Maximum Likelihood Estimation and Measuring Association

Gioele La Manno

École Polytechnique Fédérale de Lausanne (EPFL)

School of Life Science (SV)

February 2025

EPFL - BMI - UPLAMANNO

Contents

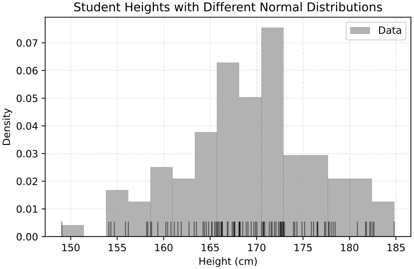
Introduction to Maximum Likelihood Estimation

The Problem - Finding Model Parameters

- We have data $x_1, ..., x_n$
- Data is assumed following a distribution family
- **Problem:** Need to determine specific parameters **but** Infinite combinations
- **Example:** Height measurements follow normal distribution
- Which μ and σ best describe our data?

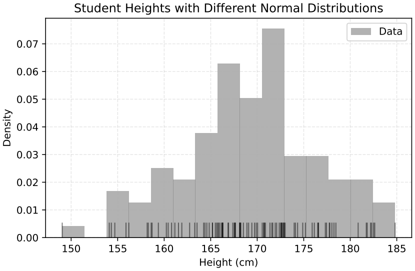
The Problem - Finding Model Parameters

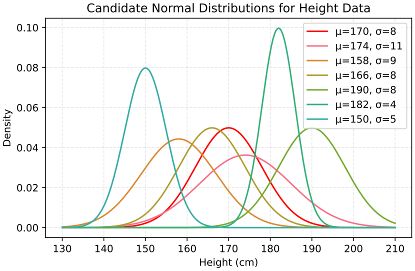
- We have data $x_1, \ldots, x_n$
- Data is assumed following a distribution family
- **Problem:** Need to determine specific parameters **but** Infinite combinations
- **Example:** Height measurements follow normal distribution
- Which μ and σ best describe our data?



The Problem - Finding Model Parameters

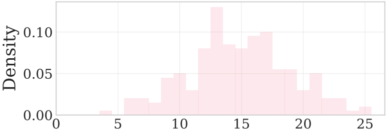
- We have data $x_1, \ldots, x_n$
- Data is assumed following a distribution family
- **Problem:** Need to determine specific parameters **but** Infinite combinations
- **Example:** Height measurements follow normal distribution
- Which μ and σ best describe our data?



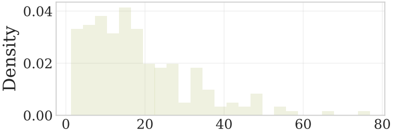


A few examples

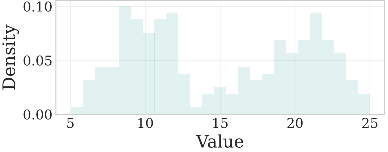
RNA Molecule Counts
(Poisson Distribution)



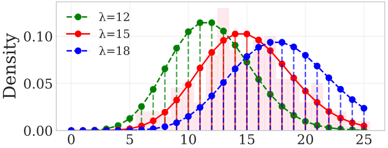
Protein Lifetimes
(Gamma Distribution)



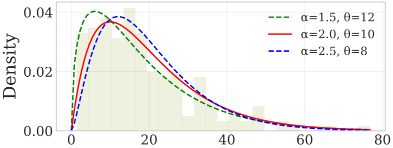
Mixed Cell Populations
(Bimodal Normal Distribution)



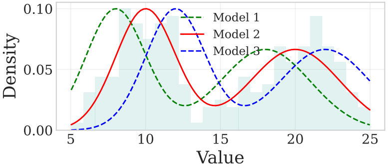
RNA Molecule Counts - Model Fitting



Protein Lifetimes - Model Fitting



Mixed Cell Populations - Model Fitting



The Likelihood Function: A New Perspective

$$L(\theta|x_1, ..., x_n) = \prod_{i=1}^{n} p(x_i|\theta)$$

- **From:** P(data — parameters) **To:** L(parameters — data)
- Treat observed data as fixed
- Parameters become variables
- "Which parameters make our data most probable?"

The Likelihood Function: A New Perspective

$$L(\boldsymbol{\theta}|x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|\boldsymbol{\theta})$$

- **From:** P(data — parameters) **To:** L(parameters — data)
- Treat observed data as fixed
- Parameters become variables
- "Which parameters make our data most probable?"

