Machine Learning for Genomics

Mathematical Foundations: Statistics and Probability in Biological Data

Sarwan Ali

Department of Computer Science
Georgia State University

Mathematical Foundations for Genomics

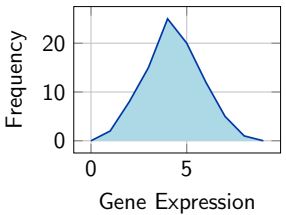
Today's Learning Journey

1. Introduction to Biological Data Statistics
2. Probability Theory Fundamentals
3. Descriptive Statistics
4. Statistical Inference
5. Confidence Intervals and Estimation
6. Machine Learning Fundamentals
7. Machine Learning for Genomics
8. Applications in Genomics

Why Statistics Matter in Genomics

Biological data is inherently noisy and uncertain:

- Measurement errors in sequencing
- Biological variation between individuals
- Technical replicates vs biological replicates
- Missing data and dropout events
- Multiple hypothesis testing challenges



Key Insight

Statistical methods help us distinguish **signal** from **noise** in biological data.

Types of Biological Data

Continuous Data:

- Gene expression levels (RNA-seq)
- Protein concentrations
- Methylation levels
- Copy number variations

Discrete Data:

- SNP genotypes (0, 1, 2)
- Read counts
- Cell types (categorical)
- Mutation presence/absence

High-Dimensional
Thousands of genes
Few samples ($n << p$)

Sparse
Many zero values
Dropout events

Heterogeneous
Different data types
Batch effects

Probability Distributions in Biology

Normal Distribution:

- Gene expression after log-transformation
- Phenotypic measurements
- Measurement errors

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

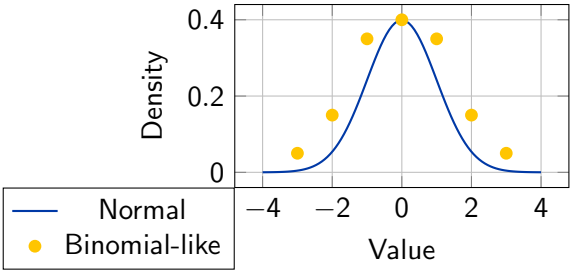
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Binomial Distribution:

- Allele frequencies
- Success/failure in experiments

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

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$



Poisson Distribution:

- RNA-seq read counts
- Mutation counts

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

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Bayes' Theorem in Genomics

Bayes' Theorem

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)}$$

Where:

- $P(H|E)$: Posterior probability (what we want to know)
- $P(E|H)$: Likelihood (probability of evidence given hypothesis)
- $P(H)$: Prior probability (initial belief)
- $P(E)$: Marginal probability (normalizing constant)

Genomics Example: Disease Risk Prediction

- H: Patient has disease
- E: Genetic variant is present
- $P(H)$: Disease prevalence in population
- $P(E|H)$: Probability of variant given disease
- $P(H|E)$: Risk of disease given variant presence

Conditional Probability and Independence

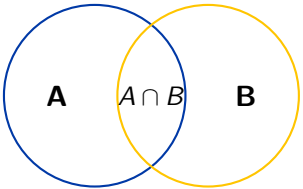
Conditional Probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

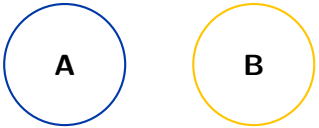
Independence: Two events A and B are independent if:

$$P(A|B) = P(A)$$

$$P(A \cap B) = P(A) \cdot P(B)$$



Dependent Events



Independent Events

Biological Examples:

- Linkage disequilibrium (non-independence)
- Hardy-Weinberg equilibrium assumptions
- Gene co-expression networks

Measures of Central Tendency

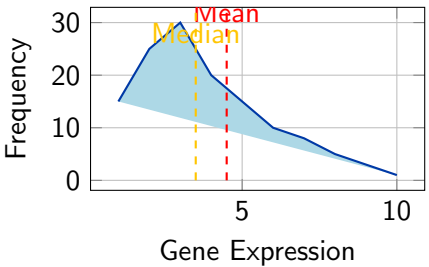
Mean (Arithmetic):

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

Median: Middle value when data is ordered
Mode: Most frequently occurring value

When to use which?

- Mean: Normal distributions
- Median: Skewed data, outliers
- Mode: Categorical data



Note: In genomics, data is often log-transformed to make it more normal.

