Statistics is the science of collecting, analyzing, interpreting, and presenting data. In data science, it provides the essential tools for understanding data, uncovering patterns, and making reliable decisions. Let's break down the core fundamentals you need to master:[^7_1][^7_2][^7_3]

Descriptive Statistics

Descriptive statistics are statistical methods used to organize, summarize, and present data in an informative way. They provide simple quantitative descriptions about the main characteristics of a dataset, such as measures of central tendency (mean, median, mode), measures of dispersion (range, variance, standard deviation), and the overall distribution or shape of the data.

Population vs. Sample

In data science, it's important to distinguish between a **population** and a **sample**:

- **Population:** This is the complete set of items or individuals you are interested in studying. In statistics, the population includes every member of the group you want to draw conclusions about. For example, if you want to analyze all customers of an online store, the population would be the entire customer base.
- **Sample:** Since it's often impractical or impossible to study an entire population, data scientists collect a **sample**—a smaller, manageable subset of the population. The sample should be representative of the population so that the findings can be generalized. For example, you might analyze a random selection of 1,000 customers from the store's database.

Why is this important in data science?
Samples allow us to estimate characteristics (parameters) of the population without the need to collect data from every individual. Proper sampling methods help ensure results are reliable and minimize bias.

Key point:
Statistical analyses are typically performed on samples, but the goal is usually to infer information about the broader population.

---

1. Measures of Central Tendency

Central tendency refers to statistical measures that identify a single value as representative of an entire dataset. These measures aim to provide an accurate summary of the typical, or "central," value in the data distribution. The three most common measures of central tendency are:

- **Mean:** The arithmetic average of all the data points.
- **Median:** The middle value when the data are sorted in order.
- **Mode:** The most frequently occurring value in the dataset.

Central tendency is fundamental for understanding summary patterns within data and is often the first step in data analysis to characterize a dataset's general behavior.

a. Mean (Average)

The **mean** (also known as the arithmetic average) is a measure of central tendency that represents the sum of all values in a dataset divided by the number of values. It provides a single value that summarizes the overall level of the data, making it useful for comparing different groups or understanding the "typical" value in a dataset. The mean is sensitive to outliers, as extremely large or small values can significantly affect its value.

Formula (Population):

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

Formula (Sample):

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

Interpretation: The arithmetic average of all data points.

b. Median

Median is a measure of central tendency that identifies the middle value in a dataset when the data points are arranged in ascending (or descending) order.

- How to calculate:
    1. the data in ascending order.
    2. If the number of data points (n) is odd, the median is the value at position $((n + 1)/2)$.
    3. If n is even, the median is the average of the two middle values at positions $(n/2)$ and $(n/2 + 1)$.
- The median divides the data into two equal halves, with 50% of the values below and 50% above (or equal to) the median.
- The median is robust to outliers and skewed data, making it useful for data distributions that are not symmetrical.
- The middle value when data is sorted.
- For odd number of points: middle value.
- For even number: average of two middle values.
- **Robust to outliers**

c. Mode

The most frequently occurring value.

- Continuous - sensitive to bins
- Categorical - highest frequency

$$ \text{Mode} = \text{arg,max}_x ; \text{frequency}(x) $$

- The most frequently occurring value in the dataset.
- Useful for categorical data.

d. Geometric Mean

The geometric mean is a measure of central tendency that is especially useful for data that are multiplicative in nature or span several orders of magnitude (e.g., rates of change, growth rates, financial returns).

- **Definition:** The geometric mean is the nth root of the product of all data points, where n is the number of points.
- **Best for:** Positively-skewed data, data involving ratios, percentages, or exponential growth.
- **Formula (Population or Sample):**

$$ \text{Geometric Mean} = \left(\prod_{i=1}^n x_i \right)^{1/n} $$

where x_i are the data points and n is the number of points.