Independence in the Likelihood

$$L(\boldsymbol{\theta}|\mathbf{x}_1, \ldots, \mathbf{x}_n) = \prod_{i=1}^{n} p(\mathbf{x}_i|\boldsymbol{\theta})$$

- **Independence:** The product is allowed only between terms corresponding to different observations
- **Example**: Nuclear and cytoplasmic diameters of the same cell - not independent - a single vector observation - in bold

The Log-Likelihood: From Products to Sums

$$\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}|x_1, \ldots, x_n) = \sum_{i=1}^{n} \log p(x_i|\boldsymbol{\theta})$$

- **Why?** Products become sums
- Preserves maximum (log is monotonic)
- Prevents numerical underflow
- Easier to optimize

The Log-Likelihood: From Products to Sums

$$\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}|x_1, \ldots, x_n) = \sum_{i=1}^{n} \log p(x_i|\boldsymbol{\theta})$$

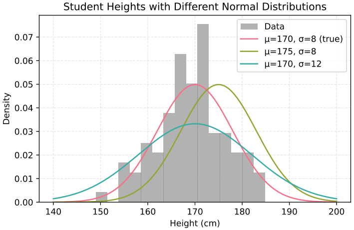
- **Why?** Products become sums
- Preserves maximum (log is monotonic)
- Prevents numerical underflow
- Easier to optimize

The Maximum Likelihood Principle

- **The Core Principle:**

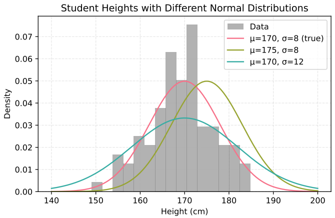
$$\hat{\theta}_{MLE} = \arg\max_{\theta} L(\theta|x_1, \dots, x_n) = \arg\max_{\theta} \ell(\theta)$$

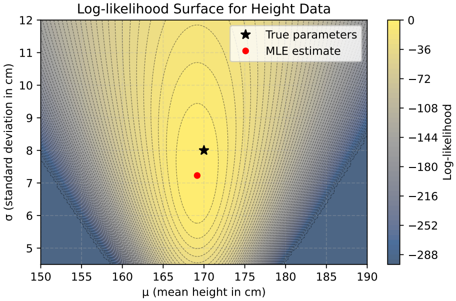
- **Intuitive Interpretation:**
  - Choose parameters that make observed data most probable
  - "Which distribution most likely generated our data?"



The Likelihood Surface

- **Bivariate Optimization:**
    - Must find optimal μ and σ simultaneously
    - Surface height indicates goodness of fit
    - Maximum corresponds to MLE solution
- **Visualization:**
    - Each point represents parameter combination
    - Brighter colors indicate higher likelihood



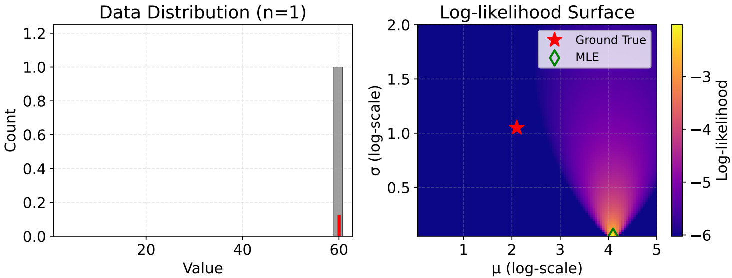


Reflecting on the Likelihood

$$L(\boldsymbol{\theta}|\mathbf{x}_1, \ldots, \mathbf{x}_n) = \prod_{i=1}^{n} p(\mathbf{x}_i|\boldsymbol{\theta})$$

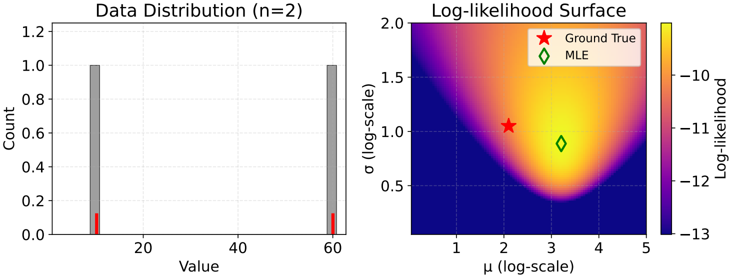
- The likelihood is primarily function of the parameters $\boldsymbol{\theta}$, because typically the data $x_1, \ldots, x_n$ was collected and can be considered fixed.
- Yet its expression depends on both the data and parameters, meaning that adding new observations changes the entire function.

