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 IQRbased rule in such cases?

In a boxplot, the whiskers show the spread of most of the data, and they are usually set using the interquartile range which is the difference between Q3 and Q1. The whiskers refer to the smallest and largest values within 1.5 times the IQR from the first and third quartiles. Any values out of this range are usually outliers and are shown as individual points. However, in skewed or heavy-tailed data, this rule may not work well because it can either miss real extreme values or mark too many points 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?

Boxplots assume a somewhat symmetrical distribution. When data is very skewed a boxplot can show too many or too few outliers which can in fact be misleading. For example, in a right-skewed distribution, many high values could be incorrectly seen as outliers. One alternative is density plots for example.

3. Explain the conceptual difference between median and mean in the context of nonsymmetric 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, but the median is more stable because it represents the middle value. A boxplot prioritizes the median since it gives a better sense of the dataset's center when we have skewness. However, this can hide important patterns like if a dataset has a long right tail, the median may look normal while the mean is actually much higher.

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?

If a boxplot is strongly right-skewed, it means the data has a lot of smaller values but not much of very large ones (like the dataset of salaries). This increases variance and makes the skewness coefficient positive. And if we are using a statistical model that assumes normality, for example a t-test, it might not work well.

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 a good option for quickly comparing multiple groups, like test scores from different schools, because they show medians and spreads side by side. But they struggle if the groups overlap a lot, as it becomes really hard to tell differences. Also, if a group has a very small

sample size, a boxplot might not be reliable as it might just show randomness instead of a real pattern.

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 bins are too wide, the histogram might miss important details, like multiple peaks in a dataset. If they are too narrow then histogram looks too messy. The same happens with KDE, like if the bandwidth is too small, the curve is too big, and it excludes key details. For example, if we track people's heights but use huge bins (like 10-15 cm each), we will miss the real distribution.

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 group numeric data into bins and show frequency within a range. A bar chart, on the other hand, is for categories and there's no binning involved. That’s why bin choice doesn’t matter in bar charts but is very important in histograms.

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 hide important patterns if the bins are too wide or too narrow. For example, if we take income data and the bins are too wide, the rich and poor might blend into the same group and we couldn't see the inequality. To fix this we can simply use KDE plots.

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 counts values in bins, and a density plot creates a smooth curve to estimate the probability of values. The challenge is to pick the right kernel function and bandwidth. If the bandwidth is too small, the plot looks too spiky and if too big, it smooths key details. Sparse datasets make this harder because there's not enough data to estimate density well.

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 total area under a density plot is always 1 because it represents a probability distribution, meaning that the sum of all possible values should be 100%. This property helps when

comparing datasets of different sizes since we're looking at relative rather than absolute frequency. It also means that the density curve reflects probabilities and not raw counts.