

AI & ML





Table of Contents

[**Statistics 2**](#_heading=h.v1t9at50aqht)

[**1. Descriptive Statistics 2**](#_heading=h.wvzqo22naoa6)

[**2. Categorical Variables 4**](#_heading=h.c0th239swy5e)

# Statistics

## 1. Descriptive Statistics

**Mean, Median, and Mode**

Let's calculate the mean, median, and mode for the list milk = [1, 2, 1, 2, 2, 1, 6, 1, 2, 2, 2, 1, 2, 10, 2, 2, 1, 2, 1, 2].

* **Mean:** The mean is the sum of all values divided by the number of values.

**Calculation:**

**Sum of values:** 1+2+1+2+2+1+6+1+2+2+2+1+2+10+2+2+1+2+1+2=461 + 2 + 1 + 2 + 2 + 1 + 6 + 1 + 2 + 2 + 2 + 1 + 2 + 10 + 2 + 2 + 1 + 2 + 1 + 2 = 461+2+1+2+2+1+6+1+2+2+2+1+2+10+2+2+1+2+1+2=46

**Number of values:** 202020

**Mean:** Mean=4620=2.3\text{Mean} = \frac{46}{20} = 2.3Mean=2046​=2.3

**Mean:** 2.3

* **Median:** The median is the middle value in a sorted list. If there’s an even number of values, the median is the average of the two middle numbers.

**Steps:**

**Sort the list:** [1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,2,2,6,10][1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 6, 10][1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,2,2,6,10]

**Find the middle values:** There are 20 values, so the median is the average of the 10th and 11th values. Median=2+22=2\text{Median} = \frac{2 + 2}{2} = 2Median=22+2​=2

**Median:** 2

* **Mode:** The mode is the value that appears most frequently in the dataset.

**Steps:**

**Identify the most frequent number:**

* + 1 appears 7 times.
  + 2 appears 10 times.
  + 6 appears 1 time.
  + 10 appears 1 time.

**Mode:** The number 2 appears the most frequently.

**Summary:**

* **Mean:** 2.3
* **Median**: 2
* **Mode**: 2

These calculations help describe the central tendency of the data.

Here's how you can calculate the mean, median, and mode for the list milk in Python:

import numpy as np

from scipy import stats

# Define the list

milk = [1, 2, 1, 2, 2, 1, 6, 1, 2, 2, 2, 1, 2, 10, 2, 2, 1, 2, 1, 2]

# Calculate the mean

mean\_value = np.mean(milk)

# Calculate the median

median\_value = np.median(milk)

# Calculate the mode

mode\_value = stats.mode(milk)[0][0]

# Display the results

print(f"Mean: {mean\_value}")

print(f"Median: {median\_value}")

print(f"Mode: {mode\_value}")

**Explanation:**

* **np.mean(milk):** This calculates the mean (average) of the list.
* **np.median(milk):** This calculates the median (middle value) of the list.
* **stats.mode(milk)[0][0]:** This calculates the mode (most frequent value) of the list. stats.mode() returns a ModeResult object, so [0][0] is used to extract the actual mode value.

**Output:** When you run this code, you should get the following output:

Mean: 2.3

Median: 2.0

Mode: 2

This matches our manual calculations from earlier.

## 2. Categorical Variables

Categorical variables, also known as qualitative variables, are variables that represent categories or groups rather than numerical values. These categories can be labels or names that distinguish different groups or levels within the data. Categorical variables are often used to represent data that can be divided into different groups or classes, such as gender, colors, or types of products.

**Types of Categorical Variables:**

1. **Nominal Variables:**
   * These are categorical variables with no inherent order or ranking among the categories.
   * **Example:**
     + Colors: Red, Green, Blue
     + Gender: Male, Female
     + Brands: Nike, Adidas, Puma
2. **Ordinal Variables:**
   * These are categorical variables with a meaningful order or ranking among the categories, but the intervals between the ranks are not necessarily equal.
   * **Example:**
     + Education Level: High School, Bachelor's, Master's, Ph.D.
     + Customer Satisfaction: Low, Medium, High