Homework 3

Location Statistics: Understanding Central Tendencies

Location statistics refer to statistical measures that describe the central point or typical value in a dataset. These measures, often called measures of **central tendency**, provide valuable insights into where most of the data is concentrated. By summarizing data with a single value, location statistics help simplify the analysis of data, making it easier to understand and compare datasets.

Three of the most commonly used measures of central tendency are the **mean**, **median**, and **mode**. Each of these statistics offers a different perspective on the "center" of the data and is useful in various contexts.

1. Mean (Arithmetic Average)

The **mean** is the sum of all data points divided by the number of data points. Mathematically, for a dataset $x_1, x_2, \ldots, x_n$, the mean μ is given by:

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$$

The mean is widely used because it takes into account all data points and provides a balanced measure. However, it is sensitive to **outliers**, which are extreme values that can skew the mean and give a misleading picture of the central tendency.

2. Median

The **median** is the middle value in a dataset when the data points are arranged in ascending or descending order. If the number of data points is **odd**, the median is the **exact middle value**. If the number of data points is **even**, the median is the **average of the two middle values**.

The median is particularly useful in skewed distributions because it is **not affected by outliers**. For instance, in the case of income distribution, the median income often gives a better indication of what most people earn, as it is not distorted by a few extreme salaries.

3. Mode

The **mode** is the value that appears most frequently in a dataset. A dataset can have no mode (if all values are unique), one mode (unimodal), or multiple modes (bimodal, multimodal).

The mode is useful in identifying the most common value in categorical or discrete data.

Comparing the Measures

Each measure of central tendency provides unique insights into the data:

- The **mean** considers all values, making it ideal for symmetrical, well-behaved datasets.

- The **median** is more robust to outliers, making it a better indicator of central tendency in skewed distributions.
- The **mode** is particularly useful for categorical data, offering insight into the most frequent occurrence.

Find a Value of c that Minimize $S(c)$

Given a set of numbers $x_1, x_2, \ldots, x_n$, we want to **find a value c that minimizes the sum of squared distances** to each x_i. So the objective is to minimize:

$$S(c) = \sum_{i=1}^{n} (x_i - c)^2$$

1. Expand the sum:

We can rewrite the formula as:

$$S(c) = \sum_{i=1}^{n} (x_i - c)^2$$

Now we can expand the square:

$$S(c) = \sum_{i=1}^{n} (x_i^2 - 2x_i c + c^2)$$

That becomes:

$$S(c) = \sum_{i=1}^{n} x_i^2 - 2c \sum_{i=1}^{n} x_i + nc^2$$

2. Differentiate with respect to c :

To find an **inflection point**, we take the derivative of $S(c)$ with respect to c and set it equal to zero:

$$\frac{d}{dc} S(c) = -2 \sum_{i=1}^{n} x_i + 2nc = 0$$

Second Derivative:

Now, let's compute the second derivative of $S(c)$ to be ensure that we are finding a local minimum and not a local maximum:

$$\frac{d^2}{dc^2} S(c) = 2n$$

We observe that the second derivative is **positive** and constant. Since the second derivative is positive, we can conclude that the function $S(c)$ is **convex**, and therefore the point we found is a **local minimum**.

3. Solve for c:

Solving for c :

$$2nc = 2\sum_{i=1}^{n} x_i \Rightarrow c = \frac{\sum_{i=1}^{n} x_i}{n}$$

Thus, the value of c that minimizes the sum of squared distances is the **average** of the numbers $x_1, x_2, \ldots, x_n$.

$$c = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Poisson Distribution

The **Poisson distribution** is a probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space, provided these events occur with a known constant rate and are independent of the time since the last event.

