Lecture 2 - **Maximum Likelihood Estimation and Measuring Association**

BIOENG-210 Course Notes
Prof. Gioele La Manno

February 2024

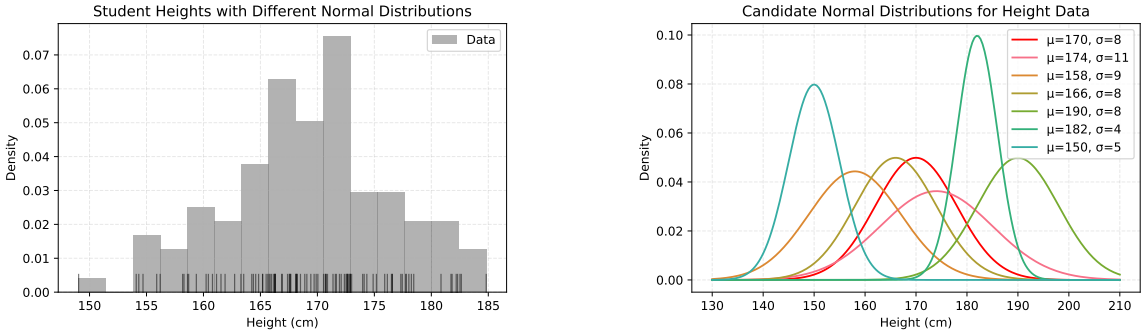
Contents

1 Introduction to Maximum Likelihood Estimation

In the realm of statistical inference, one of our fundamental challenges is to bridge the gap between our theoretical understanding of probability distributions and the practical reality of data analysis. When we observe data that appears to follow a particular distribution family, how do we determine the specific parameters that best describe our observations? This question leads us to one of statistics' most powerful and elegant tools: Maximum Likelihood Estimation (MLE).

1.1 The Problem

Imagine you are a geneticist studying the dependency of adult heights and a set of genes involved in skeletal development. After measuring a hundred of people, you notice that the heights seem to follow a normal distribution - the familiar bell-shaped curve. You decide to model this data using a normal distribution, which is characterized by two parameters: the mean μ and the standard deviation σ. But which normal distribution? Which parameter do we set? There are infinitely many possible normal distributions, each characterized by different values of μ (mean) and σ. The challenge of parameter estimation is to determine which specific values of these parameters best explain our observed data.



(a) Distribution of heights in the sample population

(b) Candidate normal distributions for height data

Figure 1: **Maximum likelihood estimation: choosing the best distribution**

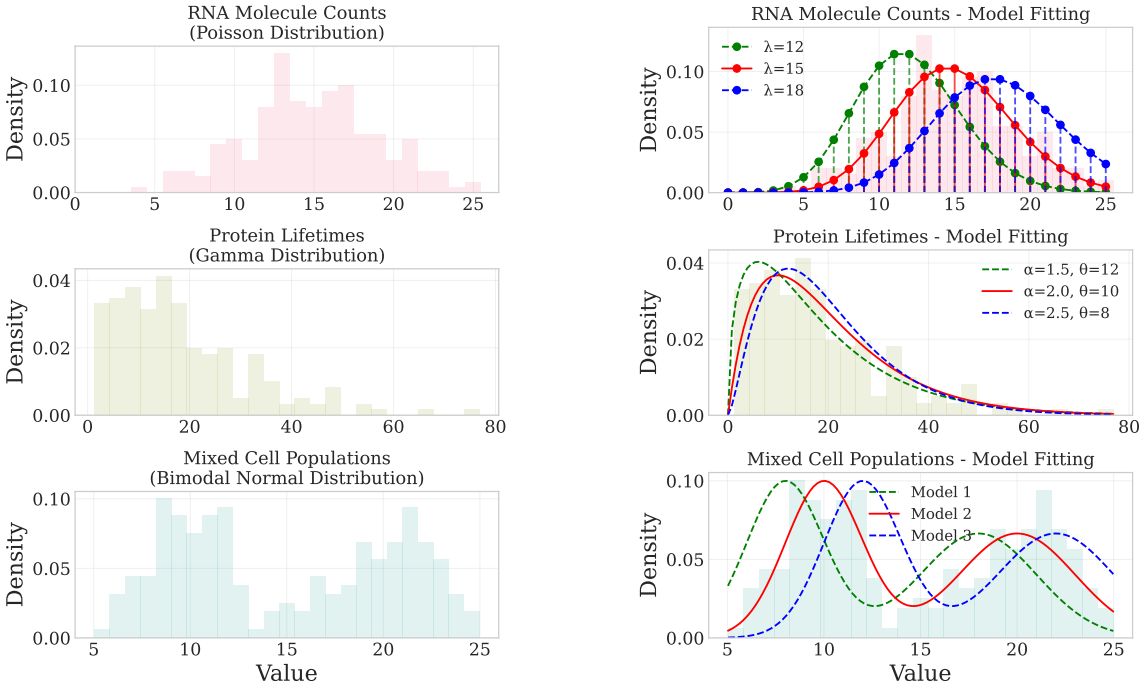
This scenario illustrates the core problem that Maximum Likelihood Estimation addresses: given a set of observations and a family of probability distributions (or, more in general, a model), how do we choose the distribution parameters that make our observed data most probable?

