Mathematical Foundations of Statistics in Python

Statistics with Python Course

February 12, 2026

Contents

1 Summary Statistics **3**

1.1 Measures of Central Tendency . . . 3

1.1.1 Arithmetic Mean . . . 3

1.1.2 Geometric Mean . . . 3

1.1.3 Median . . . 4

1.1.4 Mode . . . 4

1.2 Measures of Spread (Dispersion) . . . 4

1.2.1 Range . . . 4

1.2.2 Variance . . . 5

1.2.3 Standard Deviation . . . 5

1.2.4 Percentiles and Quartiles . . . 5

1.2.5 Interquartile Range (IQR) . . . 6

1.3 Measures of Shape . . . 6

1.3.1 Skewness . . . 6

1.3.2 Kurtosis . . . 6

2 Probability Distributions **7**

2.1 Discrete Distributions . . . 7

2.1.1 Bernoulli Distribution . . . 7

2.1.2 Binomial Distribution . . . 7

2.1.3 Poisson Distribution . . . 8

2.2 Continuous Distributions . . . 9

2.2.1 Normal (Gaussian) Distribution . . . 9

2.2.2 Student's t-Distribution . . . 10

2.2.3 Exponential Distribution . . . 10

2.2.4 Gamma Distribution . . . 11

2.2.5 Log-Normal Distribution . . . 12

2.2.6 Dirichlet Distribution . . . 13

2.3 Summary: Probability Distributions . . . 13

3 Point Estimation and Confidence Intervals **15**

3.1 Point Estimation . . . 15

3.1.1 Maximum Likelihood Estimation (MLE) . . . 15

3.1.2 Robust Estimators . . . 15

3.2 Comparison of Common Distribution Estimators . . . 17

3.3 Confidence Intervals . . . 17

3.3.1 Central Limit Theorem (CLT) . . . 17

3.4 Exact and Analytic Confidence Intervals . . . 17

3.4.1 Normal Distribution: Mean μ . . . 17

3.4.2 Normal Distribution: Variance σ^2 17
3.4.3 Binomial Proportion p 18
3.4.4 Poisson Rate λ 18
3.4.5 Exponential Rate λ 18
3.5 Summary: Confidence Interval Methods 19
3.5.1 When to Use Bootstrap 19

4 Correlation Metrics **21**
4.1 Covariance 21
4.2 Pearson Correlation Coefficient 21
4.3 Spearman Rank Correlation 22
4.4 Kendall's Tau Correlation 22
4.5 Contingency Tables for Categorical Data 23
4.5.1 Chi-Square Test of Independence 23
4.6 Cramér's V 24

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: 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 σ

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 Probability Distributions

Probability distributions describe how the values of a random variable are distributed. They are fundamental to statistical inference and modeling in life sciences.

2.1 Discrete Distributions

Discrete distributions describe random variables that can only take countable values (integers).

2.1.1 Bernoulli Distribution

Definition 2.1 (Bernoulli Distribution)**.** The **Bernoulli distribution** describes a single trial with two possible outcomes: success (1) or failure (0).

$$X \sim \text{Bernoulli}(p)$$

Probability mass function (PMF):

$$P(X = x) = p^x(1-p)^{1-x}, \quad x \in \{0, 1\}$$

Parameters:

- $p \in [0, 1]$: probability of success

Properties:

- Mean: $\mathbb{E}[X] = p$
- Variance: $\text{Var}(X) = p(1-p)$
- Maximum variance at $p = 0.5$
- Building block for the binomial distribution

Life Sciences Applications:

- Whether a patient responds to treatment (yes/no)
- Presence/absence of a genetic mutation
- Survival status (alive/dead) at a given time point
- Cell division outcome (success/failure)

2.1.2 Binomial Distribution

Definition 2.2 (Binomial Distribution)**.** The **binomial distribution** describes the number of successes in n independent Bernoulli trials.

$$X \sim \text{Binomial}(n, p)$$

Probability mass function:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k \in \{0, 1, \ldots, n\}$$

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the binomial coefficient.

