American University of Armenia, Spring 2025
CS 343 Data Visualization
Hovsepyan Sargis

Homework 02

Part 1: *Answer the questions in written form. Provide 2-3 sentences for each question, bring examples when needed. Collect your answers in a PDF file.*

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?

The whiskers are determined according to the following logic:

Lower limit = Q1 – 1.5 x IQR
Upper limit = Q3 + 1.5 x IQR

The whiskers are chosen to be the min and max of the dataset elements falling into this range. Anything outside this range is classified as outliers.

The limitations of this approach are heavy tail distributions, when the majority of the tail might be classified as outliers although they might be naturally occurring observations.

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?

As already stated in highly skewed data boxplot will give unevenly distanced whiskers and will classify many values as outliers. Having multiple peaks will not easy the misinterpretation of outliers but will most probably add to it, by not capturing all the meaningful peaks appropriately. Therefore, in datasets like this it might be better to rely on robust outlier detection methods like the Modified Z-score or adjusting the IQR multiplier to better fit the distribution. One can also use visualization techniques like violin plots to better understand data spread.

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 the arithmetic average and is sensitive to extreme values, making it shift toward skewed data points. In contrast, the median is the middle value and is more robust to outliers, providing a better central tendency measure for non-symmetric distributions.

Therefor, boxplot prioritizes the median as it effectively represents the center of a skewed distribution without distortion from outliers. However, in multi-modal distributions, the median may not capture important clusters, and using only the median could hide meaningful patterns.

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?

Well, strong right skewedness indicates that most of the points are concentrated on the left side and we have a tail of huge values on the right. Clearly the variance increases due to the extreme values buildup at the right part. Similarly, the skewness coefficient is positive. As skewness violates the assumptions of normal distributions which many model use, the initial distribution should be transformed before being fed to a model (e.g. Box-Cox Transformations).

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 useful in the mentioned matter because they provide a concise summary of distribution characteristics like median, variability, and outliers, allowing for quick visual comparisons between groups, making them ideal for spotting trends, differences, or anomalies. However, when distributions significantly overlap, boxplots may fail to distinguish between abovementioned characteristics. Similarly, for categorical variables with small sample sizes, boxplots can misrepresent variability, as their structure relies on having enough data points to accurately define quartiles and whiskers.

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)?

If too few bins are used, the histogram becomes over-smoothed, potentially hiding important features like multiple peaks or skewness. Conversely, if too many bins are used, the histogram can become too noisy, making it difficult to distinguish meaningful patterns from random fluctuations. Similarly, in KDE, the bandwidth parameter controls the smoothing level, where a large bandwidth over-smooths the distribution and may hide multimodal features, while a small bandwidth can introduce artificial noise, similar to excessive binning in a histogram.

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?

Histograms show continuous data with bars representing frequency density, where bin choice affects data interpretation. In contrast, bar charts display categorical data with discrete bars, making binning irrelevant. In histograms, bin width impacts shape and trends, while in bar charts, bar placement and order have no statistical meaning.

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 histogram can distort a dataset's distribution if bin width is too large or too small. For example, in a bimodal distribution, poor binning may merge two peaks into one, misleadingly suggesting a unimodal pattern. Alternatively, KDE or violin plots can smoothly represent density without binning issues, preserving modality of structures, providing a clearer, more continuous view of the true data distribution.

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?

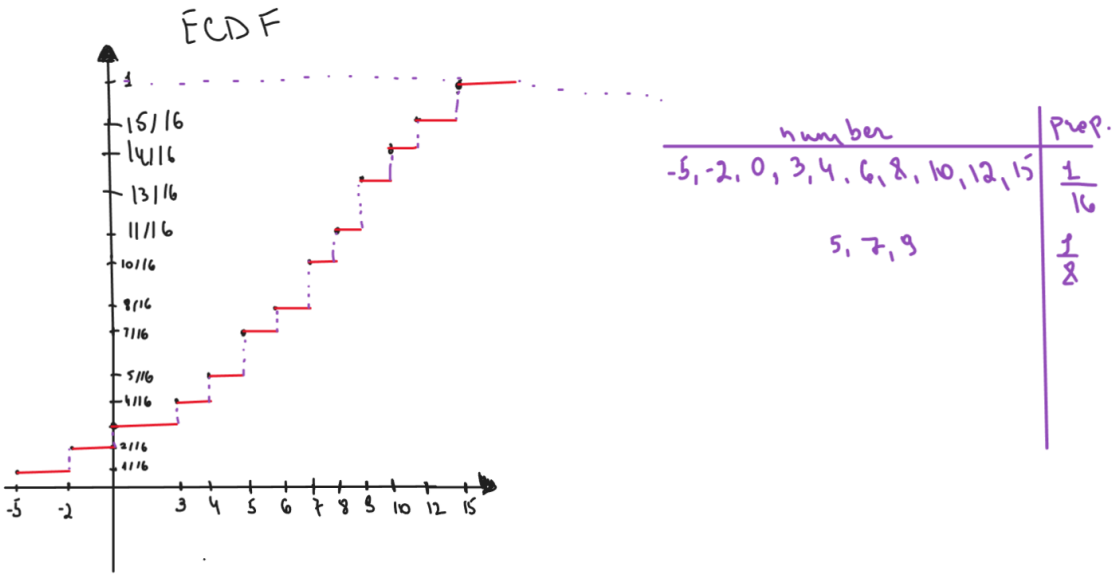
The density plot provides a continuous, smooth representation of the data distribution, while histograms provide a discrete, step-wise view. Choosing the right kernel function affects the smoothness and shape of the density estimate. The bandwidth controls the smoothing level, and selecting an inappropriate bandwidth can lead to under-smoothing (overfitting) or over-smoothing (losing detail). In sparse datasets, small sample sizes make bandwidth selection particularly challenging, as it can overestimate or misrepresent data features, potentially masking important patterns.