Reflecting on the Likelihood (1)



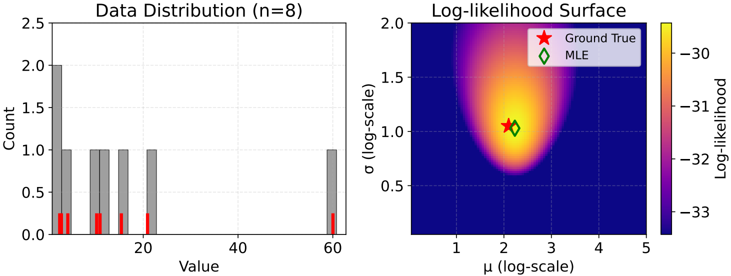
- **With one observation:** Very broad likelihood
  - Many parameter combinations explain the data well
  - High uncertainty in parameter estimates

Reflecting on the Likelihood (2)



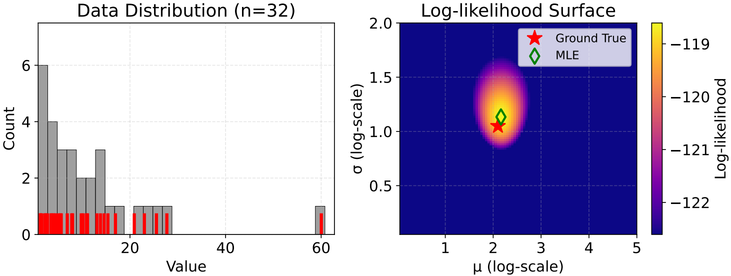
- **With eight observations:** Likelihood narrows
  - Fewer parameter combinations are compatible
  - Parameter estimates become more precise

Reflecting on the Likelihood (3)



- **With thirty-two observations:** Sharp likelihood
- Clear peak emerges
- Much more confident parameter estimates

Reflecting on the Likelihood (4)



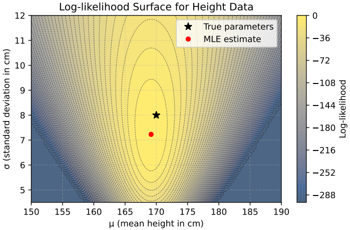
- **With 100 observations:** Likelihood narrows
  - Fewer parameter combinations are compatible
  - Parameter estimates become more precise

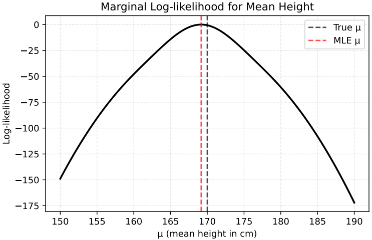
Marginal Likelihood Profiles

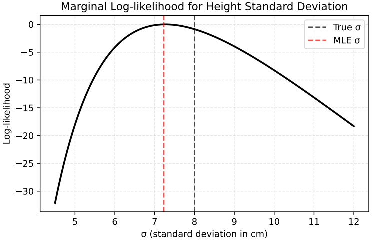
- **Marginal Likelihood:**

$$L_{\text{marginal}}(\mu) = \int_0^\infty L(\mu, \sigma)d\sigma$$

$$L_{\text{marginal}}(\sigma) = \int_{-\infty}^\infty L(\mu, \sigma)d\mu$$







Finding Maximum Likelihood Estimates: The Setup

- **The Problem:** Find μ and σ^2 that maximize likelihood
- **Starting point:** The likelihood function as product of the pdf evaluated per observation

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

  Log-likelihood function

$$\ell(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}$$

- **Strategy:**
  - Take partial derivatives
  - Set each equal to zero
  - Solve resulting equations

Finding Maximum Likelihood Estimates: The Setup

- **The Problem:** Find μ and σ^2 that maximize likelihood
- **Starting point:** The likelihood function as product of the pdf evaluated per observation

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

Log-likelihood function

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}$$

- **Strategy:**
  - Take partial derivatives
  - Set each equal to zero
  - Solve resulting equations

Finding Maximum Likelihood Estimates: The Setup

- **The Problem:** Find μ and σ^2 that maximize likelihood
- **Starting point:** The likelihood function as product of the pdf evaluated per observation

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

Log-likelihood function

$$\ell(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}$$

- **Strategy:**
  - Take partial derivatives
  - Set each equal to zero
  - Solve resulting equations

Finding Maximum Likelihood Estimates: Mean

- **Step 1:** Take derivative with respect to μ

$$\frac{\partial \ell}{\partial \mu} = \frac{\partial}{\partial \mu}\left[-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}\right]$$

- **Step 2:** Simplify

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

- **Step 3:** Set equal to zero and solve

$$\sum_{i=1}^{n}(x_i - \hat{\mu}) = 0 \implies \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- **Result:** MLE for μ is the sample mean