Parameters:

- $n \in \mathbb{N}$: number of trials
- $p \in [0, 1]$: probability of success in each trial

Properties:

- Mean: $\mathbb{E}[X] = np$
- Variance: $\text{Var}(X) = np(1 - p)$
- Mode: $\lfloor (n + 1)p \rfloor$ or $\lceil (n + 1)p \rceil - 1$
- Sum of n independent Bernoulli(p) random variables
- Approximates Normal when n is large: $X \approx N(np, np(1 - p))$

Life Sciences Applications:

- Number of patients responding to treatment in a clinical trial
- Number of successful PCR amplifications out of n attempts
- Count of cells showing a particular phenotype in a sample
- Number of offspring with a recessive trait (Mendelian genetics)

2.1.3 Poisson Distribution

Definition 2.3 (Poisson Distribution)**.** The **Poisson distribution** describes the number of events occurring in a fixed interval when events happen at a constant average rate.

$$X \sim \text{Poisson}(\lambda)$$

Probability mass function:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k \in \{0, 1, 2, \ldots\}$$

Parameters:

- $\lambda > 0$: average rate (expected number of events)

Properties:

- Mean: $\mathbb{E}[X] = \lambda$
- Variance: $\text{Var}(X) = \lambda$ (mean equals variance)
- Mode: $\lfloor \lambda \rfloor$
- Limit of Binomial(n, p) as $n \to \infty$, $p \to 0$, with $np = \lambda$
- Sum of independent Poisson variables: $X_1 + X_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$

Life Sciences Applications:

- Number of mutations in a DNA sequence
- RNA-seq read counts per gene
- Number of bacterial colonies on a plate
- Cancer incidence rates in a population
- Number of ion channel openings per unit time

Remark 2.1. In RNA-seq analysis, the Poisson distribution is often replaced by the **Negative Binomial** distribution to account for overdispersion (variance greater than mean).

2.2 Continuous Distributions

Continuous distributions describe random variables that can take any value in an interval.

2.2.1 Normal (Gaussian) Distribution

Definition 2.4 (Normal Distribution)**.** The **normal distribution** (or Gaussian distribution) is a symmetric, bell-shaped continuous distribution.

$$X \sim N(\mu, \sigma^2)$$

Probability density function (PDF):

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}$$

The **standard normal distribution** has $\mu = 0$ and $\sigma = 1$:

$$Z \sim N(0,1), \quad \phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

Parameters:

- $\mu \in \mathbb{R}$: mean (location parameter)
- $\sigma^2 > 0$: variance (scale parameter squared)

Properties:

- Mean: $\mathbb{E}[X] = \mu$
- Variance: $\mathrm{Var}(X) = \sigma^2$
- Symmetric about μ: mean = median = mode
- Skewness: 0; Kurtosis: 3 (excess kurtosis = 0)
- **68-95-99.7 rule:** 68% within $\pm 1\sigma$, 95% within $\pm 2\sigma$, 99.7% within $\pm 3\sigma$
- Linear combinations remain normal: $aX + bY \sim N(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2)$
- Standardization: $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$

Life Sciences Applications:

- Measurement errors in laboratory assays
- Human height, weight, blood pressure (within populations)
- Log-transformed gene expression levels
- Distribution of sample means (Central Limit Theorem)

2.2.2 Student's t-Distribution

Definition 2.5 (Student's t-Distribution)**.** The **Student's t-distribution** arises when estimating the mean of a normally distributed population with unknown variance.

$$T \sim t_\nu$$

Probability density function:

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad t \in \mathbb{R}$$

where $\Gamma(\cdot)$ is the gamma function.