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?

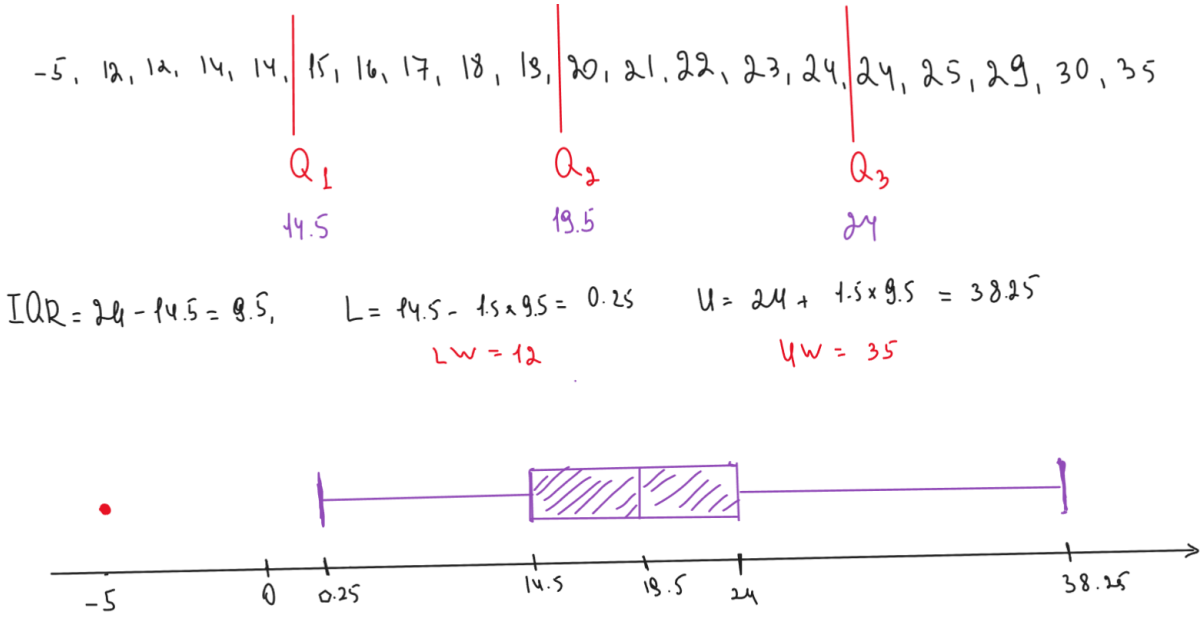
The area under a density plot is always equal to 1 because it represents the total **probability** of all possible outcomes in a distribution, as required by probability theory. This ensures that the **probability density function (PDF)** integrates to 1, signifying that some outcome must occur. When comparing distributions with different sample sizes, this property ensures that **relative shapes** are comparable, but **differences in scale** (due to sample size) should be considered when interpreting the density.

Part 2: *Hand-Drawn Graphs*

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

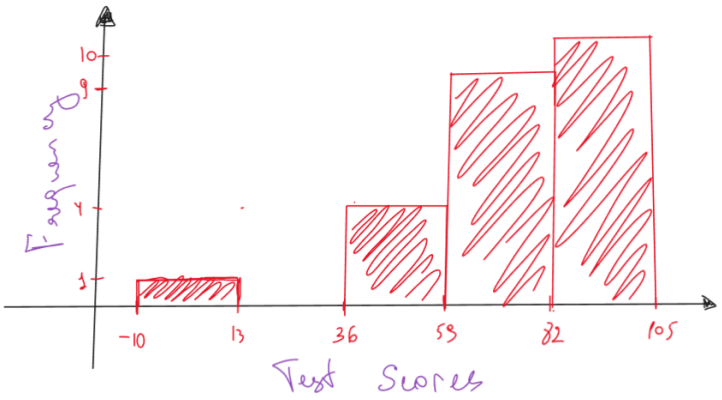


2. Given the dataset: -5, 12, 14, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 24, 25, 29, 30, 35, create a boxplot. Indicate the median, quartiles, and any potential outliers.



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.

min = -10, max = 105, bins = 5, $\Rightarrow$ bin_width = $\frac{105-(-10)}{5} = 23$



bins	freq.
[-10; 13)	1
[13; 36)	0
[36; 59)	4
[59; 82)	9
[82; 105]	10