Finding Maximum Likelihood Estimates: Mean

- **Step 1:** Take derivative with respect to μ

$$\frac{\partial \ell}{\partial \mu} = \frac{\partial}{\partial \mu}\left[-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}\right]$$

- **Step 2:** Simplify

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

- **Step 3:** Set equal to zero and solve

$$\sum_{i=1}^{n}(x_i - \hat{\mu}) = 0 \implies \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- **Result:** MLE for μ is the sample mean

Finding Maximum Likelihood Estimates: Mean

- **Step 1:** Take derivative with respect to μ

$$\frac{\partial \ell}{\partial \mu} = \frac{\partial}{\partial \mu}\left[-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}\right]$$

- **Step 2:** Simplify

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

- **Step 3:** Set equal to zero and solve

$$\sum_{i=1}^{n}(x_i - \hat{\mu}) = 0 \implies \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- **Result:** MLE for μ is the sample mean

Finding Maximum Likelihood Estimates: Variance

- **Step 1:** Take derivative with respect to σ^2

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

- **Step 2:** Set equal to zero and solve

$$-\frac{n}{2\sigma^2} + \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2(\sigma^2)^2} = 0$$

- **Step 3:** Use plug-in estimation ($\hat{\mu}$ for μ)

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

Finding Maximum Likelihood Estimates: Variance

- **Step 1:** Take derivative with respect to σ^2

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

- **Step 2:** Set equal to zero and solve

$$-\frac{n}{2\sigma^2} + \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2(\sigma^2)^2} = 0$$

- **Step 3:** Use plug-in estimation ($\hat{\mu}$ for μ)

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

Finding Maximum Likelihood Estimates: Variance

- **Step 1:** Take derivative with respect to σ^2

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

- **Step 2:** Set equal to zero and solve

$$-\frac{n}{2\sigma^2} + \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2(\sigma^2)^2} = 0$$

- **Step 3:** Use plug-in estimation ($\hat{\mu}$ for μ)

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

Finding Maximum Likelihood Estimates: Variance

- **Step 3:** Use plug-in estimation ($\hat{\mu}$ for μ)

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

- **Key Points:**
  - Sequential estimation works here
  - MLE uses n in denominator (not $n - 1$)

The Bias in MLE Variance Estimation

- Why n instead of $n-1$ in MLE?

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

- **The Reason:**
  - Using $\hat{\mu}$ forces data to be centered
  - Systematic underestimation: deviations $(x_i - \hat{\mu})^2 \leq (x_i - \mu)^2$
- **The Bias:**

$$E[\hat{\sigma}^2_{MLE}] = \frac{n-1}{n}\sigma^2$$

- **The Solution:** Unbiased estimator

$$s^2 = \frac{n}{n-1}\hat{\sigma}^2_{MLE} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

Key Point: MLEs can be biased despite their optimal properties!

The Bias in MLE Variance Estimation

- Why n instead of $n-1$ in MLE?

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

- **The Reason:**
  - Using $\hat{\mu}$ forces data to be centered
  - Systematic underestimation: deviations $(x_i - \hat{\mu})^2 \leq (x_i - \mu)^2$
- **The Bias:**

$$E[\hat{\sigma}^2_{MLE}] = \frac{n-1}{n}\sigma^2$$

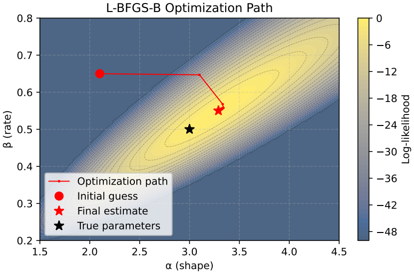
- **The Solution:** Unbiased estimator

$$s^2 = \frac{n}{n-1}\hat{\sigma}^2_{MLE} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

Key Point: MLEs can be biased despite their optimal properties!

Beyond Analytical Solutions

- **The Challenge:**
  - Some distributions and more complex models lack closed-form MLEs
  - Or too complex to solve analytically
- **Solution:**
  - Optimization algorithms (Iterative numerical methods)



Example: MLE for the Gamma Distribution

- **Problem:** Protein degradation times

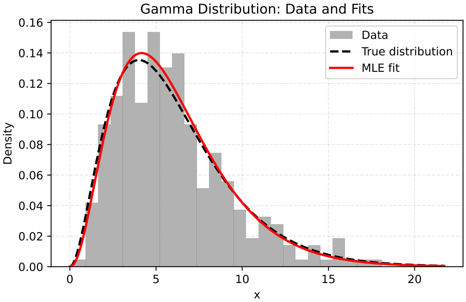
$$f(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

- **Log-likelihood:**

$$\ell(\alpha,\beta) = n\alpha\log\beta - n\log\Gamma(\alpha)$$

$$+(\alpha-1)\sum\log(x_i) - \beta\sum x_i$$

- **Challenge:** No closed-form solution!