Parameters:

- $\nu > 0$: degrees of freedom

Properties:

- Mean: $\mathbb{E}[T] = 0$ for $\nu > 1$ (undefined for $\nu \leq 1$)
- Variance: $\mathrm{Var}(T) = \frac{\nu}{\nu-2}$ for $\nu > 2$ (infinite for $1 < \nu \leq 2$)
- Symmetric about zero
- **Heavy-tailed**: more extreme values than normal distribution
- Converges to $N(0, 1)$ as $\nu \to \infty$
- Arises from: $T = \frac{Z}{\sqrt{V/\nu}}$ where $Z \sim N(0, 1)$ and $V \sim \chi^2_\nu$

Life Sciences Applications:

- Student's t-tests for comparing group means
- Confidence intervals for means with small samples
- Gene expression analysis (moderated t-statistics in limma)
- Robust regression methods

Remark 2.2 (Heavy Tails)*.* The t-distribution has **heavy tails**, meaning extreme values are more likely than in a normal distribution. This is captured by its higher kurtosis: excess kurtosis $= \frac{6}{\nu-4}$ for $\nu > 4$. Heavy-tailed distributions are important for modeling data with occasional extreme observations.

2.2.3 Exponential Distribution

Definition 2.6 (Exponential Distribution)**.** The **exponential distribution** describes the time between events in a Poisson process.

$$X \sim \text{Exponential}(\lambda) \quad \text{or} \quad X \sim \text{Exp}(\lambda)$$

Probability density function:

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0$$

Cumulative distribution function:

$$F(x) = 1 - e^{-\lambda x}$$

Parameters:

- $\lambda > 0$: rate parameter (alternative: scale $\beta = 1/\lambda$)

Properties:

- Mean: $\mathbb{E}[X] = \frac{1}{\lambda}$
- Variance: $\text{Var}(X) = \frac{1}{\lambda^2}$
- Median: $\frac{\ln 2}{\lambda}$
- **Memoryless property:** $P(X > s + t \mid X > s) = P(X > t)$
- Special case of Gamma distribution: $\text{Gamma}(1, \lambda)$
- Mode: 0 (right-skewed distribution)

Life Sciences Applications:

- Time until cell division
- Time between neural spikes
- Survival time in certain constant-hazard scenarios
- Radioactive decay (half-life)
- Drug clearance from the body (first-order kinetics)

2.2.4 Gamma Distribution

Definition 2.7 (Gamma Distribution)**.** The **gamma distribution** generalizes the exponential distribution and models waiting times.

$$X \sim \text{Gamma}(\alpha, \beta) \quad \text{or} \quad X \sim \Gamma(\alpha, \beta)$$

Probability density function:

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad x > 0$$

where $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$ is the gamma function.

Parameters:

- $\alpha > 0$: shape parameter
- $\beta > 0$: rate parameter (alternative: scale $\theta = 1/\beta$)

Properties:

- Mean: $\mathbb{E}[X] = \frac{\alpha}{\beta}$
- Variance: $\text{Var}(X) = \frac{\alpha}{\beta^2}$
- Mode: $\frac{\alpha-1}{\beta}$ for $\alpha \geq 1$
- Right-skewed (positive skewness $= \frac{2}{\sqrt{\alpha}}$)
- Sum of α independent $\text{Exp}(\beta)$ variables (for integer α)

- Special cases: Exponential ($\alpha = 1$), Chi-squared ($\alpha = \nu/2$, $\beta = 1/2$)

Life Sciences Applications:

- Waiting time until k-th event (Erlang distribution)
- Modeling variance in hierarchical Bayesian models
- Prior distribution for precision parameters
- Time to failure in reliability studies
- Protein expression levels

2.2.5 Log-Normal Distribution

Definition 2.8 (Log-Normal Distribution)**.** The **log-normal distribution** describes a variable whose logarithm is normally distributed.