- **Note:** All data points must be strictly positive for the geometric mean to be defined.

e. Harmonic Mean

The harmonic mean is another type of average that is useful when dealing with rates, ratios, or situations where the data is defined in terms of "per unit" value (e.g., speed, efficiency).

- **Definition:** The harmonic mean is the reciprocal of the arithmetic mean of the reciprocals of the data points.
- **Best for:** Averaging rates (e.g., speed: distance per unit time), ratios, or quantities where values must be aggregated in a reciprocal manner.
- **Formula (Population or Sample):**

  $$ \text{Harmonic Mean} = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}} $$

  where x_i are the data points and n is the number of points.

- **Note:** All data points must be non-zero for the harmonic mean to be defined.

Example Use Case:
If you want to compute the average speed over multiple trips with different speeds, the harmonic mean provides a more accurate average when the distance for each segment is the same.

Python Example:

```
import numpy as np

data = [4, 8, 9, 10, 6, 12, 14, 4, 5, 3, 4]

# Harmonic Mean
harmonic_mean = len(data) / sum(1/x for x in data)
print(f"Harmonic Mean: {harmonic_mean}")

# Using scipy:
```

```
# from scipy.stats import hmean
# harmonic_mean = hmean(data)
```

- **Interpretation:** The harmonic mean tends to be lower than the arithmetic mean, especially when the dataset contains large outlier values. It mitigates the influence of large numbers and gives greater weight to smaller values.

d. Comparison of Arithmetic, Geometric, and Harmonic Means

Mean Type	Formula	Best Use Cases	Sensitivity to Outliers	Interpretation & Notes
Arithmetic	$\displaystyle \overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$	General average, additive data, most common central tendency	**Highly sensitive to outliers**	Easy to compute/interpret; every value has equal weight.
Geometric	$\displaystyle \text{GM} = \left(\prod_{i=1}^n x_i\right)^{1/n}$	Growth rates, multiplicative processes, ratios, percentages	**Less sensitive to high outliers**; **More sensitive to low/zero values**	Best for percentages, ratios, normalized values, lognormal distributions.
Harmonic	$\displaystyle \text{HM} = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$	Rates ("per unit" e.g., speed), when averaging ratios or rates	**Highly sensitive to low outliers**	Mitigates influence of large values; emphasizes impact of smaller values.

Key Differences and Insights:

- **Use Cases**:
  - *Arithmetic mean* is most suitable for general, additive data (e.g., total income, total test scores).
  - *Geometric mean* is preferred when dealing with data that are products or rates of change (e.g., average growth rate, returns in finance).
  - *Harmonic mean* is ideal for averaging ratios or rates (e.g., speed, price per unit, when time or distance is constant).
- **Sensitivity to Outliers**:
  - *Arithmetic mean* can be heavily skewed by extreme high or low values.
  - *Geometric mean* reduces the influence of large outliers but is very sensitive to small values approaching zero (or zeros, which make it undefined).
  - *Harmonic mean* is the most affected by small (very low) values in the dataset; a single small value will drag the harmonic mean down significantly.
- **Interpretation**:

    - *Arithmetic mean* provides the "typical" value if every observation were the same.
    - *Geometric mean* shows the typical rate of change or normalized product per observation—often used when the data is multiplied together across periods or dimensions.
    - *Harmonic mean* highlights the aggregate rate, especially in contexts where reciprocals are meaningful (e.g., "per unit" calculations).

- **Further Notes**:

    - For any set of positive numbers:

$$ \text{Harmonic Mean} \leq \text{Geometric Mean} \leq \text{Arithmetic Mean} $$

```
Equality holds only when all data points are identical.
```

- The choice of mean can affect downstream analysis or business insights—using the inappropriate mean may misrepresent the data.

