# Mathematical Foundations of Statistics in Python

Statistics with Python Course

February 10, 2026

## Contents

# 1 Summary Statistics

Given a dataset $X = \{x_1, x_2, \ldots, x_n\}$ with $n$ observations, summary statistics provide numerical measures that describe the main features of the data.

## 1.1 Measures of Central Tendency

Central tendency measures identify a single value that represents the "center" or typical value of a distribution.

### 1.1.1 Arithmetic Mean

**Definition 1.1** (Arithmetic Mean)**.** The **arithmetic mean** (or simply mean) is the sum of all observations divided by the number of observations:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

**Properties:**

- The sum of deviations from the mean is zero: $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$
- Minimizes the sum of squared deviations: $\bar{x} = \arg\min_\mu \sum_{i=1}^{n}(x_i - \mu)^2$
- Sensitive to outliers
- For population data, denoted $\mu = \mathbb{E}[X]$

**Example 1.1.** For $X = \{2, 4, 6, 8, 10\}$:

$$\bar{x} = \frac{2 + 4 + 6 + 8 + 10}{5} = \frac{30}{5} = 6$$

### 1.1.2 Geometric Mean

**Definition 1.2** (Geometric Mean)**.** The **geometric mean** is the $n$-th root of the product of all observations:

$$\bar{x}_g = \left(\prod_{i=1}^{n} x_i\right)^{1/n} = \sqrt[n]{x_1 \cdot x_2 \cdots x_n}$$

Equivalently, using logarithms:

$$\log(\bar{x}_g) = \frac{1}{n}\sum_{i=1}^{n} \log(x_i)$$

**Properties:**

- Only defined for positive values ($x_i > 0$)
- Always less than or equal to the arithmetic mean: $\bar{x}_g \leq \bar{x}$ (AM-GM inequality)
- Appropriate for multiplicative processes (e.g., growth rates, ratios)
- Less sensitive to extreme values than arithmetic mean

**Example 1.2.** For growth rates $r = \{1.10, 1.20, 0.90\}$ (10%, 20%, -10%):

$$\bar{r}_g = \sqrt[3]{1.10 \times 1.20 \times 0.90} = \sqrt[3]{1.188} \approx 1.059$$

Average growth rate $\approx 5.9\%$

### 1.1.3 Median

**Definition 1.3** (Median)**.** The **median** is the middle value when observations are ordered from smallest to largest:

$$\text{Median} = \begin{cases} x_{(k+1)} & \text{if } n = 2k+1 \text{ (odd)} \\ \dfrac{x_{(k)} + x_{(k+1)}}{2} & \text{if } n = 2k \text{ (even)} \end{cases}$$

where $x_{(i)}$ denotes the $i$-th order statistic (the $i$-th smallest value).

**Properties:**

- Minimizes the sum of absolute deviations: $\text{Median} = \arg\min_\mu \sum_{i=1}^n |x_i - \mu|$
- Robust to outliers (resistant measure)
- The 50th percentile (divides data in half)

### 1.1.4 Mode

**Definition 1.4** (Mode)**.** The **mode** is the value that appears most frequently in the dataset:

$$\text{Mode} = \arg\max_x f(x)$$

where $f(x)$ is the frequency of value $x$.

**Properties:**

- Can be used with categorical data
- May not exist (uniform distribution) or may not be unique (multimodal)
- Unimodal: one mode; Bimodal: two modes; Multimodal: multiple modes

## 1.2 Measures of Spread (Dispersion)

Dispersion measures quantify the variability or spread of the data around the central tendency.