It is commonly used to model random events that happen with a certain average rate

Probability Mass Function (PMF):

The probability of observing exactly k events in a given interval is given by the **Poisson probability mass function (PMF)**:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Where:

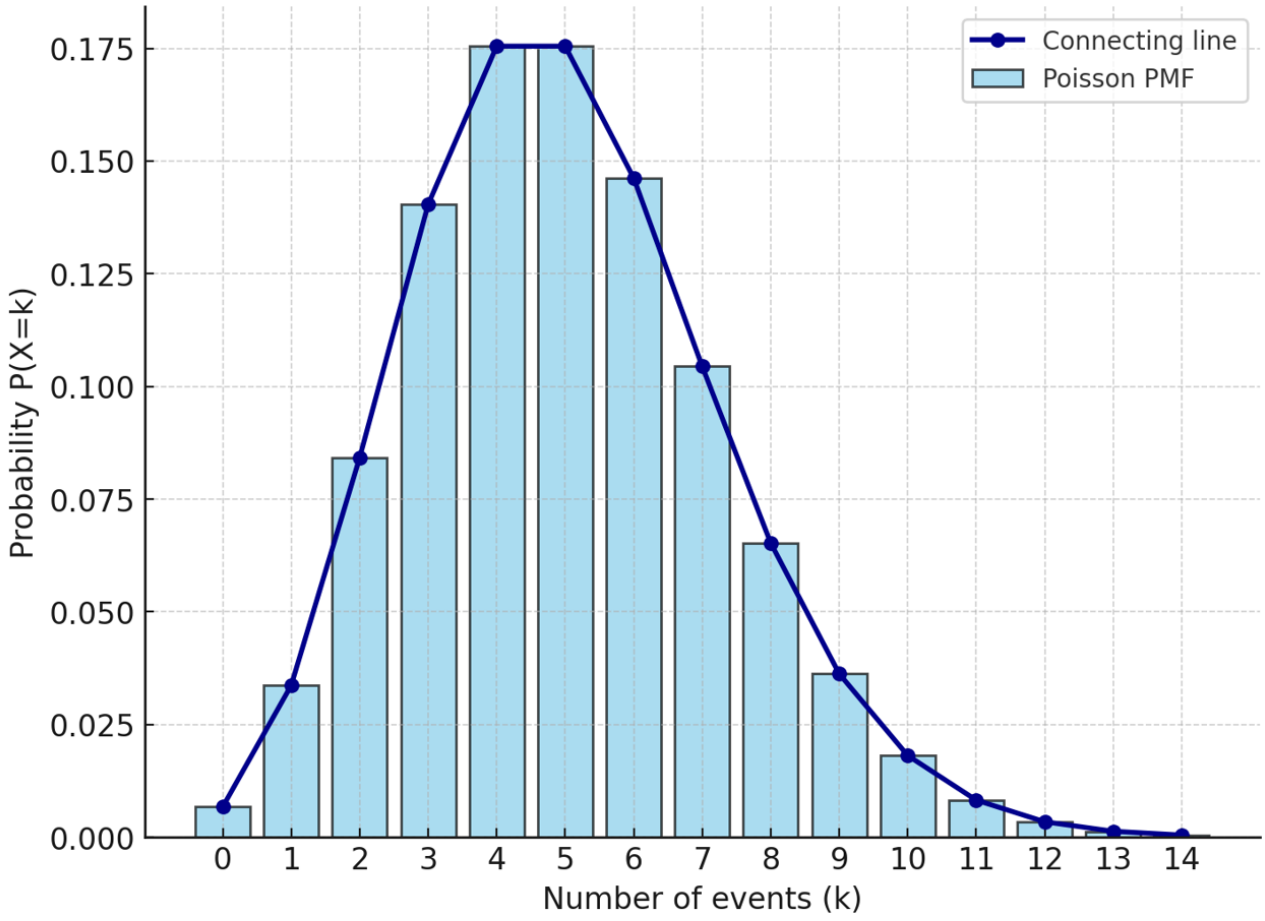
- λ: average number of events per interval (**rate of occurrence**).
- k: **actual number of events** (a non-negative integer).
- e: **Nepero's costant** that is the base of the natural logarithm (approximately 2.718).

Characteristics of the Poisson Distribution:

1. **Mean and Variance**:
- The mean of the distribution is λ.
- The variance is also λ.
2. **Assumptions**:
- Events occur independently.
- The average rate of events λ is constant over the observed period.
- Two events cannot happen at the exact same moment (they are **discrete**).

Visual Representation:

A Poisson distribution can be represented with a **discrete bar plot**, where the height of each bar represents the probability of observing a certain number of events.


Connecting line
Poisson PMF
Probability P(X=k)
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Number of events (k)