Gradient Descent for Gamma MLE

- **Compute Gradients:**

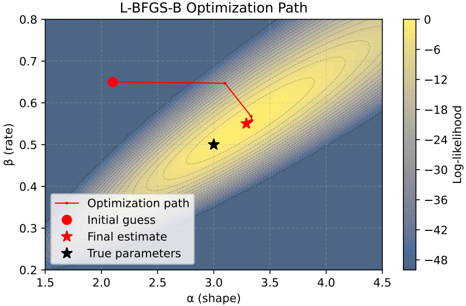
$$\frac{\partial \ell}{\partial \alpha} = n \log \beta - n\psi(\alpha) + \sum_{i=1}^{n} \log(x_i)$$

$$\frac{\partial \ell}{\partial \beta} = \frac{n\alpha}{\beta} - \sum_{i=1}^{n} x_i$$

- **Update Rules:**

$$\alpha^{(t+1)} = \alpha^{(t)} + \alpha_{rate} \frac{\partial \ell}{\partial \alpha}$$

$$\beta^{(t+1)} = \beta^{(t)} + \beta_{rate} \frac{\partial \ell}{\partial \beta}$$



Example of Exam Questions

Question 1: What is the fundamental idea behind Maximum Likelihood Estimation (MLE)?

A. Finding parameters that minimize the probability of observed data

B. Maximizing the probability of the observed data given the parameters

C. Finding parameters that match the sample mean exactly

D. Estimating parameters by averaging all possible values

E. Choosing the parameters that yield the largest standard deviation

Example of Exam Questions

Question 2: Suppose a dataset consists of independent and identically distributed (i.i.d.) observations following a normal distribution with unknown mean μ and variance σ^2. What is the likelihood function $L(\mu, \sigma^2)$ for this dataset?

A. The sum of probability densities for all observations
B. The product of probability densities for all observations
C. The difference between the probability densities of the first and last observation
D. The average of probability densities for all observations
E. The cumulative distribution function evaluated at the mean

Example of Exam Questions

Question 3

Given a dataset with n independent observations $x_1, x_2, ..., x_n$, following an exponential distribution with rate parameter λ, what is the Maximum Likelihood Estimator (MLE) for λ?

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

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

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

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

E. $\sum_{i=1}^{n} x_i$

Measuring Association

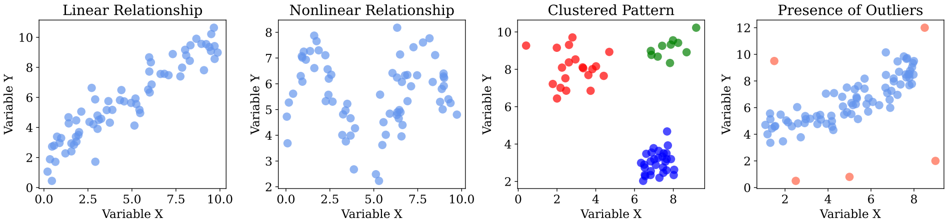
Measuring Association in Biological Data

- **Why Study Associations?**
    - Understanding relationships more insightful than isolated variables
    - Reveals how biological elements influence each other
    - Fundamental to understanding complex systems
- **The Challenge:**
    - Different/Complex types of relationships
    - Need quantitative measures

Measuring Association in Biological Data

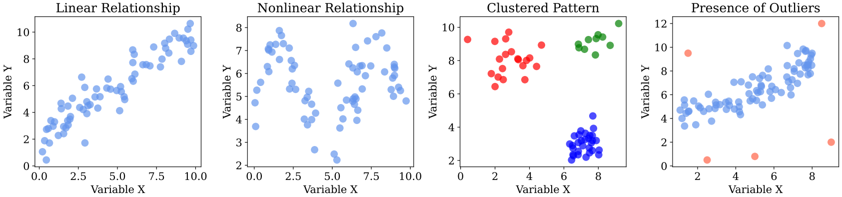
- **Why Study Associations?**
  - Understanding relationships more insightful than isolated variables
  - Reveals how biological elements influence each other
  - Fundamental to understanding complex systems

- **The Challenge:**
  - Different/Complex types of relationships
  - Need quantitative measures



Scatter Plots: Fundamental Tool for Visualization

- **What is a Scatter Plot?**
  - Each point represents one observation
  - Position determined by two variables