The Figure 2 below shows examples of different occurrences of this problem. Consider the following examples illustrating different distribution fitting scenarios:

- Discrete count data (e.g., number of RNA molecules) modeled with a Poisson distribution
- Right-skewed continuous data (e.g., protein lifetimes) modeled with a gamma distribution
- Bimodal data (e.g., mixed cell populations) modeled with a mixture of two Gaussians

1.2 The Likelihood Function: A Bridge Between Data and Theory

At the heart of Maximum Likelihood Estimation lies a profound shift in perspective. Instead of thinking about the probability of future observations given fixed parameters, we consider how likely our observed data would be under different parameter values. This leads us to the concept of the likelihood function.



(a) Data distributions

(b) Candidate distributions for data fitting

Figure 2: **Maximum likelihood estimation: fitting different data distributions**

Definition 1.1 (Likelihood Function)**.** For independent observations $x_1, ..., x_n$ and a probability model with parameter(s) $\boldsymbol{\theta}$, the likelihood function is defined as:

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

where $p(x_i|\boldsymbol{\theta})$ is the probability density (or mass) function evaluated at observation x_i.

The likelihood function reverses our usual probabilistic thinking. Rather than treating $\boldsymbol{\theta}$ as fixed and x as variable (as we do in probability calculations), we now treat our observed data x as fixed and consider $\boldsymbol{\theta}$ as variable. This subtle but crucial shift allows us to ask: "Which parameter values would make our observed data most likely?"

1.2.1 Independence of Observations

An important concept in likelihood functions is that observations $x_1, ..., x_n$ are independent. The interpretation of independence relates to the unit of observation. Consider two biological examples:

- When measuring both nuclear and cytoplasmic diameters of the same cell, these measurements are not independent and should be treated as a single vector observation
- In a blood test the measurements of the same patient are not independent, but the measurements of different patients are.

We can formalize this by writing the likelihood as:

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

where $\mathbf{x}_i$ is a vector containing related measurements that form a single observation.

1.3 The Log-Likelihood: From Products to Sums

The expression of the likelihood contains a product of many (maybe small) terms. Computationally working with products of probabilities can be challenging: numbers can become extremely small, leading to numerical underflow. Moreover, products are generally more difficult to optimize than sums (why this is important will be soon apparent). This leads us to work with the log-likelihood function:

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

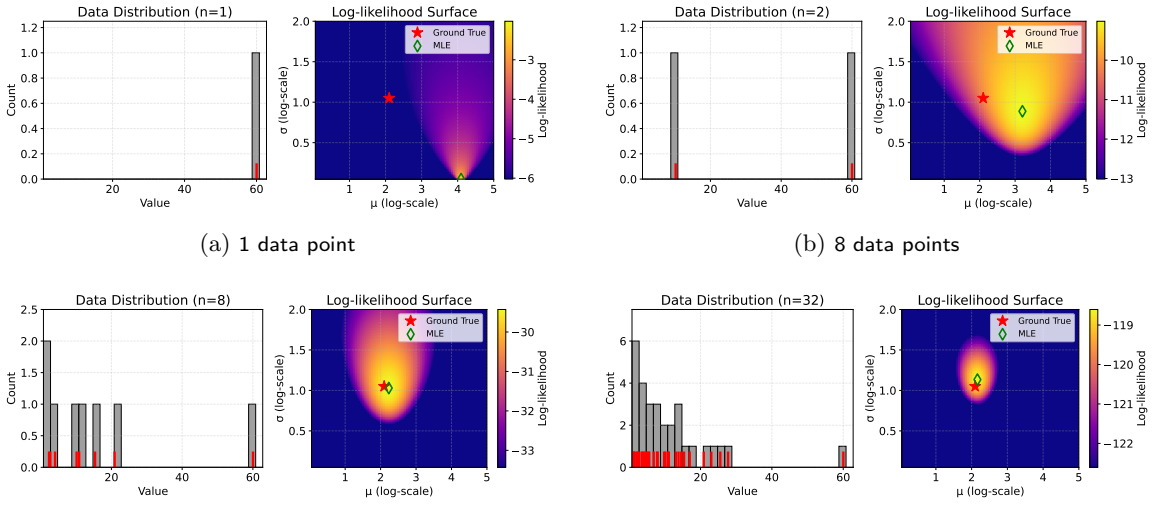
As it was anticipated, the logarithmic transformation provides several advantages:

- It converts products to sums, making calculations more manageable
- It maintains the same maximum point as the likelihood function (since log is monotonic)
- It often leads to simpler optimization problems
- It helps prevent numerical underflow in computations

1.3.1 Reflecting on the Likelihood and Log-Likelihood

Let us pause and consider what the likelihood function really represents. It is fundamentally different from a probability density function in two critical ways:

- The likelihood is primarily function of the parameters $\boldsymbol{\theta}$, because typically the data $x_1, ..., x_n$ 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.



(a) 1 data point

(b) 8 data points

(c) 16 data points

(d) 32 data points

Figure 3: **Likelihood function estimation with increasing sample size**

This inverted perspective—fixing the data and varying the parameters—is what makes the likelihood a powerful tool for inference, but also requires careful consideration. For example the Likelihood is not a probability density function, it does not integrate to 1 over the parameter space. In general, it can have multiple peaks, corresponding to different parameter values that explain the data almost equally well.