### 1.2.1 Range

**Definition 1.5** (Range)**.** The **range** is the difference between the maximum and minimum values:

$$\text{Range} = x_{(n)} - x_{(1)} = \max(X) - \min(X)$$

**Properties:**

- Simple but highly sensitive to outliers
- Only considers two data points
- Increases with sample size

### 1.2.2 Variance

**Definition 1.6** (Variance)**.** The **population variance** measures the average squared deviation from the mean:

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

The **sample variance** uses $n - 1$ in the denominator (Bessel's correction) for an unbiased estimator:

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

**Alternative computational formula:**

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

**Properties:**

- Always non-negative: $s^2 \geq 0$
- Equals zero only when all values are identical
- Units are squared (e.g., if data in meters, variance in meters$^2$)
- $\mathbb{E}[s^2] = \sigma^2$ (unbiased estimator)

### 1.2.3 Standard Deviation

**Definition 1.7** (Standard Deviation)**.** The **standard deviation** is the square root of the variance:

$$\sigma = \sqrt{\sigma^2} \quad \text{(population)}, \qquad s = \sqrt{s^2} \quad \text{(sample)}$$

**Properties:**

- Same units as the original data
- For normal distributions: approximately 68% of data within $\pm 1\sigma$, 95% within $\pm 2\sigma$
- Sample standard deviation $s$ is a biased estimator of $\sigma$

### 1.2.4 Percentiles and Quartiles

**Definition 1.8** (Percentile)**.** The ***p*-th percentile** ($P_p$) is a value below which $p$% of the observations fall:

$$P_p = x_{(\lceil n \cdot p/100 \rceil)}$$

More precisely, using linear interpolation:

$$P_p = (1 - g) \cdot x_{(j)} + g \cdot x_{(j+1)}$$

where $j = \lfloor (n-1) \cdot p/100 \rfloor$ and $g = (n-1) \cdot p/100 - j$

**Definition 1.9** (Quartiles)**.** **Quartiles** divide the ordered data into four equal parts:

$$\begin{aligned} Q_1 &= P_{25} \quad \text{(First quartile / 25th percentile)} \\ Q_2 &= P_{50} = \text{Median} \quad \text{(Second quartile)} \\ Q_3 &= P_{75} \quad \text{(Third quartile / 75th percentile)} \end{aligned}$$

### 1.2.5 Interquartile Range (IQR)

**Definition 1.10** (Interquartile Range)**.** The **interquartile range** is the difference between the third and first quartiles:

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

**Properties:**

- Contains the middle 50% of the data
- Robust to outliers
- Used to define outliers: values beyond $Q_1 - 1.5 \cdot \text{IQR}$ or $Q_3 + 1.5 \cdot \text{IQR}$

## 1.3 Measures of Shape

Shape measures describe the symmetry and tail behavior of distributions.

### 1.3.1 Skewness

**Definition 1.11** (Skewness)**. Skewness** measures the asymmetry of the distribution. The sample skewness (Fisher's definition):

$$\gamma_1 = \frac{m_3}{m_2^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}}$$

Adjusted sample skewness (for small samples):

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} \cdot \gamma_1$$

**Interpretation:**

- $\gamma_1 = 0$: Symmetric distribution
- $\gamma_1 > 0$: Right-skewed (positive skew) — long right tail, mean > median
- $\gamma_1 < 0$: Left-skewed (negative skew) — long left tail, mean < median

*Remark* 1.1. A rule of thumb: $|\gamma_1| > 1$ indicates substantial skewness.

### 1.3.2 Kurtosis

**Definition 1.12** (Kurtosis)**. Kurtosis** measures the "tailedness" of the distribution. The sample kurtosis:

$$\gamma_2 = \frac{m_4}{m_2^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2}$$

**Excess kurtosis** compares to the normal distribution:

$$\text{Excess Kurtosis} = \gamma_2 - 3$$

**Interpretation:**

- $\gamma_2 = 3$ (excess $= 0$): Mesokurtic (normal-like tails)
- $\gamma_2 > 3$ (excess $> 0$): Leptokurtic — heavy tails, more outliers
- $\gamma_2 < 3$ (excess $< 0$): Platykurtic — light tails, fewer outliers

*Remark* 1.2. Kurtosis is often misinterpreted as measuring "peakedness." It primarily measures tail weight and outlier propensity.