Summary Table:

Scenario	Use Arithmetic Mean	Use Geometric Mean	Use Harmonic Mean
Averaging simple survey scores	✓		
Percent growth or ratios		✓	
Averaging speeds (constant dist.)			✓
Data with large outliers	✗	✓	✓
Data includes zeros	✓	✗	✗

2. Measures of Dispersion (Spread)

a. Range

- **Definition:** The range is the simplest measure of dispersion; it is the difference between the largest and smallest values in a dataset.
- **Use Case:** Range is useful for quickly assessing the total spread or span of a dataset. It’s often used as an initial descriptive statistic to get a sense of variability, particularly when you want a fast, rough idea of how wide-ranging your data are. However, it is very sensitive to outliers and does not give information about the distribution of values between the extremes.
- Difference between maximum and minimum values.

$$ Range = x_{max} - x_{min} $$

b. Variance

- **Definition:** Variance is a statistical measure of dispersion that quantifies how much the values in a dataset differ from the mean (average) of that dataset. In essence, it measures the average squared deviation of each number from the mean.
- **How to Calculate:**
    1. Find the mean (average) of the dataset.
    2. Subtract the mean from each data point and square the result (this gives you the squared deviations).
    3. Take the average of those squared deviations:
        - If calculating **population variance**, divide by the number of data points (N).

    $$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

        - If calculating **sample variance**, divide by one less than the number of data points ($n-1$), which corrects bias in small samples.

    $$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \overline{x})^2$$

- **Use Case:**
  Variance is used:
    - To understand how spread out or clustered the data is around the mean.
    - In risk analysis, such as finance (portfolio volatility).
    - As a basis for standard deviation and in other statistical techniques, such as analysis of variance (ANOVA).
    - To compare the consistency (variability) across different datasets.
    - Large variance means data points are more spread out; small variance means data are closely clustered around the mean.

c. Standard Deviation

- **Definition:** Standard deviation is a measure of dispersion that tells you how much the data points in a dataset typically differ from the mean. It represents the average distance of each data point from the mean.
- **How to Calculate:**
    1. Find the mean (average) of the dataset.
    2. Subtract the mean from each data point and square the result (squared deviations).
    3. Calculate the variance—average of those squared deviations (population: divide by N, sample: divide by $n-1$).
    4. Take the square root of the variance.
- **Formula:**

  Population standard deviation:

  $$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}}$$

Sample standard deviation:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \overline{x})^2}$$

- **Why Square Root of Variance:**
  Taking the square root of the variance gives the standard deviation in the **same units** as the original data, making it directly interpretable. Variance is in squared units, which are less intuitive, while standard deviation restores the original scale.

- **Use Case:**

  - To understand how consistent or variable your data is.
  - Used in risk and volatility analysis in fields like finance and quality control.
  - Standard deviation is key in identifying outliers, constructing confidence intervals, and comparing the spread between different datasets.

- **What It Signifies:**
  A small standard deviation indicates that data points are close to the mean, suggesting little variability. A large standard deviation means that the data are spread out over a wider range of values. It's an essential summary statistic for understanding the distribution shape and expected variation within your data.

d. Interquartile Range (IQR)

Disadvantage of Range:

- The minimum and maximum are the most unreliable measures.
- What is the chance that you sample the extremes? There are very, very few values on the tails.
- Could be outliers (to be discussed).
- Safer to work with quartiles.

The solution is to rely on quartiles:

- Quartiles divide the data into four equal parts:

  - **First quartile (Q1):** 25th percentile

    $$Q_1 = \text{Value at } 25^\text{th} \text{ percentile}$$

  - **Second quartile (Q2, median):** 50th percentile

    $$Q_2 = \text{Value at } 50^\text{th} \text{ percentile}$$

  - **Third quartile (Q3):** 75th percentile

    $$Q_3 = \text{Value at } 75^\text{th} \text{ percentile}$$

Interquartile Range (IQR):

The interquartile range (IQR) is a measure of statistical dispersion and represents the range within which the central 50% of your data lie. It is calculated as the difference between the third quartile (Q3) and the first quartile (Q1):

$$ \text{IQR} = Q_3 - Q_1 $$

Quartiles are values that split your ordered data into four equal parts:

- **First quartile (Q1):** The value below which $25%$ of the data fall (25th percentile)
- **Second quartile (Q2 or median):** The value below which $50%$ of the data fall (50th percentile)
- **Third quartile (Q3):** The value below which $75%$ of the data fall (75th percentile)

The IQR is robust to outliers and provides a better sense of the spread in the center of the data than the total range, since it ignores the lowest $25%$ and highest $25%% of values and focuses on the variability of the middle half of the dataset.

Detect ouliers using IQR

To detect outliers using the Interquartile Range (IQR):

1. **Calculate Q1 and Q3**
    - Find the first quartile (Q1, 25th percentile) and third quartile (Q3, 75th percentile).
2. **Compute the IQR:**

    $$\text{IQR} = Q_3 - Q_1$$
3. **Determine outlier thresholds:**
    - **Lower bound:** $Q_1 - 1.5 \times \text{IQR}$
    - **Upper bound:** $Q_3 + 1.5 \times \text{IQR}$
4. **Flag outliers:**
    - Any data point < lower bound or > upper bound is considered an outlier.

Why?
The "1.5 × IQR" rule is a commonly used empirical threshold to identify extreme values that fall significantly below or above the central 50% of the data, which helps spot unusually distant values (potential outliers) while being robust to the influence of those very same values.

Example:
If $Q_1 = 10$ and $Q_3 = 20$, then $IQR = 10$

- Lower bound: $10 - (1.5 \times 10) = -5$
- Upper bound: $20 + (1.5 \times 10) = 35$
  Any value below -5 or above 35 is flagged as an outlier using this rule.

How Box Plots Work for Outlier Detection using IRQ: Box plots, also known as box-and-whisker plots, are graphical representations used to visualize the distribution of a dataset and identify potential outliers using the Interquartile Range (IQR) method.

- **The Box:**
  The central box represents the interquartile range (IQR), stretching from the first quartile (Q1) to the third quartile (Q3). This covers the middle $50%$ of the data.

- **The Median Line:**
  A horizontal line inside the box marks the median (Q2) of the data.
- **The "Whiskers":**
  - The whiskers extend from the edges of the box (Q1 and Q3) to the smallest and largest data points that are **not** considered outliers.
  - The end of the whisker is typically capped at the last data point within:
    - Lower whisker: $Q_1 - 1.5 \times \text{IQR}$
    - Upper whisker: $Q_3 + 1.5 \times \text{IQR}$
- **Outliers:**
  - Data points that fall outside the whiskers (below the lower bound or above the upper bound) are plotted individually, often as dots or stars.
  - These represent values flagged as outliers by the 1.5 × IQR rule.

Summary Table:

Element	Describes
Box	Q1 to Q3 (IQR: middle 50% of data)
Median line	The median (Q2, 50th percentile)
Whiskers	Data within [Q1 - 1.5×IQR, Q3 + 1.5×IQR]
Outliers	Data outside whiskers (potential outliers)

Visualization Example:

```
|----*-------|-------[=====|====|=====]-------|--------*----|
    Outlier  min     Q1   Median  Q3      max     Outlier

<---- Whisker ---->[==== BOX (IQR) ====]<---- Whisker ---->
```

Takeaway:
Box plots make it easy to visually detect outliers using the IQR rule, as potential outliers are displayed as points outside the whiskers, quickly highlighting unusual values in the dataset.

3. Moments of Distribution - Shape of Distribution

Moments of Distribution

In statistics, **moments** are quantitative measures related to the shape of a distribution. They provide significant information about its characteristics such as central tendency, spread, asymmetry, and peakedness.

What are Moments?