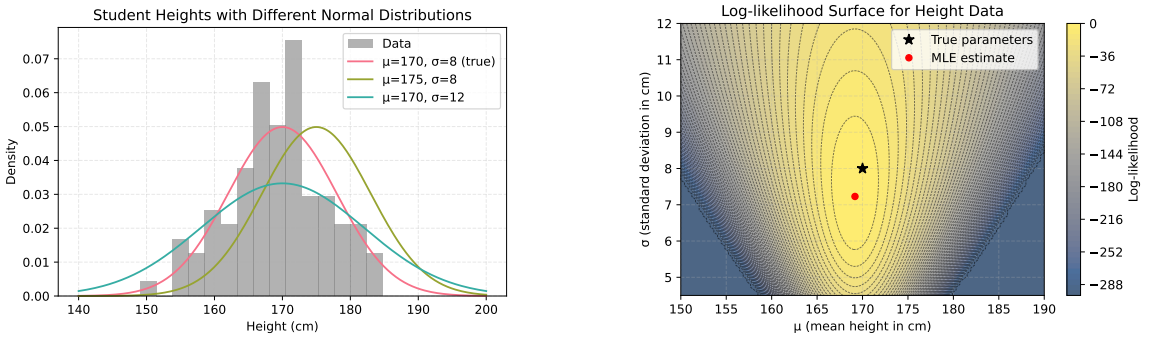
1.4 The Maximum Likelihood Principle

The fundamental principle of maximum likelihood estimation is deceptively simple: choose the parameter values that maximize the likelihood (or log-likelihood) function. Mathematically:

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

This principle has a compelling intuitive interpretation: we choose the parameter values that would make our observed data most probable. it is like asking, "If we had to bet on which probability distribution generated our data, which one should we choose?"

To make this concrete, let's examine a practical example involving height measurements. Suppose we have collected height measurements from 100 students, and we observe that these measurements appear to follow a normal distribution. The challenge lies in determining which specific normal distribution - that is, which values of μ and σ - best describes our data.



(a) Distribution of heights in the sample population

(b) Log-likelihood surface as a function of μ and σ

Figure 4: **Maximum likelihood estimation: finding the distribution that best fits the data**

The visualizations above illustrate two key aspects of our problem:

- The histogram shows our observed heights overlaid with several possible normal distributions
- The log-likelihood surface displays how well different combinations of parameters explain our data

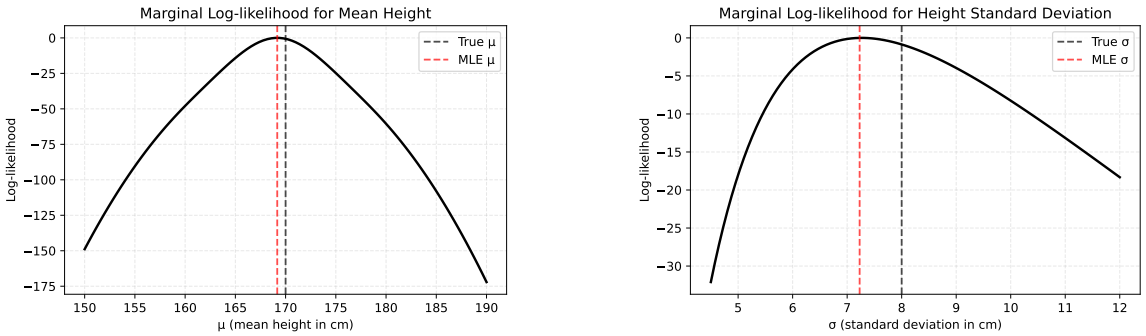
The log-likelihood surface provides a particularly illuminating view of our optimization problem. Each point on this surface represents a possible combination of μ and σ values, with the height of the surface indicating how well those parameter values explain our observed data. The MLE solution corresponds to the peak of this surface.

It is crucial to understand that our maximum likelihood problem is inherently bivariate - we must simultaneously determine both μ and σ. However, visualizing and understanding this joint structure can be challenging. To gain additional insight, we can examine the marginal likelihoods. The marginal likelihood for one parameter is obtained by integrating the joint likelihood over all possible values of the other parameter:

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

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

This marginalization process is not merely ignoring the other parameter - it is accounting for all its possible values. This gives us a way to visualize how each parameter individually affects the likelihood while properly accounting for our uncertainty in the other parameter:



(a) Marginal log-likelihood for μ

(b) Marginal log-likelihood for σ

Figure 5: **Marginal likelihood profiles for each parameter**

These marginal views help us understand how each parameter individually affects the likelihood, showing clear peaks that correspond to our maximum likelihood estimates.

1.5 Finding the Maximum Likelihood Estimates

For the normal distribution, we can find the maximum likelihood estimates analytically. Let us work through this step-by-step:

Starting with the log-likelihood function:

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

To find the maximum, we take partial derivatives with respect to both parameters and set them equal to zero:

For μ:

$$\frac{\partial \ell}{\partial \mu}\Big|_{\mu=\hat{\mu}} = \sum_{i=1}^{n}\frac{x_i - \hat{\mu}}{\sigma^2} = 0$$

This leads to our first estimate:

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

For σ^2:

$$\frac{\partial \ell}{\partial \sigma^2}\Big|_{\sigma^2=\hat{\sigma}^2} = -\frac{n}{2\hat{\sigma}^2} + \sum_{i=1}^{n}\frac{(x_i - \mu)^2}{2(\hat{\sigma}^2)^2} = 0$$

This equation still contains μ, which we do not know. However, we can use our previously derived estimate $\hat{\mu}$ - a technique known as plug-in estimation. Substituting $\hat{\mu}$ for μ and solving yields:

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

This sequential estimation process, where we first estimate μ and then plug this estimate into our formula for σ^2, is an example of how we can break down complex multiparameter estimation problems into simpler steps.