# 2 Correlation Metrics

Correlation metrics measure the strength and direction of relationships between variables. Given paired observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$:

## 2.1 Covariance

**Definition 2.1** (Covariance)**.** The **sample covariance** measures the joint variability of two variables:

$$\text{Cov}(X, Y) = s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

For populations:

$$\sigma_{XY} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)$$

**Properties:**

- $\text{Cov}(X, Y) > 0$: Positive relationship (both increase together)
- $\text{Cov}(X, Y) < 0$: Negative relationship (one increases as other decreases)
- $\text{Cov}(X, Y) = 0$: No linear relationship
- Symmetric: $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
- $\text{Cov}(X, X) = \text{Var}(X)$
- Scale-dependent (units are product of units of $X$ and $Y$)

**Computational formula:**

$$s_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}\right)$$

## 2.2 Pearson Correlation Coefficient

**Definition 2.2** (Pearson Correlation)**.** The **Pearson correlation coefficient** is the standardized covariance:

$$r = \frac{\text{Cov}(X, Y)}{s_X \cdot s_Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

For populations: $\rho = \dfrac{\sigma_{XY}}{\sigma_X \sigma_Y}$

**Properties:**

- Bounded: $-1 \leq r \leq 1$
- Dimensionless (scale-free)
- $r = 1$: Perfect positive linear relationship
- $r = -1$: Perfect negative linear relationship
- $r = 0$: No linear relationship (but may have nonlinear relationship!)
- Invariant under linear transformations: $\text{Corr}(aX + b, cY + d) = \text{sign}(ac) \cdot \text{Corr}(X, Y)$

**Coefficient of determination:**

$$R^2 = r^2$$

represents the proportion of variance in $Y$ explained by the linear relationship with $X$.

*Remark* 2.1. Pearson correlation measures **linear** relationships only. A correlation of zero does not imply independence—the variables may have a strong nonlinear relationship.

## 2.3 Spearman Rank Correlation

**Definition 2.3** (Spearman Correlation)**.** The **Spearman rank correlation** is the Pearson correlation applied to the ranks of the data:

$$r_s = \frac{\text{Cov}(R_X, R_Y)}{s_{R_X} \cdot s_{R_Y}}$$

where $R_X$ and $R_Y$ are the ranks of $X$ and $Y$.

When there are no tied ranks:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

where $d_i = R_{x_i} - R_{y_i}$ is the difference between ranks.

**Properties:**

- Bounded: $-1 \leq r_s \leq 1$
- Measures monotonic relationships (not just linear)
- Robust to outliers (uses ranks, not values)
- Appropriate for ordinal data
- $r_s = 1$: Perfect monotonically increasing relationship
- $r_s = -1$: Perfect monotonically decreasing relationship

**Example 2.1.** For data: $(1, 10), (2, 30), (3, 20), (4, 50), (5, 40)$

Ranks: $R_X = (1, 2, 3, 4, 5)$, $R_Y = (1, 3, 2, 5, 4)$

$d = (0, -1, 1, -1, 1)$, $\sum d^2 = 4$

$r_s = 1 - \frac{6\times 4}{5(25-1)} = 1 - \frac{24}{120} = 0.8$

## 2.4 Kendall's Tau Correlation

**Definition 2.4** (Kendall's Tau)**. Kendall's Tau ($\tau$)** measures ordinal association based on concordant and discordant pairs:

$$\tau = \frac{n_c - n_d}{\binom{n}{2}} = \frac{n_c - n_d}{\frac{n(n-1)}{2}}$$

where:

- $n_c$ = number of **concordant pairs**: $(x_i - x_j)(y_i - y_j) > 0$
- $n_d$ = number of **discordant pairs**: $(x_i - x_j)(y_i - y_j) < 0$

**With ties (Tau-b):**

$$\tau_b = \frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}}$$

where $n_0 = \frac{n(n-1)}{2}$, $n_1 = \sum_i \frac{t_i(t_i-1)}{2}$ (ties in $X$), $n_2 = \sum_j \frac{u_j(u_j-1)}{2}$ (ties in $Y$).