Moments are calculated with respect to the mean (central moments) or with respect to the origin (raw moments). The r-th moment about the mean for a random variable X is defined as:

$$\mu_r = E[(X - \mu)^r]$$

where E is the expectation operator, μ is the mean of X, and r is the order of the moment.

Main Types and Their Uses

Moment (Order)	Name	Formula	Use/Interpretation
1st Moment	Mean (Average)	$\mu_1 = E[X]$	Central tendency (location)
2nd Central Moment	Variance	$\mu_2 = E[(X - \mu)^2]$	Spread or dispersion
3rd Central Moment	Skewness	$\gamma_1 = \frac{E[(X - \mu)^3]}{\sigma^3}$	Asymmetry of distribution
4th Central Moment	Kurtosis	$\gamma_2 = \frac{E[(X - \mu)^4]}{\sigma^4}$	Peakedness ("tailedness")

- **First Moment (Mean):**
  - Indicates the average or center of the data.
  - Formula: $\displaystyle \mu = \frac{1}{n} \sum_{i=1}^n x_i$
- **Second Moment (Variance):**
  - Measures variability/spread from the mean.
  - Formula:

$$\displaystyle \sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

- **Third Moment (Skewness):**
  - Measures the degree of asymmetry of the distribution.
  - Formula:

$$\displaystyle \text{Skewness} = \frac{E[(X - \mu)^3]}{\sigma^3}$$

  - Interpretation: Positive value (right-skewed), Negative value (left-skewed)
- **Fourth Moment (Kurtosis):**
  - Measures the heaviness of tails and sharpness of peak.
  - Formula:

$$\displaystyle \text{Kurtosis} = \frac{E[(X - \mu)^4]}{\sigma^4}$$

  - Interpretation: High value (leptokurtic, heavy tails), Low value (platykurtic, light tails)

Use Cases

- **Mean:** Understand the typical/central value.
- **Variance:** Gauge spread—risk & volatility estimation in finance.
- **Skewness:** Detect tendency for extreme values on one side (e.g., outlier-prone data).
- **Kurtosis:** Reveal if outliers/extreme values (heavy tails) are more/less likely than for a normal distribution (important in fields like finance, quality control, etc.).

Summary Table:

Moment	What it Measures	Common Uses
Mean	Center of data	Averaging, summarizing data
Variance	Spread of data	Variability, risk assessment
Skewness	Asymmetry	Detecting outlier-prone distributions
Kurtosis	Tails/peakedness	Detecting propensity for outliers

> **Moments provide a foundation for understanding dataset characteristics and choosing appropriate statistical methods.**

4. Python Code Examples (Using pandas and numpy)

```
import numpy as np
import pandas as pd

# Sample data
data = [4, 8, 9, 10, 6, 12, 14, 4, 5, 3, 4]

# Convert to pandas Series
series = pd.Series(data)

# Mean
mean = series.mean()
print(f"Mean: {mean}")

# Median
median = series.median()
print(f"Median: {median}")

# Mode
mode = series.mode().tolist()  # mode() returns Series
print(f"Mode(s): {mode}")

# Range
data_range = series.max() - series.min()
print(f"Range: {data_range}")

# Variance (sample)
variance = series.var()
print(f"Variance: {variance}")
```

```python
# Standard Deviation (sample)
std_dev = series.std()
print(f"Standard Deviation: {std_dev}")

# Interquartile Range (IQR)
Q1 = series.quantile(0.25)
Q3 = series.quantile(0.75)
IQR = Q3 - Q1
print(f"IQR: {IQR}")

# Skewness
skewness = series.skew()
print(f"Skewness: {skewness}")

# Kurtosis
kurtosis = series.kurtosis()
print(f"Kurtosis: {kurtosis}")
```

---

5. Summary

- **Mean** is sensitive to outliers; **median** is robust.
- **Variance** and **standard deviation** quantify spread.
- **IQR** is useful for understanding spread without influence of outliers.
- **Skewness** and **kurtosis** describe distribution shape.