1.6 The Bias in MLE Variance Estimation

An interesting subtlety emerges in our variance estimation. Notice that our MLE for σ^2 uses n in the denominator rather than the more familiar n-1 that appears in the sample variance formula. This isn't a mistake - it reveals a fundamental property of maximum likelihood estimation.

The reason for this discrepancy lies in how we use the sample mean $\hat{\mu}$ in our variance calculation. When we substitute $\hat{\mu}$ for the true (unknown) population mean μ, we're forcing our data to be centered around $\hat{\mu}$. This makes the squared deviations $(x_i - \hat{\mu})^2$ slightly smaller than they would be if we used the true mean $(x_i - \mu)^2$. As a result, our MLE systematically underestimates the true variance.

This bias can be quantified:

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

This relationship explains why we often prefer the unbiased estimator:

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

Understanding this bias is crucial in practice. While MLEs have many desirable properties, they aren't always unbiased. This example serves as a reminder that we must always think carefully about the properties of our estimators and choose them based on our specific needs.

1.7 Beyond Analytical Solutions: Numerical Methods for MLE

While the normal distribution provides an elegant case where we can find MLEs analytically, most real-world problems require numerical optimization. The likelihood functions for many distributions do not yield closed-form solutions when we set their derivatives to zero. In such cases, we must rely on iterative numerical methods to find the maximum likelihood estimates.

1.7.1 Common Numerical Optimization Methods

Several approaches are available for finding the maximum of the likelihood function:

The Newton-Raphson method iteratively refines our parameter estimates using both first and second derivatives:

$$\theta^{(t+1)} = \theta^{(t)} - \left[\frac{\partial^2 \ell}{\partial \theta^2}\right]^{-1} \frac{\partial \ell}{\partial \theta}$$

This method typically converges quickly when we're close to the maximum, but it requires computing second derivatives and can be sensitive to starting values.

Gradient descent offers a simpler alternative, using only first derivatives:

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

where α is a learning rate that controls step size. While this method converges more slowly, it's more robust and easier to implement.

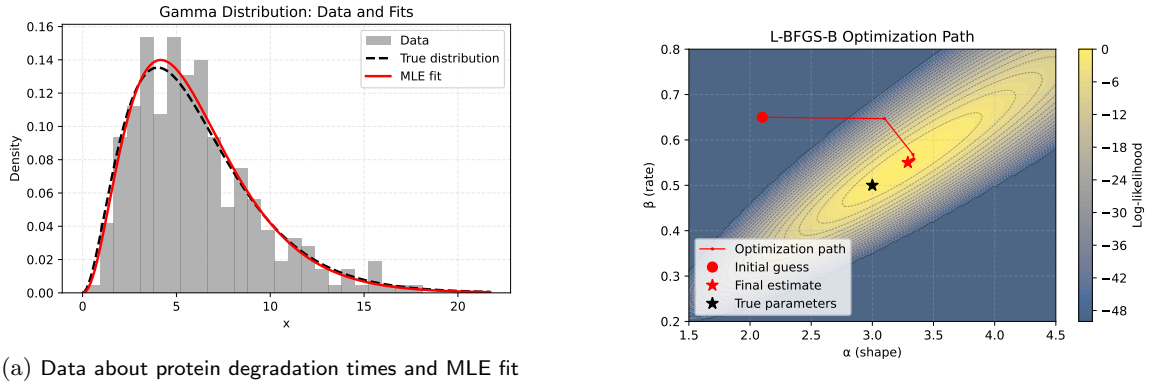
1.7.2 Example: MLE for the Gamma Distribution

Let's examine a practical example using protein degradation rates in cells. Protein degradation times often follow a gamma distribution, which has two parameters: shape (α) and rate (β). The probability density function is:

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

For a sample $x_1, ..., x_n$, the log-likelihood is:

$$\ell(\alpha, \beta) = n\alpha \log \beta - n \log \Gamma(\alpha) + (\alpha - 1) \sum_{i=1}^{n} \log(x_i) - \beta \sum_{i=1}^{n} x_i$$



(a) Data about protein degradation times and MLE fit of gamma distribution

(b) Optimization path using gradient descent

Figure 6: **Numerical optimization of gamma distribution parameters**

The left panel shows the log-likelihood surface, with brighter colors indicating higher likelihood values. The right panel demonstrates how gradient descent navigates this surface, starting from an initial guess and converging to the maximum likelihood estimates.

Let us examine how the gradient descent optimization proceeds:

1. Initialize parameters at reasonable values (often using method of moments estimates) 2. Compute gradients:

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

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

where $\psi(\alpha)$ is the digamma function

3. Update parameters using gradient descent:

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

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

4. Repeat until convergence (when parameter changes become sufficiently small)

1.7.3 Practical Considerations

Several challenges often arise in numerical optimization:

- Multiple local maxima may exist, requiring multiple starting points
- Parameters may have natural constraints (e.g., $\alpha, \beta > 0$ for gamma)
- Step sizes must be chosen carefully to ensure convergence
- Numerical instabilities can occur, especially near parameter constraints

Modern statistical softwares handle many of these issues automatically, but understanding the underlying process helps us diagnose problems and ensure reliable results. In the exercise section, we will implement these methods ourselves and explore their behavior with real biological data.

2 Measuring Association

In biological data science, understanding relationships between variables often provides deeper insights than studying variables in isolation. When we observe gene expression patterns, protein interactions, or cellular responses to stimuli, we're fundamentally interested in how these elements influence and relate to each other. The challenge lies in quantifying these relationships in meaningful ways.

