ECE M148 Homework 1

Sahiti Gabrani

Spring 2023

Question 1

Consider the following data set $A = 1, 1, 5, 9, 9$. What are the mean and median of A? Now, consider $B = 1, 1, 5, 9, 9, 11$. What are the mean and median of B? Using the mean and median, compare A and B.

Total elements in $A = 5$
Total elements in $B = 6$
Both A and B are arranged in ascending order.

Mean of $A = \frac{1+1+5+9+9}{5} = \frac{25}{5} = 5$

Median of $A = 5$

Mean of $B = \frac{1+1+5+9+9+11}{6} = \frac{36}{6} = 6$

Median of $B = \frac{5+9}{2} = 7$

Comparison

The mean of B is greater than the mean of A and the median of B is greater than the median of A. This suggests that a value (or some values) in B are, on average, higher than A. However, the difference in means or medians could be driven by outliers or extreme values in one list but not the other. In addition, the variability or spread of the data in each list could be different, which could impact how the data should be interpreted.

Question 2

In class, we discussed different ways to sample data. Explain in 1-2 sentences each the advantages and disadvantages of:
(a) Random sampling
(b) Stratified sampling
(c) Systematic sampling
(d) Cluster sampling

(a) Random sampling

Advantages

Each member of the population has an equal chance of being selected, and it reduces the risk of bias.

Disadvantages

It may not be representative of the population if the sample size is small, and it can be difficult to ensure that the sample truly represents the population.

(b) Stratified sampling

Advantages

It ensures that the sample is representative of the population by dividing the population into homogeneous subgroups, and it can increase the precision of estimates.

Disadvantages

It can be more time-consuming and expensive than random sampling, and it requires prior knowledge of the population to create appropriate strata.

(c) Systematic sampling

Advantages

It's a simple and quick method that can be applied to large populations, and it can provide a representative sample if the population is uniformly distributed.

Disadvantages

It can introduce bias if there is a pattern in the population that's not reflected in the sampling interval, and it's vulnerable to periodicity in the population.

(d) Cluster sampling

Advantages

It can be more efficient than other sampling methods if the population is naturally divided into clusters, and it's easier to implement in the field.

Disadvantages

It can introduce bias if the clusters are not representative of the population, and it may require a larger sample size than other methods to achieve the same level of precision.

Question 3

As discussed in class, many real-world datasets will contain missing or null values in the data. List four different strategies you could reasonably use to address null values. For each, clarify what the advantages and disadvantages to it are.

1. Delete the data

Definition: In this approach, the rows or columns containing missing values are deleted from the dataset. This method can be applied when the number of missing values is small compared to the size of the dataset, or when the missing values do not have much impact on the analysis.
Advantage: The advantage of this method is that it is straightforward to implement and can reduce the complexity of the analysis.
Disadvantage: However, the disadvantage is that it may result in a loss of information, especially if the missing values are not missing at random. Additionally, it may reduce the statistical power of the analysis.

2. Map to the closest point using some feature

Definition: In this approach, the missing values are replaced with the values from the nearest neighbor in the feature space. This method can preserve the local structure of the data and can work well for continuous features with a smooth distribution.
Advantage: The advantage is that it can be effective for high-dimensional datasets where other imputation methods may not work well.
Disadvantage: However, the disadvantage is that it may introduce bias if the feature space is not well-defined, and it may not work well for categorical or sparse features.

3. Populate using a random choice from the data set

Definition: In this approach, the missing values are replaced with randomly chosen values from the same feature in the dataset. This method can preserve the sample size and the distribution of the data, and can work well for features with a uniform or random distribution.
Advantage: The advantage is that it is easy to implement and can be effective for small datasets.
Disadvantage: However, the disadvantage is that it may introduce variability and noise into the analysis, and it may not work well for features with a skewed or non-random distribution.