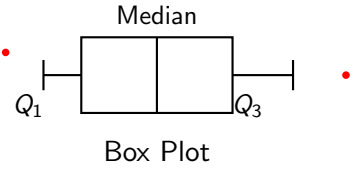
Measures of Variability

Variance:

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

Standard Deviation:

$$\sigma = \sqrt{\sigma^2}$$

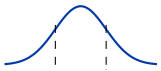


Range:

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

Interquartile Range (IQR):

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



Coefficient of Variation:

$$CV = \frac{\sigma}{\mu} \times 100\%$$

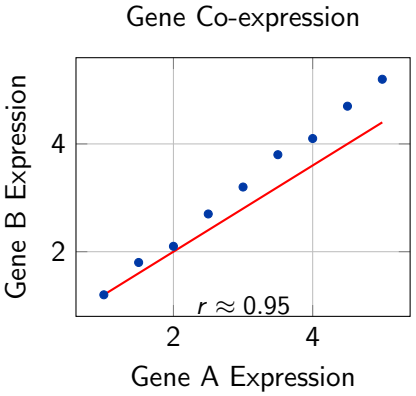
Correlation and Covariance

Covariance:

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

Pearson Correlation: $r = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$
Spearman Correlation: Correlation of ranks (non-parametric)
Interpretation:

- $r = 1$: Perfect positive correlation
- $r = 0$: No linear correlation
- $r = -1$: Perfect negative correlation



Genomics Application

Gene co-expression networks use correlation to identify functionally related genes.

Hypothesis Testing Framework

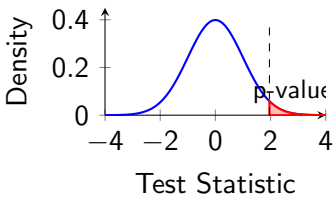
The Process:

1. **Null Hypothesis** (H_0): No effect/difference
2. **Alternative Hypothesis** (H_1): There is an effect
3. Choose significance level ($\alpha = 0.05$)
4. Calculate test statistic
5. Determine p-value
6. Make decision: Reject or fail to reject H_0

	H_0 True	H_0 False
Reject H_0	Type I	Correct
Fail to Reject	Correct	Type II

Types of Errors:

- Type I Error: False positive (α)
- Type II Error: False negative (β)
- Power: $1 - \beta$



Common Statistical Tests in Genomics

t-tests:

- One-sample: $H_0 : \mu = \mu_0$
- Two-sample: $H_0 : \mu_1 = \mu_2$
- Paired: Before/after treatment

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Chi-square test:

- Goodness of fit
- Independence (contingency tables)
- Hardy-Weinberg equilibrium

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

ANOVA: Compare multiple groups

$$F = \frac{\text{MS}_{\text{between}}}{\text{MS}_{\text{within}}}$$

Non-parametric tests:

- Mann-Whitney U test
- Wilcoxon signed-rank test
- Kruskal-Wallis test

Example

Testing differential gene expression between cancer and normal samples using t-test.

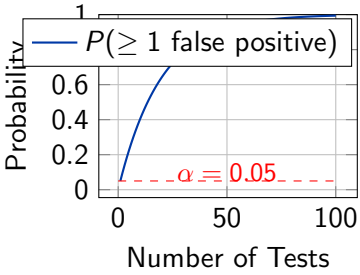
Multiple Testing Problem

The Problem: When testing thousands of genes simultaneously:

$$P(\text{at least one false positive}) = 1 - (1 - \alpha)^m$$

For $m = 20,000$ genes and $\alpha = 0.05$:

$$P \approx 1 - 0.95^{20000} \approx 1$$



Solutions:

- **Bonferroni**: $\alpha_{adj} = \frac{\alpha}{m}$
- **Benjamini-Hochberg (FDR)**: Control false discovery rate
- **q-values**: Bayesian approach to FDR

Method	Threshold
Bonferroni	2.5×10^{-6}
FDR (5%)	Variable

Confidence Intervals

Definition: A confidence interval provides a range of plausible values for a parameter.

