DS 116 / CS 343 Data Visualization   Spring 2025

Homework N2

Part 1: Theoretical Questions

1. In a dataset with a non-normal distribution and potential extreme values, how are the whiskers in a boxplot determined, and what are the limitations of the standard IQR based rule in such cases?

- In a boxplot, the whiskers stretch to the smallest and largest values that fall within **1.5 times the IQR** from **Q1 and Q3**, while anything beyond that is flagged as an outlier. This method works well for many distributions but can be misleading for **skewed data** or datasets with **legitimate extreme values**, as it might wrongly classify them as outliers.

2. Given a dataset with heavy skewness and multiple peaks, how can a boxplot misrepresent outliers, and what alternative methods exist for identifying them more accurately? A boxplot can misclassify outliers in **skewed or multi-modal data**, either marking valid extreme values as outliers or missing actual anomalies. Better alternatives include **log transformation** (to reduce skew), **Modified Z-Score** (using median/MAD), **DBSCAN** (density-based detection), and **Isolation Forest** (ML-based anomaly detection). **Histograms or KDE plots** can also help visualize unusual patterns.
3. Explain the conceptual difference between median and mean in the context of non symmetric distributions. Why does a boxplot prioritize the median, and in what cases could this choice obscure important data characteristics?

- The **mean** is sensitive to extreme values, pulling toward skewed tails, while the **median** is robust, reflecting the dataset’s central position without being influenced by outliers. A **boxplot prioritizes the median** because it better represents the data’s true center in **non-symmetric distributions**. However, this can obscure **important data characteristics**, such as when **a long tail or multiple peaks** shift the mean significantly, hiding insights about skewness or bimodal trends.

4. If a boxplot exhibits strong right skewness, what can you infer about the underlying probability distribution? How would this skewness affect statistical measures such as variance, skewness coefficient, and potential model assumptions?

- A **strong right-skewed boxplot** suggests the underlying **probability distribution** is **positively skewed**, meaning a long right tail with extreme high values.
  - **Variance** increases due to large deviations on the higher end.
  - **Skewness coefficient** is **positive**, confirming asymmetry.
  - **Model assumptions** (e.g., normality in regression) may be violated, requiring **log transformation** or **non-parametric methods** for better analysis.

5. Why are boxplots particularly useful for comparing multiple groups in high-dimensional data? What are the limitations of boxplots when dealing with overlapping distributions or categorical variables with small sample sizes?

- **Boxplots are great for comparing multiple groups in high-dimensional data** since they highlight key stats like median, IQR, and outliers while keeping things visually clean. But they fall short when distributions overlap, sample sizes are tiny, or the data has multiple peaks, making it easy to miss important patterns. In those cases, violin plots or swarm plots do a better job of showing the full picture.

6. What are the theoretical consequences of selecting an inappropriate number of bins in a histogram, particularly in datasets with varying density regions or multimodal distributions? How does bin width selection affect kernel density estimation (KDE)?

- Too few bins can oversimplify the data, while too many can introduce noise and obscure patterns. For KDE, bandwidth selection plays a similar role; too small a bandwidth leads to overfitting, while too large smooths out important details.

7. Histograms and bar charts both use rectangular bars to display data. How does the interpretation of frequency differ in these two visualizations, and why is bin choice irrelevant in bar charts but crucial in histograms?

- In histograms, bars represent frequency for continuous intervals, so bin size affects interpretation.
- Bar charts, on the other hand, display counts of categorical data where binning doesn't apply.

8. Under what conditions might a histogram distort the perception of a dataset's distribution? Provide an example where binning choices lead to misleading conclusions, and explain how alternative visualizations (e.g., KDE or violin plots) could address these distortions.

- A poorly chosen bin width might create false peaks or hide patterns. For example, using too few bins could make a multimodal distribution appear unimodal. KDE or violin plots smooth the distribution, giving a clearer picture of its shape.

9. How does a density plot differ from a histogram in terms of its mathematical foundation and interpretability? What challenges arise when choosing a kernel function and bandwidth for density estimation, particularly in sparse datasets?

- Density plots estimate the probability density function using kernel functions and bandwidth, while histograms simply count frequencies. Choosing the wrong bandwidth can lead to oversmoothing (missing important details) or undersmoothing (capturing noise), especially in sparse datasets.

10. Explain why the area under a density plot is always equal to 1. How does this property relate to probability theory, and what implications does it have for comparing distributions with different sample sizes?