**Properties:**

- Bounded: $-1 \leq \tau \leq 1$
- More robust than Spearman for small samples
- Has a more intuitive probabilistic interpretation:

$$\tau = P(\text{concordant}) - P(\text{discordant})$$

- Generally $|\tau| < |r_s|$ for the same data

## 2.5 Contingency Tables for Categorical Data

**Definition 2.5** (Contingency Table)**.** A **contingency table** (cross-tabulation) displays the frequency distribution of categorical variables:

| | $Y_1$ | $Y_2$ | $\cdots$ | Total |
|---|---|---|---|---|
| $X_1$ | $n_{11}$ | $n_{12}$ | $\cdots$ | $n_{1\cdot}$ |
| $X_2$ | $n_{21}$ | $n_{22}$ | $\cdots$ | $n_{2\cdot}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| Total | $n_{\cdot 1}$ | $n_{\cdot 2}$ | $\cdots$ | $n$ |

where:

- $n_{ij}$ = observed frequency in cell $(i, j)$
- $n_{i\cdot} = \sum_j n_{ij}$ = row marginal
- $n_{\cdot j} = \sum_i n_{ij}$ = column marginal

**Normalizations:**

- **Row normalization:** $p_{j|i} = \dfrac{n_{ij}}{n_{i\cdot}}$ gives $P(Y = j \mid X = i)$
- **Column normalization:** $p_{i|j} = \dfrac{n_{ij}}{n_{\cdot j}}$ gives $P(X = i \mid Y = j)$
- **Total normalization:** $p_{ij} = \dfrac{n_{ij}}{n}$ gives joint probability

### 2.5.1 Chi-Square Test of Independence

**Definition 2.6** (Chi-Square Statistic)**.** The **chi-square statistic** tests whether two categorical variables are independent:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where:

- $O_{ij} = n_{ij}$ = observed frequency
- $E_{ij} = \dfrac{n_{i\cdot} \cdot n_{\cdot j}}{n}$ = expected frequency under independence

Under the null hypothesis of independence, $\chi^2 \sim \chi^2_{(r-1)(c-1)}$.

## 2.6 Cramér's V

**Definition 2.7** (Cramér's V)**. Cramér's V** is a normalized measure of association for categorical variables:

$$V = \sqrt{\frac{\chi^2}{n \cdot (k-1)}}$$

where:

- $\chi^2$ = chi-square statistic
- $n$ = total sample size
- $k = \min(r, c)$ = minimum of number of rows and columns

**Properties:**

- Bounded: $0 \leq V \leq 1$
- $V = 0$: Complete independence
- $V = 1$: Perfect association
- Symmetric: same value regardless of which variable is row/column
- For $2 \times 2$ tables, equals the absolute value of the phi coefficient: $V = |\phi|$

**Interpretation guidelines:**

| Cramér's V | Interpretation |
|---|---|
| 0.00 – 0.10 | Negligible association |
| 0.10 – 0.20 | Weak association |
| 0.20 – 0.40 | Moderate association |
| 0.40 – 0.60 | Relatively strong association |
| 0.60 – 0.80 | Strong association |
| 0.80 – 1.00 | Very strong association |

**Example 2.2.** Consider a $2 \times 2$ contingency table:

| | Improved | Not Improved | Total |
|---|---|---|---|
| Treatment | 80 | 20 | 100 |
| Control | 30 | 70 | 100 |
| Total | 110 | 90 | 200 |

Expected values under independence:

$$E_{11} = \frac{100 \times 110}{200} = 55, \quad E_{12} = \frac{100 \times 90}{200} = 45$$

Chi-square:

$$\chi^2 = \frac{(80-55)^2}{55} + \frac{(20-45)^2}{45} + \frac{(30-55)^2}{55} + \frac{(70-45)^2}{45} = 50.51$$

Cramér's V:

$$V = \sqrt{\frac{50.51}{200 \times 1}} = \sqrt{0.253} = 0.503$$

This indicates a moderately strong association between treatment and outcome.

## Summary: Choosing the Right Correlation Measure

| Measure | Data Type | Relationship Type |
|---|---|---|
| Pearson ($r$) | Continuous | Linear |
| Spearman ($r_s$) | Continuous/Ordinal | Monotonic |
| Kendall ($\tau$) | Continuous/Ordinal | Monotonic (small samples) |
| Cramér's V | Categorical | Any association |

**Key reminders:**

1. Correlation $\neq$ causation
2. Zero correlation $\neq$ independence (may have nonlinear relationships)
3. Always visualize data before interpreting correlation coefficients
4. Consider the nature of your data when choosing a correlation measure