If $Y = \ln(X)$ and $Y \sim N(\mu, \sigma^2)$, then:

$$X \sim \text{LogNormal}(\mu, \sigma^2)$$

Probability density function:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right), \quad x > 0$$

Parameters:

- $\mu \in \mathbb{R}$: mean of $\ln(X)$
- $\sigma^2 > 0$: variance of $\ln(X)$

Properties:

- Mean: $\mathbb{E}[X] = e^{\mu+\sigma^2/2}$
- Variance: $\text{Var}(X) = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$
- Median: e^{μ}
- Mode: $e^{\mu-\sigma^2}$
- Always positive: $X > 0$
- Right-skewed; **heavy-tailed** for large σ
- Product of log-normal variables is log-normal
- Geometric mean of data follows log-normal distribution

Life Sciences Applications:

- Gene expression levels (microarray intensities, RNA-seq counts)
- Protein concentrations
- Cell sizes and organism body masses
- Drug concentrations in pharmacokinetics
- Survival times in certain medical contexts
- Species abundance in ecological studies

Remark 2.3. The log-normal distribution arises naturally when a quantity is the product of many independent positive random factors (multiplicative processes), just as the normal distribution arises from additive processes (Central Limit Theorem).

2.2.6 Dirichlet Distribution

Definition 2.9 (Dirichlet Distribution)**.** The **Dirichlet distribution** is a multivariate generalization of the Beta distribution, describing probability distributions over probability vectors.

$$\mathbf{X} = (X_1, \ldots, X_K) \sim \text{Dirichlet}(\boldsymbol{\alpha}) \quad \text{where } \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$$

Probability density function:

$$f(\mathbf{x}; \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} x_k^{\alpha_k - 1}$$

where $\alpha_0 = \sum_{k=1}^{K} \alpha_k$ and the simplex constraint: $\sum_{k=1}^{K} x_k = 1$, $x_k \geq 0$.

Parameters:

- $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ with $\alpha_k > 0$: concentration parameters
- K: number of categories

Properties:

- Support: $(K-1)$-dimensional simplex (probabilities sum to 1)
- Mean: $\mathbb{E}[X_k] = \frac{\alpha_k}{\alpha_0}$
- Variance: $\text{Var}(X_k) = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)}$
- Marginals: $X_k \sim \text{Beta}(\alpha_k, \alpha_0 - \alpha_k)$
- Conjugate prior for the multinomial distribution
- When $\alpha_k = 1$ for all k: uniform distribution on the simplex
- Larger α_0 = more concentrated distribution; smaller α_0 = more spread

Life Sciences Applications:

- Modeling cell type proportions in tissue deconvolution
- Allele frequencies in population genetics
- Microbiome composition (relative abundance of taxa)
- Probabilistic topic models for biological text mining
- Bayesian inference for categorical outcomes

2.3 Summary: Probability Distributions

Distribution	**Type**	**Support**	**Heavy Tails**	**Life Sciences Example**
Bernoulli	Discrete	$\{0, 1\}$	No	Treatment response (yes/no)
Binomial	Discrete	$\{0, \ldots, n\}$	No	# patients responding
Poisson	Discrete	$\{0, 1, 2, \ldots\}$	No	RNA-seq counts
Normal	Continuous	$\mathbb{R}$	No	Measurement errors
t-distribution	Continuous	$\mathbb{R}$	**Yes**	Small-sample t-tests
Exponential	Continuous	$[0, \infty)$	No	Time between events
Gamma	Continuous	$(0, \infty)$	Medium	Waiting times
Log-Normal	Continuous	$(0, \infty)$	**Yes**	Gene expression levels
Dirichlet	Continuous	Simplex	No	Cell type proportions

Key relationships between distributions:

- Bernoulli is Binomial with $n = 1$
- Binomial → Poisson as $n \to \infty$, $p \to 0$, $np = \lambda$
- Binomial → Normal as $n \to \infty$ (CLT)
- Exponential is Gamma with $\alpha = 1$
- t-distribution → Normal as $\nu \to \infty$
- Log-Normal: X is log-normal iff $\ln(X)$ is normal
- Dirichlet marginals are Beta distributions