- **What can one see?**
  - **Direction**: Increase together, decrease together, or move in opposite directions
  - **Form**: Whether the relationship appears linear, curved, or follows some other pattern
  - **Strength**: How closely the points adhere to a pattern
  - **Outliers**: Points that deviate substantially from the overall pattern

Example - Proteomics Data

- **Data Structure:**
  - Matrix format
  - Rows: Proteins (thousands)
  - Columns: Samples/Patients
  - Values: Protein abundance

Protein	**P1**	**P2**	**P3**	**P4**
Protein A	10.2	15.3	8.7	12.1
Protein B	5.6	7.2	4.3	6.8
Protein C	0.3	0.2	3.5	0.4

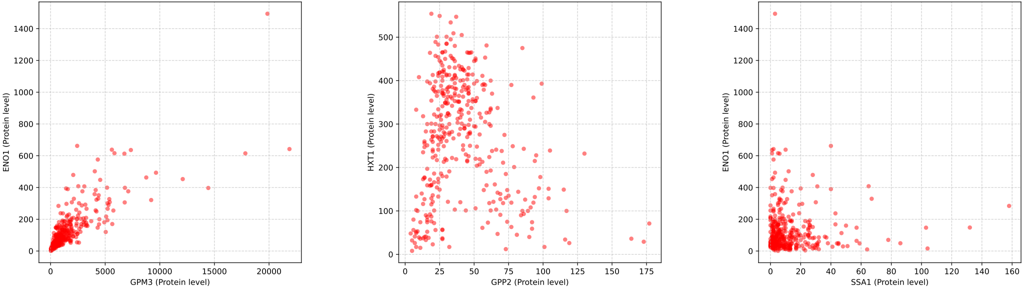
- **Values Represent:**
  - Mass spectrometry intensities
  - Antibody-based measurements
  - Normalized abundance values

Analysis Approaches in Proteomics

- **Two Key Perspectives:**
  - Protein-Protein correlations
  - Sample-Sample comparisons
- **Biological Questions:**
  - Which proteins are co-regulated?
  - Do protein complexes show coordinated expression?
  - How do disease states differ?
  - Are there distinct patient subgroups?

Scatter: Protein-Protein View

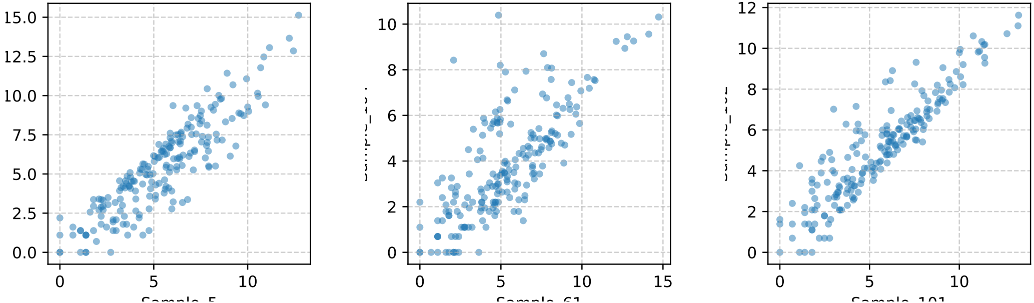
- **Points:** Each represents one sample



Different patterns of protein co-expression

Scatter Plots in Proteomics: Sample-Sample View

- **Points:** Each represents one protein



Different patterns in sample comparisons

Covariance: The Basic Measure

Definition (Covariance)

The covariance between random variables X and Y is:

$$\mathrm{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$

- Sample estimate:

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

Correlation: A Normalized Measure

Definition (Pearson Correlation)