3. Populate based on statistics such as mean, median, mode of other data points

Definition: In this approach, the missing values are replaced with the statistical summary (e.g., mean, median, mode) of the other non-missing values in the same feature. This method can preserve the distribution and central tendency of the data, and can work well for features with a well-defined central tendency.
Advantage: The advantage is that it is easy to implement and can be effective for continuous or categorical features.
Disadvantage: However, the disadvantage is that it may introduce bias if the feature distribution is not well-defined, and it may not work well for features with outliers or extreme values. Additionally, it may not work well if the missing values are clustered or patterned in a specific way.

Question 4

Consider the following sampling scenarios and determine which type of sampling bias is being demonstrated and explain your answer.
(a) Bob is a wealthy CEO who thinks taxes are too high. To confirm this hypothesis, he asks all his wealthy CEO friends their opinion.
(b) Sally is a teacher who wants to know how her class is performing. She sends out a survey with the following question: "Do you feel like you will get an A in the course or are you failing?"
(c) Constantine wants to know people's opinion about his website. He posts a survey link on his website asking for responses.
You may choose among the following options for the type of bias:
i) Response Bias
ii) Voluntary Bias
iii) Convenience Bias
iv) Under-coverage Bias
v) Over-coverage Bias
vi) Non-response bias

(a) This is an example of **over-coverage bias**. Bob is only sampling individuals who share his characteristics (wealthy CEOs), which can lead to an overrepresentation of this group's opinions and a biased result.

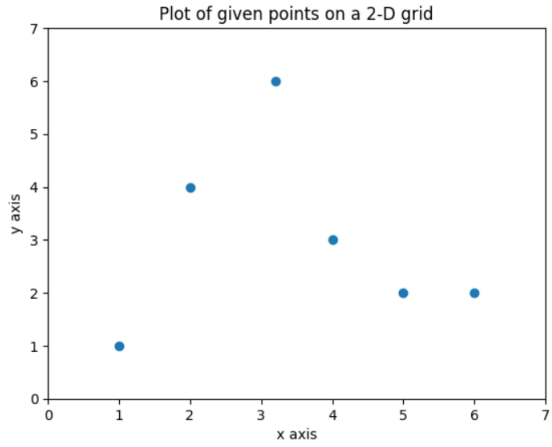
(b) This is an example of **response bias**. By asking students whether they feel like they will get an A or are failing, the survey question creates a dichotomy that may not accurately represent the distribution of grades. Additionally, students who are more confident may be more likely to respond, leading to a biased result.

(c) This is an example of **voluntary bias**. Since the survey is only posted on Constantine's website, only visitors to the website are able to respond, potentially leading to a non-representative sample. Individuals who are more likely to visit the website and respond may differ systematically from the general population, leading to a biased result.

