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?

The whiskers in a boxplot indicate the expected variation of the data. They extend 1.5 times the Interquartile Range (IQR) from the top and bottom of the box. If the data do not reach the ends of the whiskers, then the whiskers will extend to the minimum and maximum data values.

The IQR is calculated as the difference between the third quartile (Q3) and the first quartile (Q1), which corresponds to the differences between the 75th and 25th percentiles. Typically, the whiskers extend to the most extreme data points within the range of Q1 - 1.5×IQR (for the lower whisker) and Q3 + 1.5×IQR (for the upper whisker). For instance, if a dataset has Q1 = 10 and Q3 = 20, the whiskers would span from 5 to 25.

However, the standard IQR-based rule has limitations, especially in non-normal distributions. This is particularly true for skewed data or data with extreme values. For example, in right-skewed distributions, many high values may be incorrectly classified as outliers, as seen in the case of house prices in a city where luxury homes greatly exceed average values. To overcome these limitations, adjusted boxplots or robust measures like the Median Absolute Deviation (MAD) can be useful.

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 uses the interquartile range (IQR) to identify outliers. However, in datasets with heavy skewness and multiple peaks, this method can misrepresent outliers. For instance, in skewed data, many points on the longer tail may be labeled as outliers even though they are part of the distribution. Alternative methods for identifying outliers include the Modified Z-Score, which uses the median and median absolute deviation (MAD) and is less affected by skewness. Another useful approach is Kernel Density Estimation (KDE), as it visualizes the entire distribution and helps identify outliers in low-density regions, particularly in multimodal datasets.

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 (average) is calculated by adding all values and dividing by the number of observations. This method can be significantly influenced by a few extremely large or small values in a skewed distribution. The median, on the other hand, is the middle value when all data points are arranged from smallest to largest. This characteristic makes the median more robust, as extreme values on one side do not affect it as dramatically. For example, in a skewed salary dataset, one very high income might inflate the mean while leaving the median mostly unchanged. Therefore, the median often serves as a better indicator of a "typical" salary.

A boxplot illustrates the median as a central line, allowing us to quickly identify the "middle" of the data. Because the median is not easily influenced by outliers, it provides a stable representation of where most values fall in skewed distributions. However, there are situations where relying solely on the median can obscure important data characteristics:

1. Multiple Peaks: If the dataset contains clusters of data (for instance, one group clustered around a lower value and another around a higher value), the boxplot may conceal this "bimodality" because it only displays one median line.

2. Extreme Outliers' Influence: While the median minimizes the effect of outliers, some of these outliers may be significant (such as extremely high incomes in a salary dataset). A boxplot's emphasis on the median may overlook how these outliers impact the overall average.

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?

Right-skewed distributions indicate that the majority of the data is located on the left side of the graph, while the mean (average) is greater than the median. When a boxplot shows significant right skewness, it suggests that the underlying probability distribution is positively skewed. This means most of the data is concentrated on the left side, with a long tail extending to the right, indicating the presence of extreme values that are greater than the median.

Due to these extreme values, the variance is typically high, contributing to an overall wider spread of the data. The skewness coefficient will be positive, reflecting the asymmetry and the longer tail on the right. Such skewness can violate important model assumptions, such as normality, which is particularly crucial for techniques like linear regression. To address these

issues, transformations such as logarithmic or square root transformations may be applied to reduce skewness and help ensure that statistical assumptions are met.

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?

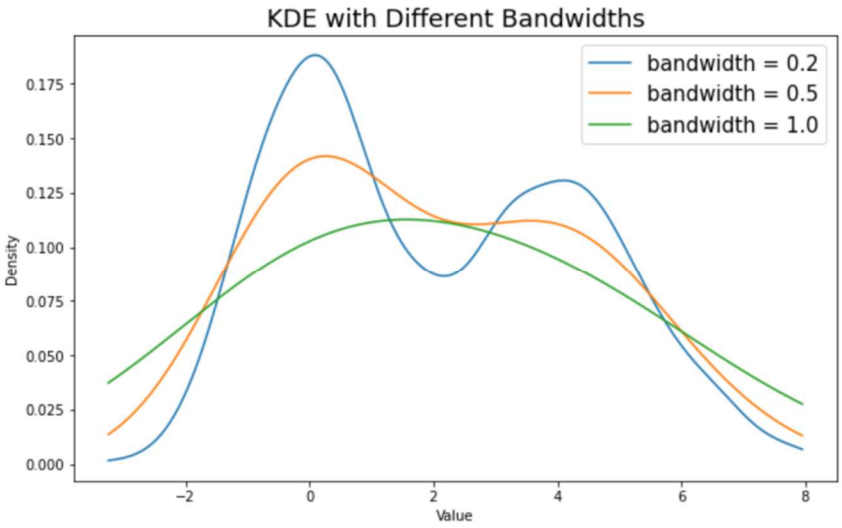
Box plots are valuable tools for comparing the distributions of multiple datasets or groups. They offer a clear visual summary of central tendency, spread, and outliers, which makes it easy to identify patterns and differences. This is particularly useful in high-dimensional data analysis, where multiple groups or variables need to be compared at the same time. Box plots enable a quick visual assessment of variations across groups without becoming cluttered.