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

Key Properties of Correlation

- **Symmetry:**

$$\rho_{X,Y} = \rho_{Y,X}$$

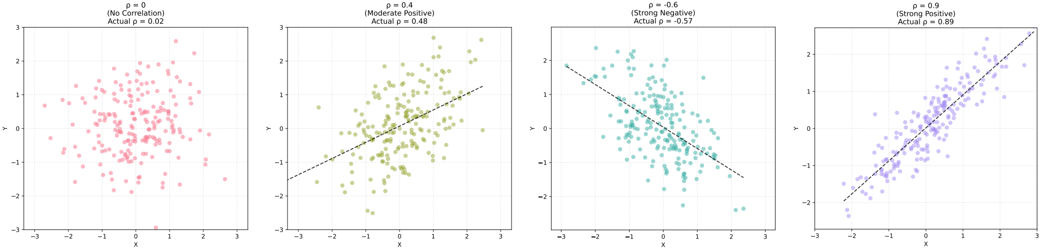
- **Range:**

$$-1 \leq \rho_{X,Y} \leq 1$$

- **Scale Invariance:** For $a, c \neq 0$

$$\rho_{aX+b,cY+d} = \text{sign}(ac) \cdot \rho_{X,Y}$$

Visual Guide to Correlation Values


ρ = 0
(No Correlation)
Actual ρ = 0.02

ρ = 0.4
(Moderate Positive)
Actual ρ = 0.48

ρ = -0.6
(Strong Negative)
Actual ρ = -0.57

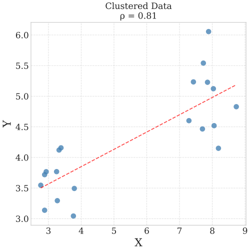
ρ = 0.9
(Strong Positive)
Actual ρ = 0.89

- Values near ±1: Strong linear relationship
- Values near 0: No linear relationship

Correlation - Avoiding Misconceptions

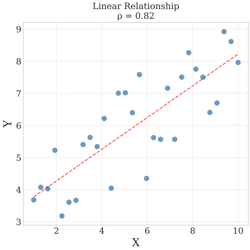
- **Correlation $\neq$ Causation**
- **Linear Relationships Only**
- **Zero Correlation $\neq$ Independence**
- **Outlier Sensitivity**

Why Visualize? Same Correlation, Different Stories



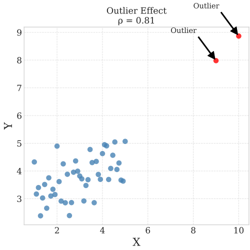
Clusters

$\rho = 0.57$



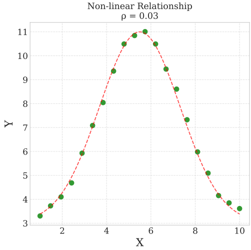
Linear

$\rho = 0.57$



Outliers

$\rho = 0.57$



Nonlinear

$\rho \approx 0$

- First three: Same correlation, radically different patterns
- Last one: Strong relationship, but correlation near zero

Geometric Interpretation of Correlation

- For centered vectors $\mathbf{x}_c$ and $\mathbf{y}_c$:

$$\rho_{X,Y} = \cos(\theta) = \frac{\mathbf{x}_c \cdot \mathbf{y}_c}{||\mathbf{x}_c||||\mathbf{y}_c||}$$

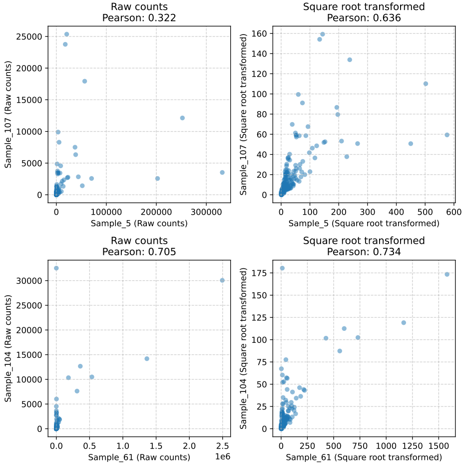
- θ is angle between vectors in $\mathbb{R}^n$
- Correlation = cosine of angle

Geometric Properties of Correlation

- **Perfect positive:**
    - $\rho = 1$
    - $\theta = 0$
    - Same direction
- **Perfect negative:**
    - $\rho = -1$
    - $\theta = 180$
    - Opposite directions
- **No correlation:**
    - $\rho = 0$
    - $\theta = 90$
    - Perpendicular vectors

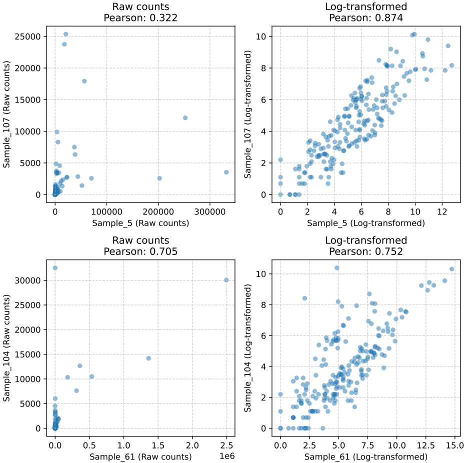
Data Transformations in Correlation Analysis

- **Common Challenges:**
    - Non-linear relationships
    - Skewed distributions
    - Multiple orders of magnitude
- **Transformations:**
    - **Square root**
    - Logarithm
    - Logit
- **Goals:**
    - Reduce skewness
    - Linearize relationships
- **Warning:**
    - Different transformations
    - → Different correlations