3 Point Estimation and Confidence Intervals

3.1 Point Estimation

Definition 3.1 (Point Estimator)**.** A **point estimator** $\hat{\theta}$ is a statistic (function of the sample data) used to estimate an unknown population parameter θ. Key properties of estimators include:

- **Unbiasedness:** $\mathbb{E}[\hat{\theta}] = \theta$
- **Consistency:** $\hat{\theta} \xrightarrow{p} \theta$ as $n \to \infty$
- **Efficiency:** Achieves minimum variance among unbiased estimators

3.1.1 Maximum Likelihood Estimation (MLE)

Definition 3.2 (Maximum Likelihood Estimator)**.** Given observations $x_1, \ldots, x_n$ from a distribution with parameter θ, the **maximum likelihood estimator** maximizes the likelihood function:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} L(\theta; \mathbf{x}) = \arg\max_{\theta} \prod_{i=1}^{n} f(x_i; \theta)$$

or equivalently, the log-likelihood:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \ell(\theta; \mathbf{x}) = \arg\max_{\theta} \sum_{i=1}^{n} \ln f(x_i; \theta)$$

Properties of MLE:

- Consistent: $\hat{\theta}_{\text{MLE}} \xrightarrow{p} \theta_0$ (true value)
- Asymptotically efficient: achieves Cramér-Rao lower bound
- Asymptotically normal: $\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta_0) \xrightarrow{d} N(0, I(\theta_0)^{-1})$
- Invariant under reparameterization: if $\hat{\theta}$ is MLE of θ, then $g(\hat{\theta})$ is MLE of $g(\theta)$

3.1.2 Robust Estimators

Robust estimators are less sensitive to outliers and deviations from model assumptions.

Definition 3.3 (Median Absolute Deviation (MAD))**.** The **MAD** is a robust estimator of scale:

$$\text{MAD} = \text{median}(|X_i - \text{median}(X)|)$$

To estimate the standard deviation of a normal distribution:

$$\hat{\sigma}_{\text{MAD}} = 1.4826 \times \text{MAD}$$

Definition 3.4 (Trimmed Mean)**.** The **α-trimmed mean** removes the lowest and highest $\alpha\%$ of observations:

$$\bar{x}_{\text{trim}} = \frac{1}{n-2k} \sum_{i=k+1}^{n-k} x_{(i)}$$

where $k = \lfloor \alpha n \rfloor$ and $x_{(i)}$ are order statistics.

Table 1: Comparison of estimators and confidence interval methods for common distributions

Distribution	**Parameters**	**MLE**	**Robust Alternative**	**CLT CI OK?**	**When to Prefer Other CI**
Normal (μ, σ^2)	Mean μ, Variance σ^2	$\hat{\mu} = \bar{x}$, $\hat{\sigma}^2 = s^2$	Median (for μ), MAD (for σ)	Yes	Heavy tails, outliers → robust or bootstrap
Bernoulli (p)	Probability p	$\hat{p} = \bar{x}$	Trimmed proportion if contamination	Yes (large n, p not near $0/1$)	Small n or p near $0/1$ → Clopper-Pearson or Wilson CI
Binomial (n, p)	n known, p unknown	$\hat{p} = k/n$	Same as Bernoulli	Yes (large n)	Small n or rare events → exact binomial CI
Poisson (λ)	Rate λ	$\hat{\lambda} = \bar{x}$	Median-based or trimmed mean	Yes (large n)	Small counts or overdispersion → exact or bootstrap
t (ν)	df ν (location-scale: μ, σ)	Numerical MLE for ν	Median (location), robust scale	Approx (large n)	Small samples → use t-based CI
Exponential (λ)	Rate λ	$\hat{\lambda} = 1/\bar{x}$	Median-based: $\hat{\lambda} = \ln 2/\text{med}$	Yes (via CLT)	Small n → exact CI via Gamma
Gamma (α, β)	Shape α, Rate β	Numerical MLE for α	Median-based, robust moments	Approx (large n)	Small n or skewed → bootstrap
Log-Normal (μ, σ^2)	μ, σ^2 of $\log X$	Mean, variance of $\log X$	Median of $\log X$	Yes (on log scale)	Back-transformed CI → bootstrap