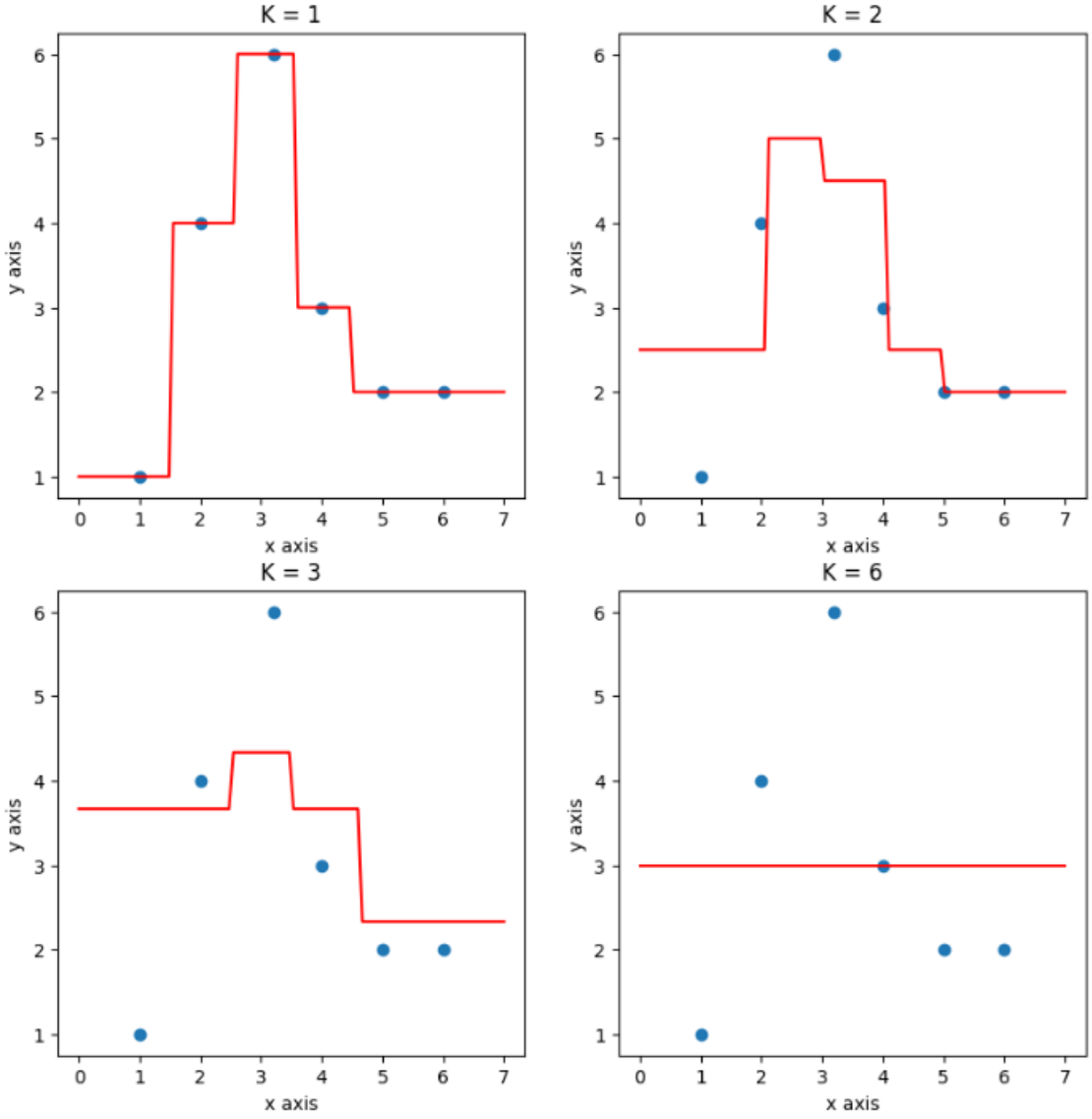
Question 5

Perform KNN Regression on the following data set for different values of K: $(x, y) = (1, 1), (2, 4), (3.2, 6), (4, 3), (5, 2), (6, 2)$**. Start by plotting the given points on a 2-D grid and then fitting a KNN regressor for the different values of K:**
Make sure to draw the regression plot from 0 **to** 7**.**
- $K = 1$
- $K = 2$
- $K = 3$
- $K = 6$

Contrast and compare your findings over various choices of K. Is a larger K always better? Is $K = 1$ **always better? Why or why not? Comment on what you think about the KNN performing regression on all** $x < 1$**.**



Plots obtained after fitting a KNN regressor for the different values of K:



.

Comparison

We can see that as the value of K increases, the regression line becomes smoother and less sensitive to individual data points. However, a larger K value may also result in underfitting, where the regression line fails to capture the true underlying relationship between x and y. In this case, we can see that $K = 1$ results in a poor fit, while $K = 6$ results in a straight line that fails to capture the curvature of the data. Overall, the choice of K depends on the specific dataset and the trade-off between bias and variance. .

KNN regression on $x < 1$.

Regarding the KNN regression on all $x < 1$, we cannot make any meaningful predictions for these values, as they fall outside the range of the input data. In practice, we would need to extrapolate beyond the range of the input data, which may lead to unreliable predictions. Therefore, we should be cautious when using KNN regression to make predictions outside the range of the input data.