2.1 Types of Association

The relationships we observe in biological systems span a spectrum of complexity. The simplest form is linear association, where changes in one variable are proportional to changes in another. Consider a simple enzymatic reaction under ideal conditions: as we increase enzyme concentration, we might see a proportional increase in the rate of product formation. Yet biology rarely confines itself to such simplicity. More often, we encounter non-linear relationships, where variables might increase together but not at a constant rate, or display even more intricate patterns of dependency that change across different ranges of measurement.

2.2 Scatter Plots: Visualizing Relationships Between Variables

Before attempting to quantify associations, it is crucial to visualize the relationships between variables. The scatter plot is the most fundamental graphical tool for this purpose, revealing patterns that summary statistics alone might miss.

Definition 2.1 (Scatter Plot)**.** A scatter plot displays the values of two variables for a set of data points, with each point positioned at the coordinates defined by its values on both variables.

Scatter plots provide immediate visual information about several aspects of the relationship between variables:

- **Direction**: Whether the variables tend to 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

The examples below illustrate how scatter plots reveal different types of relationships that might exist between variables.

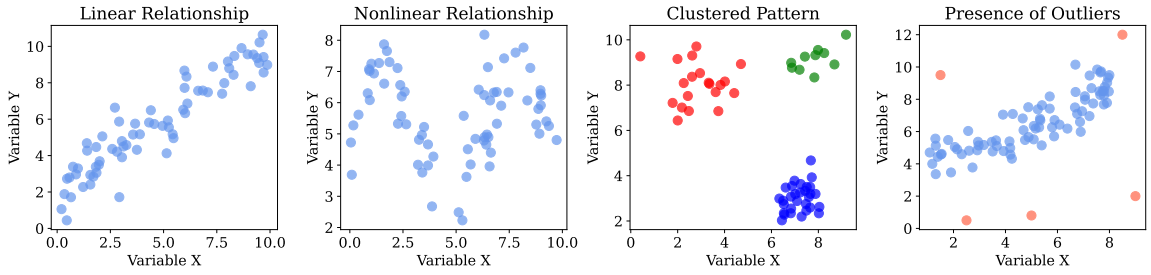


Figure 7: Common patterns in scatter plots

In the first plot, we see a clear linear relationship that would be well-captured by correlation coefficients. The second plot shows a nonlinear relationship that correlation coefficients might underestimate. The third plot reveals clustered data that might indicate distinct subpopulations. The fourth plot demonstrates how outliers can influence our perception of relationships.

2.3 Proteomics Data: A Biological Context for Association Analysis

To understand how scatter plots and association measures are applied in biological research, let's introduce proteomic data—a common data type in modern biology.

2.3.1 Structure of Proteomic Data

Proteomic data typically comes in a matrix format where:

- Rows represent proteins (often thousands)
- Columns represent samples (patients or conditions)
- Each cell value represents the abundance of a protein in that sample

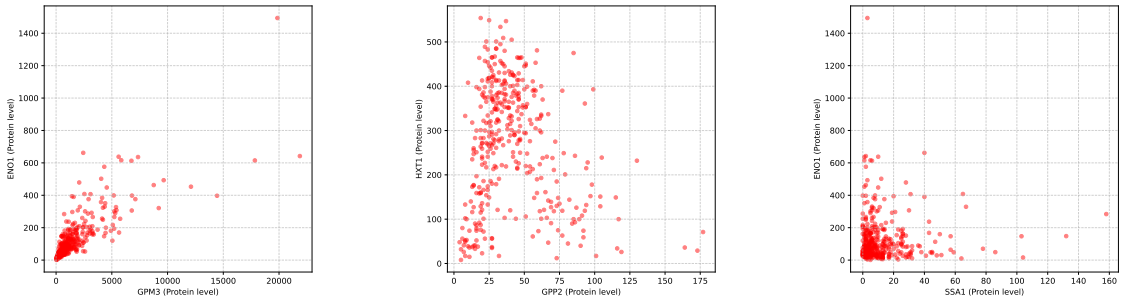
Here's a simplified example of what a proteomic matrix might look like:

Protein/Sample	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5
Protein A	10.2	15.3	8.7	12.1	9.8
Protein B	5.6	7.2	4.3	6.8	5.1
Protein C	0.3	0.2	3.5	0.4	2.8
Protein D	8.7	9.3	2.1	8.9	3.2

These values typically represent normalized intensities from mass spectrometry measurements, or relative abundance measures from antibody-based assays.

2.3.2 Scatter Plots for Proteomic Data

To examine the relationship between two proteins, we create a scatter plot where each point represents a patient, with coordinates determined by the abundance values of the two proteins in that sample.



(a) (b) (c)

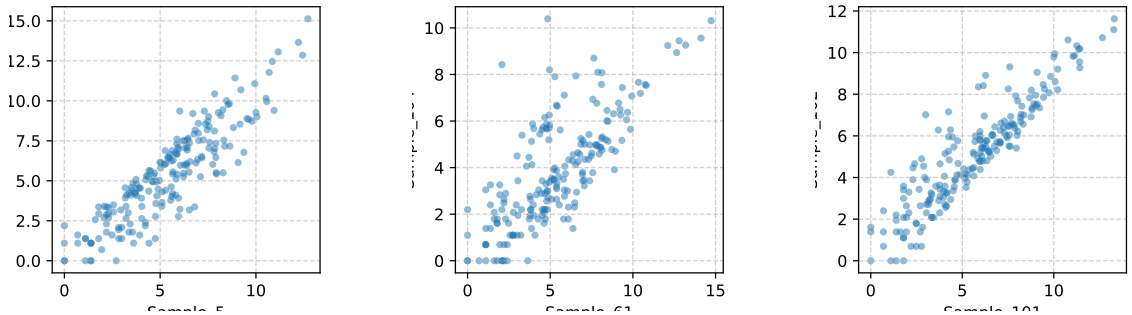
Figure 8: **Visualizing protein-protein relationships through scatter plots**

In this scatter plot, each point represents one patient, positioned according to the abundance levels of Protein A (x-axis) and Protein B (y-axis). Despite being different, all the displayed patterns suggest a positive association—as Protein A's abundance increases, Protein B's abundance tends to increase as well.