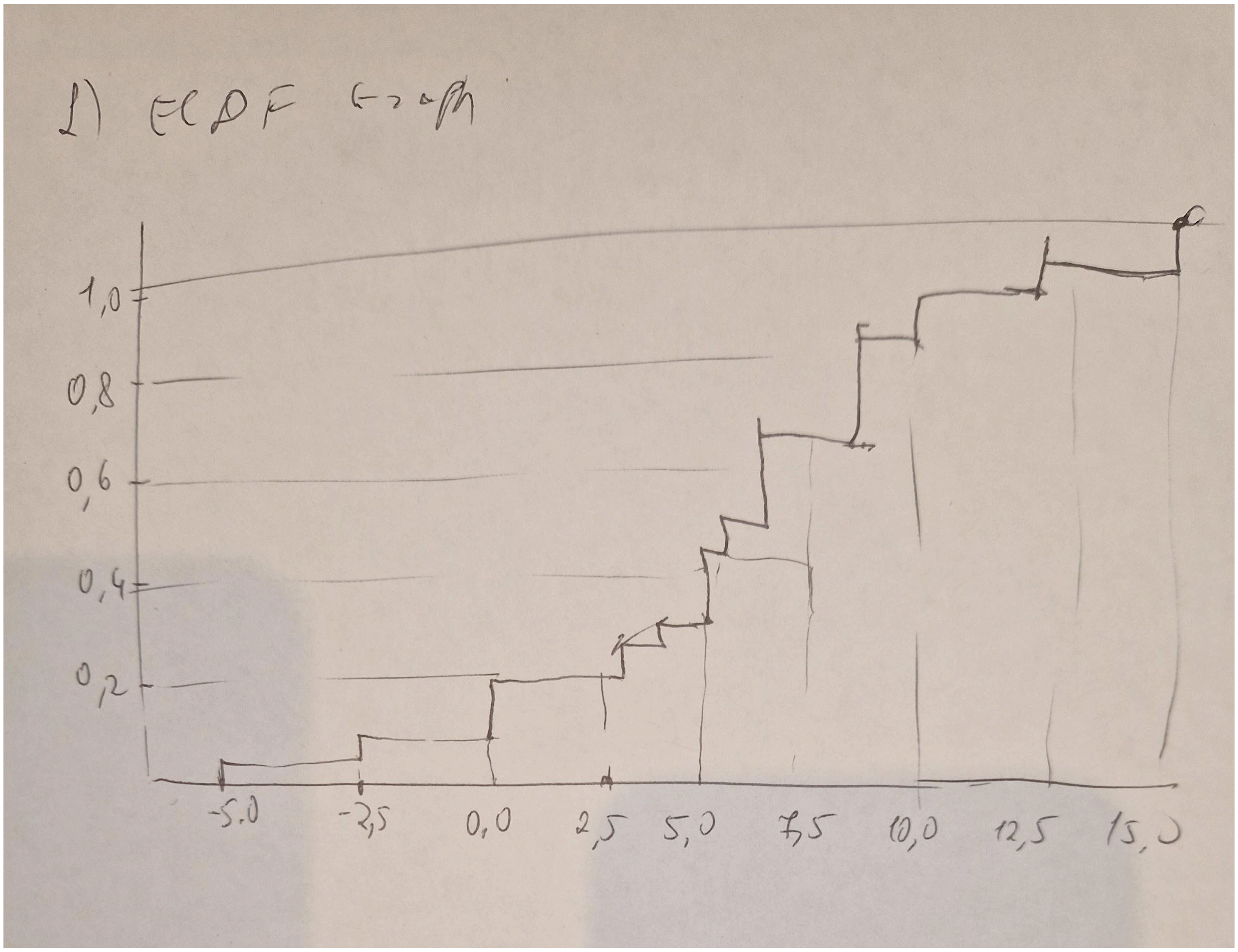
- A density plot represents a probability distribution, so the area under the curve equals

1 by definition. This allows comparison across distributions with different sample sizes since the total probability remains consistent.