3.2 Comparison of Common Distribution Estimators

3.3 Confidence Intervals

Definition 3.5 (Confidence Interval)**.** A $(1-\alpha)$ **confidence interval** for parameter θ is an interval $[L, U]$ such that:

$$P(L \leq \theta \leq U) = 1 - \alpha$$

Common confidence levels: 90% ($\alpha = 0.10$), 95% ($\alpha = 0.05$), 99% ($\alpha = 0.01$).

3.3.1 Central Limit Theorem (CLT)

Definition 3.6 (Central Limit Theorem)**.** For i.i.d. random variables $X_1, \ldots, X_n$ with mean μ and variance $\sigma^2 < \infty$:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1) \quad \text{as } n \to \infty$$

When CLT applies for confidence intervals:

- Sample size n is "large enough" (rule of thumb: $n \geq 30$)
- For proportions: $np \geq 5$ and $n(1-p) \geq 5$
- For Poisson: $\lambda n \geq 5$
- Distribution not too skewed or heavy-tailed

3.4 Exact and Analytic Confidence Intervals

3.4.1 Normal Distribution: Mean μ

Case 1: σ known (Z-interval):

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

where $z_{\alpha/2}$ is the $(1-\alpha/2)$ quantile of $N(0,1)$.

Case 2: σ unknown (t-interval):

$$\bar{x} \pm t_{\alpha/2,n-1} \frac{s}{\sqrt{n}}$$

where $t_{\alpha/2,n-1}$ is the $(1-\alpha/2)$ quantile of t_{n-1}.

Derivation: The statistic $T = \frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$ follows a t-distribution with $n-1$ degrees of freedom.

3.4.2 Normal Distribution: Variance σ^2

$$\left[\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}\right]$$

Derivation: $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$.

3.4.3 Binomial Proportion p

Wald (Normal approximation):

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Use when: $n\hat{p} \geq 5$ and $n(1-\hat{p}) \geq 5$.

Wilson Score Interval (preferred):

$$\frac{\hat{p} + \frac{z^2}{2n} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}$$

where $z = z_{\alpha/2}$.

Clopper-Pearson (Exact): Based on the relationship between binomial and beta/F distributions. Conservative but guaranteed coverage.

3.4.4 Poisson Rate λ

Normal approximation (large n):

$$\hat{\lambda} \pm z_{\alpha/2}\sqrt{\frac{\hat{\lambda}}{n}}$$

where $\hat{\lambda} = \bar{x}$ is the sample mean.

Exact confidence interval: Using the relationship between Poisson and chi-square distributions. For observed count X:

$$\left[\frac{\chi^2_{2X,\alpha/2}}{2}, \frac{\chi^2_{2(X+1),1-\alpha/2}}{2}\right]$$

3.4.5 Exponential Rate λ

Exact confidence interval: Using the relationship $2\lambda\sum X_i \sim \chi^2_{2n}$:

$$\left[\frac{\chi^2_{2n,\alpha/2}}{2\sum x_i}, \frac{\chi^2_{2n,1-\alpha/2}}{2\sum x_i}\right]$$

Equivalently, using $\bar{x}$:

$$\left[\frac{\chi^2_{2n,\alpha/2}}{2n\bar{x}}, \frac{\chi^2_{2n,1-\alpha/2}}{2n\bar{x}}\right]$$