For a mean (known σ):

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

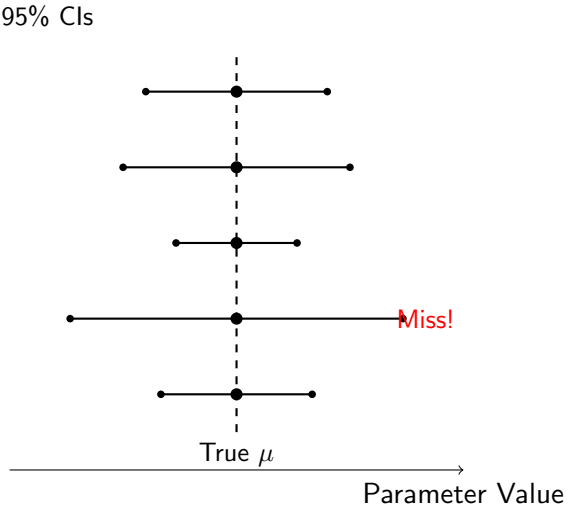
For a mean (unknown σ):

$$CI = \bar{x} \pm t_{\alpha/2,df}\frac{s}{\sqrt{n}}$$

For a proportion:

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

Interpretation: 95% CI means that if we repeated the study many times, 95% of the intervals would contain the true parameter.



Bootstrap and Resampling Methods

Bootstrap Principle:

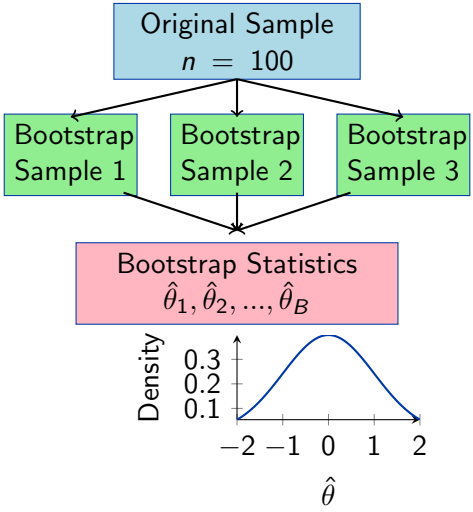
1. Resample with replacement from original data
2. Calculate statistic of interest
3. Repeat many times (e.g., 1000)
4. Use distribution of statistics for inference

Advantages:

- No distributional assumptions
- Works for complex statistics
- Provides uncertainty estimates

Genomics Applications:

- Gene set enrichment analysis
- Phylogenetic tree confidence
- Machine learning model validation



Machine Learning Paradigms

Supervised Learning:

- **Classification**: Predict discrete outcomes
- **Regression**: Predict continuous outcomes
- Examples: Disease diagnosis, gene expression prediction

Unsupervised Learning:

- **Clustering**: Group similar samples
- **Dimensionality Reduction**: Reduce feature space
- Examples: Cell type identification, pathway analysis

Semi-supervised Learning:

- Combines labeled and unlabeled data
- Useful when labels are expensive to obtain

Supervised Learning
Input: (X, Y)
Output: $f : X \to Y$

Unsupervised Learning
Input: X
Output: Hidden patterns

Semi-supervised
Input: $(X_l, Y_l), X_u$
Output: $f : X \to Y$

Bias-Variance Tradeoff

Decomposition of Prediction Error:

$$\text{Error} = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$

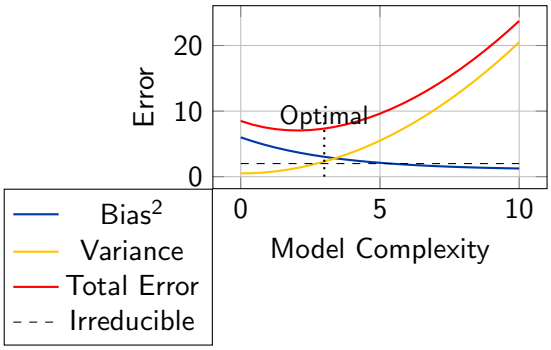
Bias:

- Error from oversimplifying assumptions
- High bias $\rightarrow$ underfitting
- Example: Linear model for nonlinear data

Variance:

- Error from sensitivity to training data
- High variance $\rightarrow$ overfitting
- Example: Very deep decision trees

Goal: Find optimal balance between bias and variance



Cross-Validation

K-Fold Cross-Validation:

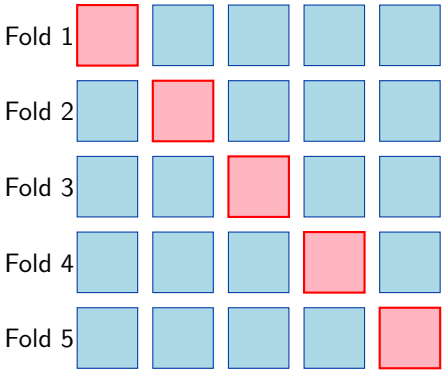
1. Split data into k folds
2. Train on $k-1$ folds
3. Test on remaining fold
4. Repeat for all folds
5. Average performance across folds