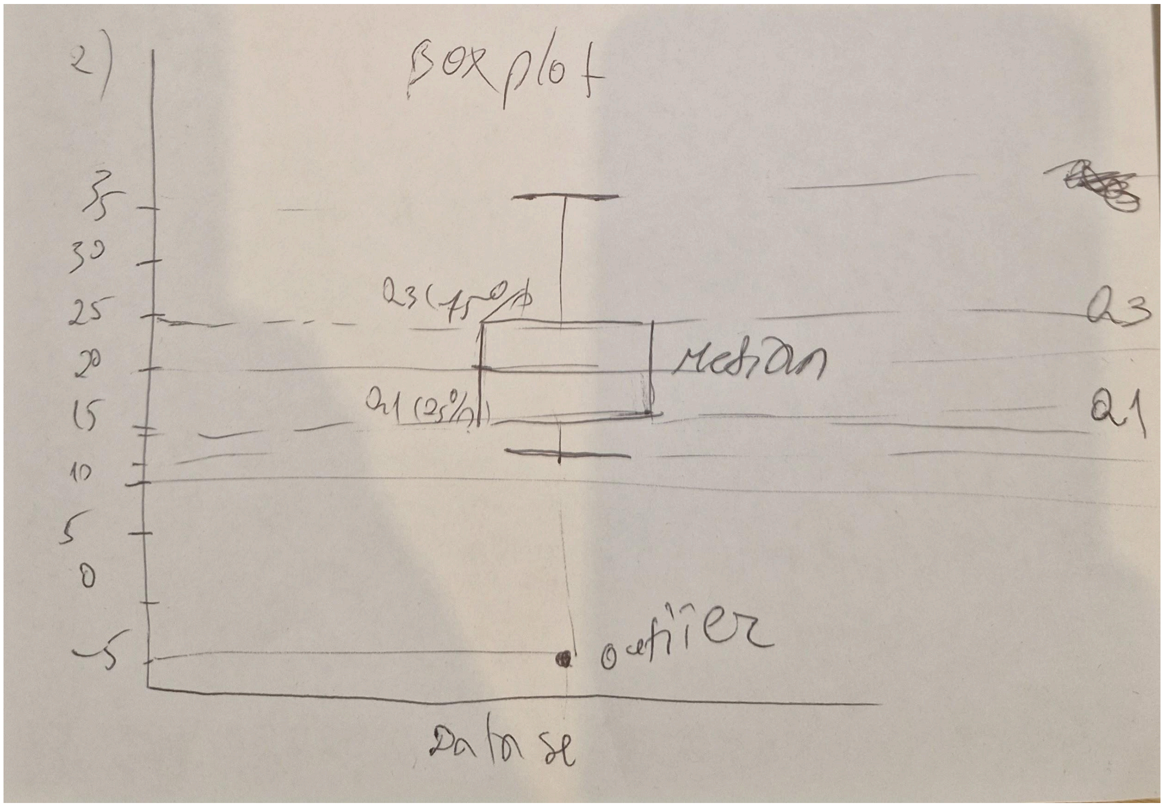
<u>Part 2: Hand-Drawn Graphs</u>

Create graphs by hand using the provided datasets.

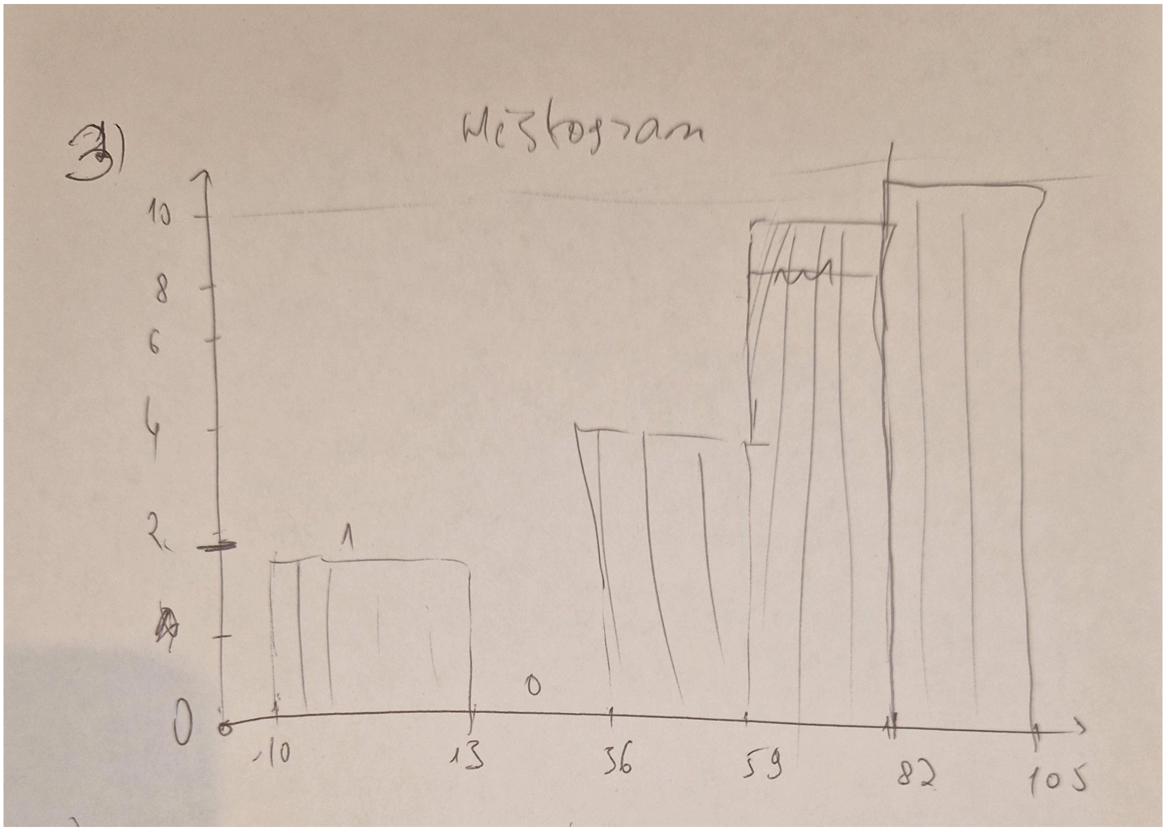
1. Given the numbers: -5, -2, 0, 3, 4, 5, 5, 6, 7, 7, 8, 9, 9, 10, 12, 15, draw an ECDF plot.



2. Given the dataset: -5, 12, 14, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 24, 25, 29, 30, 35, create a boxplot. Indicate the median, quartiles, and any potential outliers.



3. Given the test scores: -10, 45, 50, 55, 55, 60, 62, 65, 68, 70, 73, 74, 80, 80, 82, 85, 88, 90, 91, 92, 94, 97, 100, 105, create a histogram using 5 bins and label the axes.


3)
Histogram