This visualization approach allows us to:

- Identify proteins that appear to be co-regulated
- Detect potential protein complexes or pathways
- Discover complex patterns of protein interaction
- Identify outlier samples that may represent different disease states or technical issues

Note that this is not the only possible view of the data. We might want to look at the other axis comparing two samples (patients) across all the proteins.



(a) (b) (c)

Figure 9: **Visualizing sample-sample relationships through scatter plots**

Biological interpretation of scatter plots: In a cancer proteomics study, researchers created scatter plots comparing the abundance of key signaling proteins. Samples with high EGFR levels consistently showed high abundance of downstream effectors like ERK and AKT (positive correlation), suggesting active signaling pathways. Meanwhile, tumor suppressors like p53 showed opposite patterns (negative correlation), consistent with dysregulation of normal cellular control mechanisms.

Once we've visualized the relationship using scatter plots, we can then quantify the strength and direction of the association using correlation measures. The choice between Pearson correlation (for linear relationships) and Spearman correlation (for monotonic relationships) often depends on the patterns observed in the scatter plot and our knowledge of the biological context.

2.4 Covariance and Correlation: Measuring Linear Relationships

When variables are not independent, we often want to quantify the strength of their relationship. The most basic measures of association are covariance and correlation.

Definition 2.2 (Covariance)**.** The covariance between random variables X and Y is:

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

Note that this is expressed in terms of operator on random variables, but in practice, we estimate it from data using the sample covariance formula:

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

While covariance measures the direction and magnitude of linear relationships, its scale dependence makes it difficult to interpret. This leads us to correlation:

Definition 2.3 (Pearson Correlation Coefficient)**.** The Pearson correlation coefficient between X and Y is:

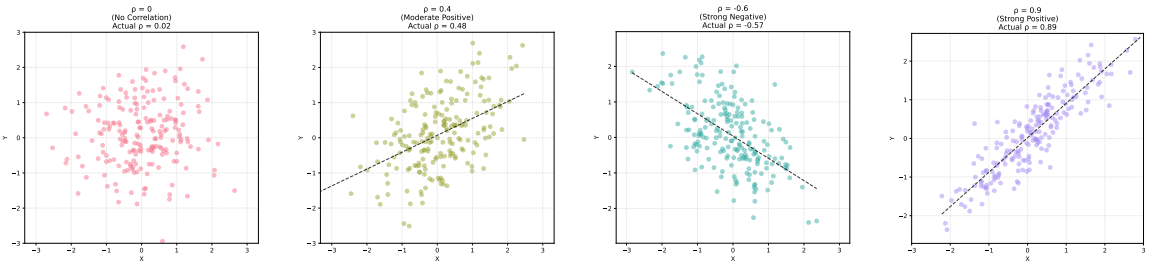
$$\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

Also here it is much more useful to estimate it from data using the sample correlation formula:

$$r_{\mathbf{x},\mathbf{y}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

The correlation coefficient has several important properties:

- Values near 1 indicate strong positive linear relationship
- Values near -1 indicate strong negative linear relationship



(a) Correlation coeff. = 0 (b) Correlation coeff. = 0.4 (c) Correlation coeff. = -0.6 (d) Correlation coeff. = 0.9

Figure 10: **Scatter plots with different correlation coefficients**

2.5 Properties and Limitations of Correlation Coefficients

Correlation coefficients have specific mathematical properties that determine their behavior and interpretation. Understanding these properties—and their limitations—is essential for correct application in biological data analysis.

2.5.1 Key Mathematical Properties

- **Symmetry**: $\rho_{X,Y} = \rho_{Y,X}$ (the order of variables doesn't matter)
- **Range**: $-1 \leq \rho_{X,Y} \leq 1$ (bounded between perfect negative and perfect positive correlation)
- **Scale Invariance**: For constants a, b, c, d where $a, c \neq 0$:

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

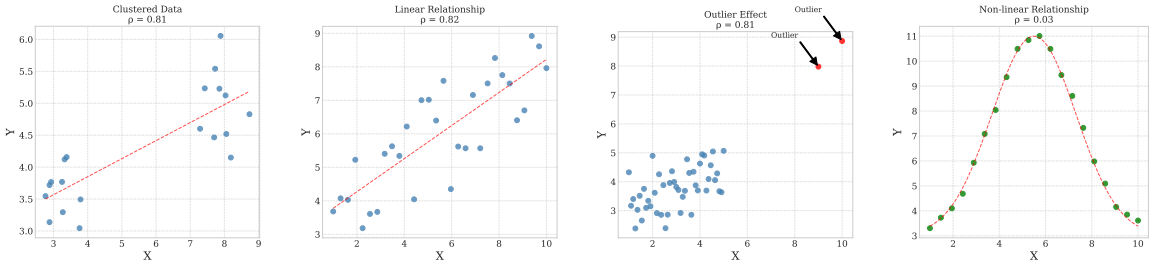
  This means linear transformations of variables (like changing units) preserve correlation magnitude
- **Standard Normal Transformation**: For variables $Z_X = \frac{X-\mu_X}{\sigma_X}$ and $Z_Y = \frac{Y-\mu_Y}{\sigma_Y}$:

  $$\rho_{X,Y} = \rho_{Z_X,Z_Y} = E[Z_X Z_Y]$$

2.5.2 Common Misconceptions and Limitations

Despite their utility, correlation coefficients are frequently misinterpreted:

- **Correlation $\neq$ Causation**: Strong correlation between variables does not imply that one causes the other
- **Linear Relationships Only**: Pearson correlation measures only linear relationships, potentially missing important non-linear patterns
- **Zero Correlation $\neq$ Independence**: Uncorrelated variables may still have strong non-linear dependencies
- **Outlier Sensitivity**: Pearson correlation can be strongly distorted by a small number of extreme values



(a) Clusters (b) Linear (c) Outliers (d) Nonlinear

Figure 11: **Correlation limitations - it remains important to look at the data**

The figure illustrates a critical limitation: four different datasets with identical correlation coefficients but fundamentally different relationships. This highlights why visual examination of data through scatter plots should always accompany correlation analysis. Depending solely on the correlation coefficient can mask important patterns in the data.

2.6 Geometric Interpretation of Correlation

Consider our data vectors $\mathbf{x} = (x_1, ..., x_n)$ and $\mathbf{y} = (y_1, ..., y_n)$, where n is the number of observations. When we center these vectors by subtracting their means ($\mathbf{x}_c = \mathbf{x} - \bar{x}$ and $\mathbf{y}_c = \mathbf{y} - \bar{y}$), the sample correlation coefficient equals:

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

where θ is the angle between the centered vectors in $\mathbb{R}^n$ space.
This geometric interpretation explains key properties of correlation:

- **Perfect positive correlation ($\rho = 1$)**: The vectors point in the same direction ($\theta = 0$)
- **Perfect negative correlation ($\rho = -1$)**: The vectors point in opposite directions ($\theta = 180$)
- **No correlation ($\rho = 0$)**: The vectors are perpendicular ($\theta = 90$)

When variables are standardized (centered and scaled to unit variance), they become unit vectors on a hypersphere. The correlation then directly represents how similarly these standardized vectors are oriented in space, explaining why correlation is invariant to linear transformations of the variables.

2.7 Data Transformations and Correlation Measures

When analyzing biological data, we often encounter relationships that deviate from linearity or datasets with skewed distributions. These characteristics can significantly impact correlation analysis and require careful consideration of data transformations.

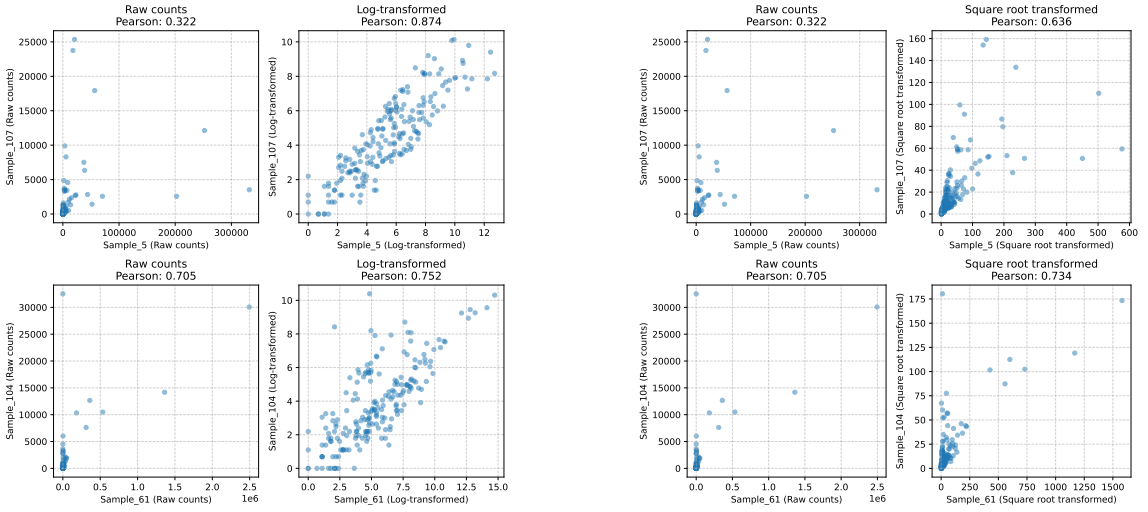
2.7.1 The Challenge of Finding Appropriate Transformations

When relationships are non-linear or data distributions are skewed, we often need to transform the natural space of the data to better capture underlying patterns. Common transformations include:

- **Log transformations**: Commonly used for gene expression, protein abundances, and other biological measurements that span multiple orders of magnitude
- **Square root transformations**: Often applied to count data with variance proportional to the mean

Finding the "right" transformation is often challenging and subjective. An appropriate transformation should:

- Reduce skewness in marginal distributions
- Linearize the relationship between variables
- Stabilize variance across the range of measurements
- Have biological interpretability



(a) Effect of log transformations on correlation

(b) Effect of square root transformations on correlation

Figure 12: **Effect of transformations on Pearson correlation**

The figure illustrates how different transformations can dramatically affect Pearson correlation values for the same underlying relationship. This transformation-dependence is a fundamental limitation when analyzing biological data, where the "true" scale of measurement is often unknown.