Leave-One-Out CV (LOOCV):

- Special case where $k = n$
- Maximum use of data
- Computationally expensive

Stratified CV:

- Maintains class proportions in each fold
- Important for imbalanced datasets



Genomic Data Characteristics

High-Dimensional Data:

- Thousands of genes, few samples
- $p \gg n$ problem
- Curse of dimensionality

Sparsity:

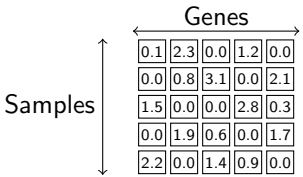
- Many zero values in gene expression
- Dropout events in single-cell data
- Sparse regulatory networks

Heterogeneity:

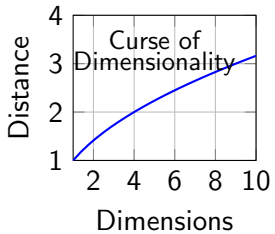
- Batch effects, Different cell types
- Technical vs biological variation

Noise:

- Measurement errors
- Biological stochasticity
- Systematic biases



Gene Expression Matrix



Feature Selection and Dimensionality Reduction

Feature Selection Methods:

- **Filter**: Statistical tests (t-test, chi-square)
- **Wrapper**: Forward/backward selection
- **Embedded**: LASSO, Ridge regression

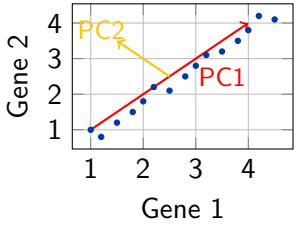
Principal Component Analysis (PCA):

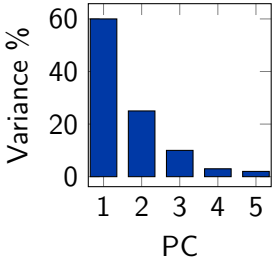
$\text{PC}_i = \sum_{j=1}^{p} w_{ij} X_j$

- Finds directions of maximum variance
- Linear transformation
- Orthogonal components

t-SNE:

- Non-linear dimensionality reduction
- Preserves local structure
- Good for visualization





Classification Algorithms

Logistic Regression:

$$P(Y = 1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\ldots+\beta_p X_p)}}$$

Support Vector Machines (SVM):

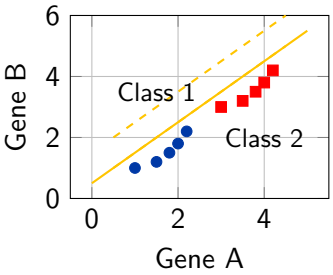
- Finds optimal separating hyperplane
- Kernel trick for non-linear boundaries
- Robust to outliers

Random Forest:

- Ensemble of decision trees
- Bootstrap aggregating (bagging)
- Feature importance measures

Neural Networks:

- Deep learning architectures
- Automatic feature learning
- Requires large datasets



Clustering Algorithms

K-means Clustering:

1. Initialize k cluster centers
2. Assign points to nearest center
3. Update centers to cluster means
4. Repeat until convergence

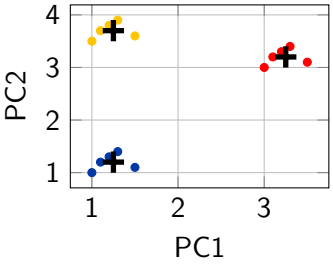
Minimize: $\sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2$

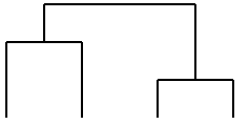
Hierarchical Clustering:

- Agglomerative (bottom-up)
- Divisive (top-down)
- Produces dendrogram
- No need to specify k

Applications:

- Cell type identification
- Gene co-expression modules
- Sample subgroups





Dendrogram

Model Evaluation Metrics

Classification Metrics:

	Predicted +	**Predicted -**
Actual +	TP	FN
Actual -	FP	TN

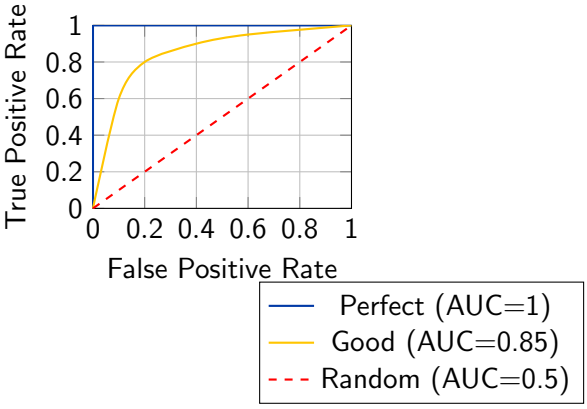
Accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$

