episodes) and wants to understand how viewer ratings differ between anime TV series and movies released after 2015. The goal is to determine which format generally receives better audience reception to inform licensing and promotion strategies.
  - *Research Question:* How do audience ratings compare between anime TV series and movies released after 2015, and which format generally receives higher ratings?
- **Dataset 3: Algorithm Performance (Synthetic)**
  - *Attributes:* Algorithm, Epoch, Accuracy, Trial Number

- *Scenario:* You are testing two reinforcement learning (RL) algorithms on a sequential decision task. To avoid overfitting and simulate real-world noise, you shuffle the dataset for each trial and run 10 independent trials per algorithm. For each trial, you track the accuracy across 10 training epochs (one pass through a dataset). Due to how you shuffle your data and algorithmic stochasticity, accuracy results vary across trials.
- *Research Question:* Which algorithm performs more accurately on average across epochs, and how does the use of a visualization help you assess reliability and variation of each algorithm?

Part A: Data Cleaning and Preprocessing

First, filter your dataset so that only the variables critical for your analysis remain. Then clean your data so that there is consistency in variable types, capitalization, and handle any missing or invalid values.

Part B: Generate Three Visualizations per Dataset

For each dataset, produce the following types of plots:

- **Error Bar Plot:** Show the mean and variability (e.g., standard error or 95% confidence intervals) of the numerical variable across each category.
- **Barcode Chart:** Also known as a strip plot or rug plot. Shows individual data points across categories.
- **Histogram:** Plot the distribution of the numerical variable, grouped by the categorical variable (using hue or facet).

Part C: Evaluate and Justify Visualization

For each dataset:

- Discuss the advantages and disadvantages of each visualization type.
- Decide which visualization is best for the research question.
- Support your answer with evidence from the plots and reasoning based on dataset size, shape, or structure.

Task 2: Fitting and Comparing Distributions

In this task, you will explore how different types of real-world datasets may follow different distributions. You will need to develop a set of hypotheses and perform experiments to validate your own hypotheses.

Part A: Developing Hypotheses

Based on your understanding of the properties of each distribution, identify and collect four real-world datasets, each of which you hypothesize follows one of these distributions:

- Normal distribution
- Uniform distribution
- Power law distribution
- Exponential distribution

In other words, find one dataset (multiple observations of a single variable) that you think follows the Normal Distribution, one for the Uniform Distribution, one for the Power-law distribution, and one for the Exponential distribution. Please be clear about the reasoning behind your hypothesis and be specific about the source of the dataset. (If you are unsure where to find datasets, Kaggle and Google Dataset Search are good places to start)

Part B: Fitting Distributions

For this exercise, we will call each of the four different theoretical distributions (normal, uniform, power law, exponential) a "model". For each dataset, fit it (i.e., estimate the model parameters) against each model (not just the one you hypothesized) using maximum likelihood estimation (or using any technique you think is appropriate; make sure to comment on the validity of your approach). This should result in a total of **16 parameter sets**. Report the estimated parameters in the following tabular format:

		Model			
Dataset	***# Observations***	**Normal**	**Uniform**	**Power law**	**Exponential**
Dataset 1	n_1	μ_1, σ_1	a_1, b_1	$\alpha_1, x_{\min_1}$	λ_1
Dataset 2	n_2	μ_2, σ_2	a_2, b_2	$\alpha_2, x_{\min_2}$	λ_2
Dataset 3	n_3	μ_3, σ_3	a_3, b_3	$\alpha_3, x_{\min_3}$	λ_3
Dataset 4	n_4	μ_4, σ_4	a_4, b_4	$\alpha_4, x_{\min_4}$	λ_4

Be sure to show the code you used to arrive at your final estimates clearly.

Part C: Comparing Real and Synthetic Data

- For each fitted distribution (there will be 16 of them, four for each dataset $1 \leq i \leq 4$, each corresponding to a different model), generate a synthetic sample (n_i data points) using the respective model parameters you inferred from the real dataset.

- Compare the real vs. synthetic data distributions using methods you think are the most appropriate, including visualizations. So, for each dataset, we compare the original dataset to four synthetic datasets, all with equal number of observations, but each synthetic dataset is generated using a different model.
- For each dataset, identify the synthetic dataset (which corresponds to a model) that is most similar to the original data in terms of its distribution.
- Now revisit your initial hypothesis. For each dataset: Did the dataset behave as expected, or was another model (assumed distribution) a better fit to the dataset? Reflect on why the observed results may differ from your expectations.