Data Visualiztion

Homework 2

David Aslanyan

Theoretical Questions.

1. On a boxplot, whiskers go up to 1.5 times the interquartile range (IQR) on each side from the first and third quartile. In normal distributions with outlier values, such a rule won't be exactly the representation of data since, under such data distributions, either the rule doesn't capture every possible outlier value or it defines the valid points of data as an outlier value. Other forms of methods like modified boxplots or robust measures in statistics overcome such a condition.
2. In skewed and multimodal distributions, boxplots may misrepresent outliers because the IQR method assumes a roughly normal distribution. If the data is heavily skewed, the upper whisker might extend further than expected, and multimodal distributions might not be well-represented. Alternatives include violin plots or kernel density estimates, which better visualize the distribution shape.
3. The mean is sensitive to extreme values, whereas the median is a more robust measure of central tendency. Boxplots prioritize the median because it provides a stable measure of the dataset’s center. However, relying solely on the median can obscure critical features such as long tails or bimodal distributions, which require additional visualization techniques.
4. A right-skewed boxplot suggests that the underlying distribution has a longer tail on the right side, often indicating positive skewness. This skewness can increase variance and impact statistical models, making parametric assumptions like normality unreliable. Right-skewness may also indicate the presence of extreme values or an underlying exponential-like distribution.
5. Boxplots are useful for comparing multiple groups, particularly in high-dimensional datasets, as they summarize distributions efficiently. However, they have limitations when distributions overlap significantly or when sample sizes are small, which can lead to misleading conclusions. Alternatives like violin plots or swarm plots can help visualize more details.
6. Selecting too few bins may oversimplify the data, hiding important patterns, while too many bins can create excessive noise. This issue is critical in varying density regions or multimodal distributions, where inappropriate bin selection can obscure trends. In kernel density estimation (KDE), the choice of bandwidth has a similar effect—too large a bandwidth smooths details, while too small creates spurious peaks.
7. Histograms represent continuous data and use bins to group values, whereas bar charts represent categorical data, where each bar corresponds to a distinct category. Bin choice is crucial in histograms because it determines how data is grouped, while in bar charts, the bar width is arbitrary and has no impact on interpretation.
8. Histograms can mislead if the bin width is poorly chosen. For example, if bin edges do not align with natural data groupings, it can create artificial peaks or gaps. A dataset with age groups might appear different if bins are not aligned with these groups. KDE or violin plots provide a smoother visualization that avoids such issues.
9. Density plots use kernel smoothing to estimate the probability density function of a dataset, unlike histograms, which count frequencies within bins. The choice of kernel and bandwidth affects density estimation, especially in sparse datasets, where inappropriate settings can introduce bias or excessive variance.
10. The total area under a density plot equals 1, ensuring it represents a valid probability distribution. This property enables comparisons across datasets of different sample sizes, as densities remain comparable despite differences in absolute frequency counts.