Precision: $\frac{TP}{TP+FP}$

Recall (Sensitivity): $\frac{TP}{TP+FN}$

Specificity: $\frac{TN}{TN+FP}$

F1-Score: $\frac{2\times \text{Precision}\times \text{Recall}}{\text{Precision}+\text{Recall}}$



Regression Metrics:

- **MSE:** $\frac{1}{n}\sum(y_i - \hat{y}_i)^2$, **RMSE:** $\sqrt{MSE}$
- **MAE:** $\frac{1}{n}\sum|y_i - \hat{y}_i|$, **$R^2$:** $1 - \frac{SS_{res}}{SS_{tot}}$

Gene Expression Analysis

Differential Expression Analysis:

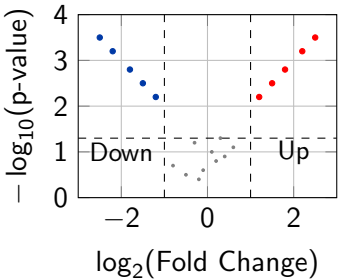
- Compare expression between conditions
- Statistical tests (t-test, limma, DESeq2)
- Multiple testing correction
- Effect size interpretation

Gene Set Enrichment Analysis (GSEA):

- Identify enriched pathways
- Rank-based approach
- Functional interpretation

Co-expression Networks:

- WGCNA (Weighted Gene Co-expression Network Analysis)
- Module identification
- Hub gene detection



Single-Cell Genomics

Challenges:

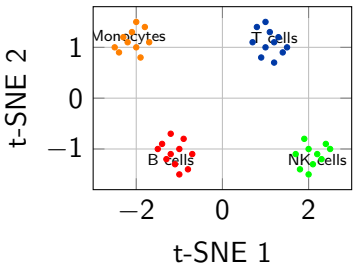
- High sparsity (dropout events)
- Technical noise
- Batch effects
- Computational scalability

Preprocessing:

- Quality control filtering
- Normalization (CPM, TPM, scran)
- Batch correction (ComBat, Harmony)
- Feature selection

Analysis Steps:

1. Dimensionality reduction (PCA, t-SNE, UMAP)
2. Clustering (Louvain, Leiden)
3. Cell type annotation
4. Trajectory analysis, Differential expression



Genomic Variant Analysis

Types of Variants:

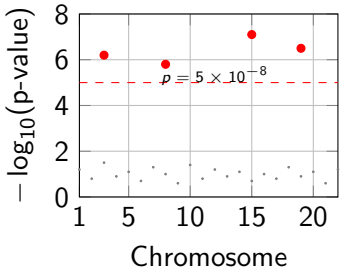
- Single Nucleotide Polymorphisms (SNPs)
- Insertions/Deletions (InDels)
- Copy Number Variations (CNVs)
- Structural Variants (SVs)

Variant Calling Pipeline:

1. Read alignment (BWA, Bowtie2)
2. Variant calling (GATK, FreeBayes)
3. Quality filtering, Annotation (VEP, ANNOVAR)
4. Functional impact prediction

Population Genetics:

- Allele frequency analysis
- Hardy-Weinberg equilibrium
- Linkage disequilibrium
- GWAS (Genome-Wide Association Studies)



Phylogenetic Analysis

Sequence Alignment:

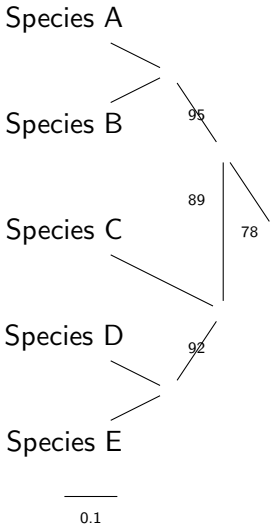
- Multiple sequence alignment (MSA)
- Tools: MUSCLE, ClustalW, MAFFT
- Gap penalties and scoring matrices

Tree Construction Methods:

- Distance-based: UPGMA, Neighbor-joining
- Character-based: Maximum parsimony
- Model-based: Maximum likelihood, Bayesian

Applications:

- Species relationships
- Evolutionary history
- Pathogen tracking
- Horizontal gene transfer



Questions?

 Thank You! 

Contact: sali85@student.gsu.edu