Data Transformations in Correlation Analysis

- **Common Challenges:**
  - Non-linear relationships
  - Skewed distributions
  - Multiple orders of magnitude
- **Transformations:**
  - Square root
  - **Logarithm**
  - Logit
- **Goals:**
  - Reduce skewness
  - Linearize relationships
- **Warning:**
  - Different transformations
  - → Different correlations



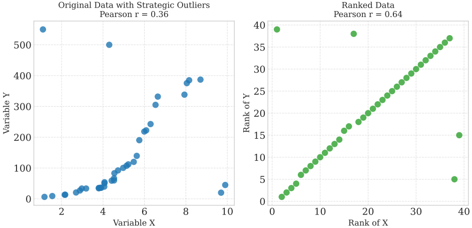
Spearman's Rank Correlation: Beyond Linear Relationships

- **Formula:**

$$\rho_s = \rho(\mathrm{rank}(X), \mathrm{rank}(Y))$$

- **Key Idea:** Replace values with ranks

- **Features:**
  - Detect non-linear relationships
  - Less sensitive to outliers

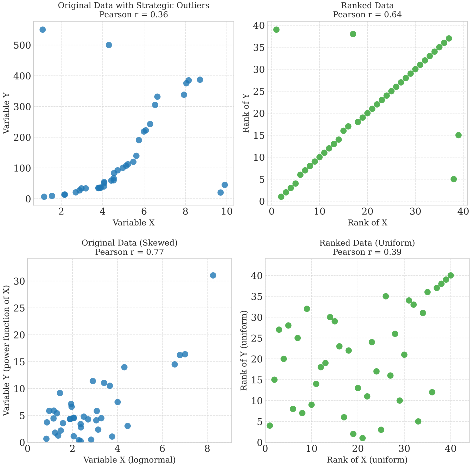


Spearman's Rank Correlation: Beyond Linear Relationships

- **Formula:**

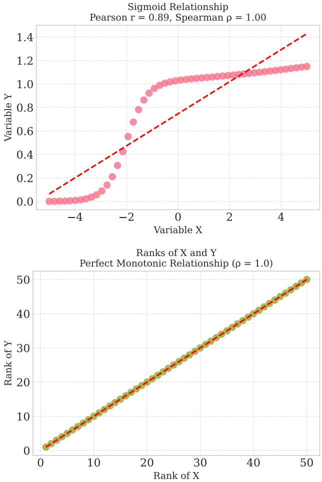
$$\rho_s = \rho(\mathrm{rank}(X), \mathrm{rank}(Y))$$

- **Key Idea:** Replace values with ranks
- **Features:**
  - Detect non-linear relationships
  - Less sensitive to outliers



Transformation Invariance: A Key Advantage

- **Invariance Property:**
  - If $Y = f(X)$ where f is monotonic: $\rho_s(X, Y) = \pm 1$
  - Same value for raw or transformed data
- **Benefits:**
  - No need to choose "correct" scale
  - Platform-independent comparisons
  - Reveals relationships automatically



When to Use Spearman Correlation in Biological Data Analysis

- **Use Spearman when:**
    - Data contains influential outliers
    - Comparing different measurement platforms
    - Expecting monotonic but non-linear relationships
    - Uncertain about measurement scale
- **Key Advantage:**
    - Reveals biological relationships
    - No subjective data transformation needed

Example of Exam Questions

Question 4: A researcher is testing whether two continuous variables X and Y are associated. She computes their Pearson correlation coefficient and finds $r = 0.85$. What does this indicate?

A. X and Y have a strong negative linear relationship
B. X and Y are independent
C. X and Y have a strong positive linear relationship
D. X is the cause of Y
E. X and Y follow a nonlinear relationship

Example of Exam Questions

Question 5 Which of the following best describes the key difference between Pearson and Spearman correlation coefficients?

A. Pearson measures linear relationships, while Spearman measures monotonic relationships.
B. Pearson is only used for categorical data, while Spearman is for continuous data.
C. Spearman considers the mean of the data, whereas Pearson does not.
D. Spearman correlation requires normally distributed data, while Pearson does not.
E. Pearson correlation can only be positive, while Spearman can be negative.

Example of Exam Questions

Question 6 Which is the mathematical expression of Cov(X, Y) for two random variables X and Y?

A. $E[XY] - E[X]E[Y]$

B. $E[X]E[Y] - E[XY]$

C. $E[X] - E[Y]$

D. $E[X] + E[Y]$

E. $E[X]E[Y]$