2.8 Spearman's Rank Correlation: Beyond Linear Relationships

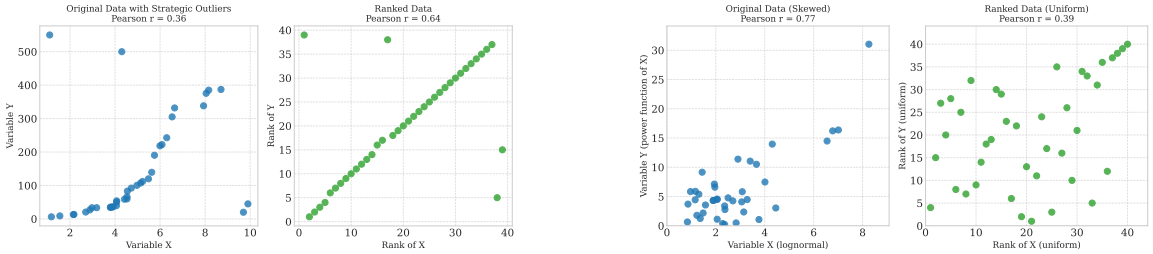
Spearman's rank correlation provides an elegant solution to the transformation challenge by focusing on monotonicity rather than linearity. By replacing values with their ranks, Spearman correlation becomes invariant to any strictly increasing transformation.

2.8.1 From Values to Ranks

Spearman's rank correlation coefficient (ρ_s) is calculated by applying Pearson's formula to the ranks of the data rather than the actual values:

$$\rho_s = \rho(\text{rank}(X), \text{rank}(Y)) = \frac{\sum_{i=1}^{n}(R(x_i) - \overline{R(x)})(R(y_i) - \overline{R(y)})}{\sqrt{\sum_{i=1}^{n}(R(x_i) - \overline{R(x)})^2 \sum_{i=1}^{n}(R(y_i) - \overline{R(y)})^2}}$$

where $R(x_i)$ and $R(y_i)$ are the ranks of observations x_i and y_i and $\overline{R(x)}$ and $\overline{R(y)}$ are the mean of the ranks.



(a) Spearman reveals better a non linear relationship

(b) Spearman is insensitive to outliers

Figure 13: **Spearman correlation as Pearson correlation of ranks**

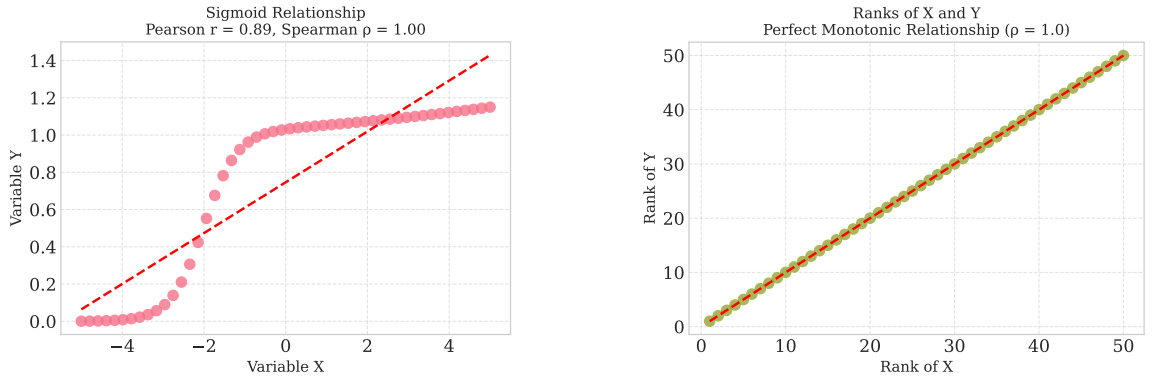
2.8.2 Transformation Invariance: A Key Advantage

Spearman correlation has several key advantages over Pearson correlation:

- If $Y = f(X)$ where f is any strictly increasing function, then $\rho_s(X, Y) = 1$ if decreasing, $\rho_s(X, Y) = -1$
- The same Spearman correlation value is obtained regardless of whether you analyze raw data, log-transformed data, or any other monotonic transformation
- This invariance is especially valuable when the "natural" scale of biological measurements is unknown or when comparing measurements across different experimental platforms

The most significant advantage of Spearman correlation is its invariance to monotonic transformations. This property eliminates the need to identify the "correct" scale of measurement

In the figure above, a sigmoid relationship appears non-linear on the original scale.



(a) Computation of Spearman and Pearson correlation

(b) The ranks of the same data scattered

Figure 14: **Spearman's measure better reveal the strong correlation**

While Pearson correlation underestimates the strength of this relationship ($\rho \approx 0.89$), Spearman correctly identifies it as perfectly correlated ($\rho_s = 1.0$) because the ranks perfectly align. No transformation is needed to reveal this strong monotonic relationship.

2.8.3 When to Use Spearman Correlation in Biological Data Analysis

Spearman correlation is particularly valuable when:

- Data contains outliers that might disproportionately influence Pearson correlation
- Different measurement techniques or platforms are being compared
- The biological mechanism suggests a monotonic but potentially non-linear relationship
- You want to avoid making assumptions about the "correct" scale of measurement

By focusing on the ranks rather than the values themselves, Spearman correlation often reveals biological relationships without requiring subjective decisions about data transformation.