However, box plots do have some limitations. When distributions overlap significantly, the box plots might look very similar, making it challenging to distinguish between groups. Furthermore, with categorical variables that have small sample sizes (fewer than 20 data points), box plots may not accurately reflect the underlying distribution. In such cases, the limited data can lead to misleading conclusions about variability, central tendency, and potential outliers.

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)?

Selecting an inappropriate number of bins in a histogram can significantly impact data interpretation. Too few bins may oversimplify the dataset, obscuring important details such as multimodal distributions or variations in density. Conversely, too many bins can introduce noise, making it difficult to identify meaningful patterns due to excessive fragmentation. In datasets with varying density regions, improper bin selection can either mask local density changes or exaggerate minor fluctuations, leading to misleading conclusions.

KDE's bandwidth parameter requires careful consideration. This parameter determines how much each individual Gaussian curve spreads out, which in turn affects the smoothness of our final estimate.



- With a small bandwidth (0.2, blue line), we see very detailed features but possibly over-fitting to our sample
- With a medium bandwidth (0.5, orange line), we maintain the major features while smoothing out some noise
- With a large bandwidth (1.0, green line), we get a very smooth curve that clearly shows the bimodal nature of our data but might be over-smoothing some important features

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?

A bar graph displays different categories using separate bars, while a histogram illustrates the distribution of continuous data by grouping values into intervals. Bar graphs are designed for categorical data, where each bar represents a distinct category, and the height of the bar indicates its frequency or proportion. In contrast, histograms are intended for numerical data; the bars represent intervals (or bins) along a continuous scale, and the area of each bar reflects the frequency of data points within that interval.

Choosing the right bin width is crucial in histograms because different widths can affect how the data distribution appears, which in turn influences the interpretation of patterns such as skewness or modality. However, bin width is not a concern in bar charts since the bars represent fixed categories rather than intervals, and the spacing between bars is usually equal and unrelated to the data values.

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.

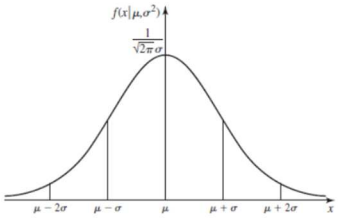
Histograms can distort data distributions if the bin sizes are not chosen correctly. Using overly broad bins may hide important details, while excessively narrow bins can introduce noise. For instance, when visualizing a dataset of individual heights, wide bins might make the distribution appear uniform, while narrow bins could reveal multiple peaks, leading to different interpretations. Alternative visualizations, such as Kernel Density Estimation (KDE) or violin plots, provide smoother and more continuous representations of data. These methods reduce the reliance on binning choices and offer clearer insights into the underlying 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?

A histogram displays information by dividing data points into fixed ranges or bins, while a density plot offers a more accurate estimate of the probability density function. Selecting the right kernel, such as Gaussian, and determining the appropriate bandwidth can be challenging, especially with sparse datasets. A small bandwidth may lead to overfitting, resulting in too many peaks, while a large bandwidth can overly smooth the distribution, obscuring important details.

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 entire area under the curve for every probability density function (PDF) is always equal to 1. This is so because the** likelihood** of all events is represented by the **area **under the curve, and a probability distribution requires that the total of all probabilities be equal to 1. A density plot represents the probability distribution of a continuous random variable, where the total probability must sum (or integrate) to 1. Graphically, this means that if you take the integral (area) under the density curve from $-\infty$ to $+\infty$, it adds up to 1. For example, the bell-shaped curve of the Normal distribution always encloses an area of exactly 1, ensuring that all possible outcomes together account for 100% probability.



$$Let, z = \frac{x-\mu}{\sigma\sqrt{2}} ; dx = \sigma\sqrt{2}\ dz$$

$$\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)^2} dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} e^{-z^2} dz$$

Evaluating

$$\int e^{-x^2}\, dx,$$

$$\int_0^{\infty} e^{-x^2}\, dx$$

$$= \frac{1}{2}\int_0^{\infty} e^{-t} t^{\left(\frac{1}{2}-1\right)}\, dt \text{ [substituting } x^2 = t \text{]}$$

$$= \frac{1}{2}\Gamma\left(\frac{1}{2}\right)$$

$$= \frac{\sqrt{\pi}}{2} \text{ [Using } \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi} \text{]}$$

Therefore

$$\int_{-\infty}^{\infty} e^{-x^2}\, dx = \sqrt{\pi}$$

$$So\ , \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)^2} dx = 1$$

... (more)