3.5 Summary: Confidence Interval Methods

Model	Parameter	Statistic	Distribution	CI Type
Normal (σ known)	μ	$Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$	Normal	Exact
Normal (σ unknown)	μ	$T = \frac{\bar{x}-\mu}{s/\sqrt{n}}$	t_{n-1}	Exact
Normal	σ^2	$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$	χ^2_{n-1}	Exact
Binomial	p	Count X	Binomial/Beta	Exact (C-P)
Binomial (large n)	p	Z-statistic	Normal (approx)	CLT-based
Poisson	λ	Count X	χ^2 relationship	Exact
Poisson (large n)	λ	$\bar{x}$	Normal (approx)	CLT-based
Exponential	λ	$2\lambda \sum X_i$	χ^2_{2n}	Exact

Remark 3.1 (Bootstrap Confidence Intervals). When exact or CLT-based intervals are not appropriate (small samples, skewed distributions, complex estimators), **bootstrap methods** provide an alternative:

1. Resample with replacement B times (typically $B = 1000$–10000)
2. Compute the statistic for each resample
3. Use percentiles of the bootstrap distribution as CI bounds

Bootstrap is especially useful for: medians, ratios, coefficients from regression, and back-transformed parameters.

3.5.1 When to Use Bootstrap

Bootstrap is appropriate when:

- **Small sample sizes:** When $n < 30$ and CLT assumptions are questionable
- **Non-standard statistics:** Medians, trimmed means, ratios, correlation coefficients, regression coefficients
- **Unknown or complex distributions:** When the sampling distribution of the estimator has no closed form
- **Skewed data:** Heavy-tailed or asymmetric distributions where normal approximation fails
- **Back-transformed parameters:** Log-normal means, odds ratios, hazard ratios
- **Robust estimators:** MAD, Huber M-estimators, where analytic variances are complex
- **Dependent data:** Block bootstrap for time series, cluster bootstrap for hierarchical data

Bootstrap may NOT be appropriate when:

- **Very small samples:** $n < 10$–15, where bootstrap distribution poorly represents the true distribution
- **Extreme quantiles:** Estimating tail probabilities or extreme percentiles
- **Non-i.i.d. data:** Standard bootstrap assumes independence; modifications needed for dependent data

- **Discontinuous statistics:** Mode, or statistics with jumps in their distribution
- **Population parameters at boundaries:** p near 0 or 1 for proportions

Common bootstrap variants:

- **Percentile bootstrap:** Use quantiles $[\hat{\theta}^*_{\alpha/2}, \hat{\theta}^*_{1-\alpha/2}]$ directly
- **BCa (Bias-corrected and accelerated):** Adjusts for bias and skewness; generally preferred
- **Parametric bootstrap:** Resample from fitted parametric distribution rather than empirical distribution
- **Block bootstrap:** For time series data; resamples blocks of consecutive observations

4 Correlation Metrics

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

4.1 Covariance

Definition 4.1 (Covariance)**.** The **sample covariance** measures the joint variability of two variables:

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

For populations:

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

Properties:

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

Computational formula:

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

4.2 Pearson Correlation Coefficient

Definition 4.2 (Pearson Correlation)**.** The **Pearson correlation coefficient** is the standardized covariance:

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

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: $\mathrm{Corr}(aX + b, cY + d) = \mathrm{sign}(ac) \cdot \mathrm{Corr}(X, Y)$

Coefficient of determination:

$$R^2 = r^2$$

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

Remark 4.1. Pearson correlation measures **linear** relationships only. A correlation of zero does not imply independence—the variables may have a strong nonlinear relationship.

4.3 Spearman Rank Correlation

Definition 4.3 (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 4.1. For data: $(1, 10), (2, 30), (3, 20), (4, 50), (5, 40)$

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

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

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

4.4 Kendall's Tau Correlation

Definition 4.4 (Kendall's Tau)**. Kendall's Tau (τ)** 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

4.5 Contingency Tables for Categorical Data

Definition 4.5 (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

4.5.1 Chi-Square Test of Independence

Definition 4.6 (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, $\chi^2 \sim \chi^2_{(r-1)(c-1)}$.

4.6 Cramér's V

Definition 4.7 (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:

- χ^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 2×2 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 4.2. Consider a 2×2 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 (τ)	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