Extended Syllabus

This course covers probability theory and statistical methods essential for analyzing biological data. Topics range from fundamental probability concepts to advanced statistical modeling techniques, with a focus on applications in biological systems.

1 Lecture 1: Probability Theory Foundations

1.1 From Deterministic to Random Variables

Introduction to the shift from deterministic mathematical relationships to probabilistic frameworks in biological systems. Understanding randomness as inherent to biological processes rather than merely measurement uncertainty.

1.2 Formal Definition of Random Variables

Understanding random variables as functions mapping outcomes from a sample space to real numbers, with measurability conditions ensuring that probabilities can be meaningfully assigned to sets of outcomes.

1.3 Support and Types of Random Variables

Characterization of discrete vs. continuous random variables and their supports. Discrete random variables take countable values while continuous random variables take values from intervals.

1.4 Probability Mass and Density Functions

Mathematical representations of probability distributions through PMFs (discrete) and PDFs (continuous), defining how probability is distributed across possible outcomes.

1.5 Cumulative Distribution Function

Properties and applications of CDFs in representing probability distributions, providing a unified framework for both discrete and continuous random variables.

1.6 Quantile Function

Understanding inverse CDFs and their applications in finding percentiles and generating random samples from distributions.

1.7 Expectation

Understanding average behavior through expectations, including properties like linearity and additivity, and their interpretation in biological contexts.

1.8 Variance and Moments

Quantifying spread and shape of distributions through variance and higher-order moments, characterizing different aspects of randomness in biological data.

1.9 Central Moments

Understanding the variance as the second central moment and properties of higher central moments that describe distribution shapes.

1.10 Parametrization of Distributions

Mathematical approaches to describing distributions through parameters and their biological interpretations, connecting statistical formalism with biological meaning.

1.11 Common Distributions in Biology

Applications of distributions (normal, log-normal, gamma, beta, Poisson, negative binomial) in modeling various biological phenomena from cell counts to gene expression.

1.12 Operations on Random Variables

Understanding how to combine random variables through convolution, mixture, and transformation operations to model complex biological processes.

2 Lecture 2: Maximum Likelihood Estimation and Measuring Association

2.1 The Likelihood Function

Understanding the likelihood as a bridge between data and theoretical distributions, reversing the usual probabilistic perspective to estimate parameters from observations.

2.2 Log-Likelihood Function

Benefits of working with log-likelihood and its mathematical properties, transforming products into sums for computational advantage.

2.3 Maximum Likelihood Principle

Choosing parameter values that maximize the likelihood (or log-likelihood) of observed data, providing a principled approach to parameter estimation.

2.4 Finding Maximum Likelihood Estimates

Analytical and numerical approaches to finding MLEs, with examples showing how to derive closed-form solutions when possible.

2.5 Bias in MLE Variance Estimation

Understanding why the MLE for variance uses n rather than n-1 in the denominator, revealing subtle aspects of estimation theory.

2.6 Numerical Methods for MLE

Approaches like Newton-Raphson and gradient descent for cases without analytical solutions, necessary for many complex biological models.

2.7 Scatter Plots and Data Visualization

Using scatter plots to visualize relationships between variables, revealing patterns that summary statistics might miss.

2.8 Proteomics Data Analysis

Application of association measures to understand protein-protein relationships in complex biological datasets.

2.9 Covariance and Correlation

Measuring linear relationships between variables through covariance and Pearson correlation, quantifying how variables change together.

2.10 Properties and Limitations of Correlation

Understanding correlation's mathematical properties and common misconceptions, including the distinction between correlation and causation.

2.11 Geometric Interpretation of Correlation

Correlation as the cosine of the angle between centered data vectors, providing geometric intuition for this statistical measure.

2.12 Data Transformations and Spearman's Rank Correlation

Beyond linear relationships using transformations and rank-based methods, capturing monotonic associations regardless of linearity.

3 Lecture 3: Joint and Bivariate Distributions

3.1 Joint Probability Distributions

Mathematical framework for understanding relationships between multiple random variables, extending univariate distributions to multiple dimensions.

3.2 Marginal Distributions

Recovering information about individual variables from joint distributions through integration or summation.

3.3 Conditional Probability and Dependencies

Understanding how variables influence each other through conditional distributions, revealing complex relationships in biological systems.

3.4 Statistical Independence

Properties and implications when random variables provide no information about each other, and methods to test for independence.

3.5 Bivariate Normal Distribution

Properties and applications of the multivariate extension of the normal distribution, a fundamental model for correlated biological measurements.

3.6 Geometric Interpretation of Bivariate Distributions

Understanding level sets as ellipses and their properties, providing visual intuition for correlation structure.

3.7 Conditioning and Marginalization in Bivariate Normal

Special properties of conditional and marginal distributions for normal variables, showing how they maintain normality under these operations.

3.8 Histogram 2D

Using two-dimensional histograms to visualize and estimate joint distributions from data points.

3.9 Parametric Estimation of Joint Distributions

Fitting parametric models (especially bivariate normal) to multivariate data using maximum likelihood.

3.10 Kernel Density Estimation in Two Dimensions

Non-parametric approaches to estimating joint distributions when parametric forms are unsuitable.

3.11 Bandwidth Selection in KDE

Methods for determining appropriate smoothing parameters in kernel density estimation to balance bias and variance.

3.12 Mutual Information

Measuring general statistical dependencies between variables using information theory concepts, capturing both linear and non-linear relationships.

4 Lecture 4: Statistical Tests and Regression

4.1 Introduction to Statistical Testing

Understanding the fundamental challenge of drawing conclusions from limited data samples and the framework for distinguishing between true effects and random variation.

4.2 Sampling Distributions

Statistical properties of sample statistics (like means) as random variables, including how sample size affects the precision of estimates.

4.3 The Central Limit Theorem

How the sampling distribution of the mean approaches a normal distribution regardless of the original distribution shape, given sufficient sample size.

4.4 Fisher's Approach to Significance Testing

Testing compatibility between observed data and a null hypothesis through p-values and their interpretation as probability of observing data as extreme by chance.

4.5 One-Sample Tests

Methods for testing whether a sample mean differs from a reference value, including z-tests (known variance) and t-tests (unknown variance).

4.6 Two-Sample Tests

Statistical approaches for comparing means between independent groups, including Welch's t-test for unequal variances.

4.7 Paired Sample Tests

Testing methods for dependent observations that account for within-subject correlation to increase statistical power.

4.8 Type I and Type II Errors

Understanding the different types of mistakes in hypothesis testing and their implications for scientific inference.

4.9 The Neyman-Pearson Framework

Extending Fisher's approach to decision-making with controlled error rates through significance levels and power.

4.10 Multiple Testing Problem

How conducting many tests simultaneously inflates the false discovery rate and requires correction.

4.11 Family-Wise Error Rate (FWER)

Controlling the probability of making at least one false discovery, including the Bonferroni correction.

4.12 False Discovery Rate (FDR)

A more practical approach for high-throughput biology that controls the proportion of false positives among discoveries, including the Benjamini-Hochberg procedure.

5 Lecture 5: Simple Linear Regression

5.1 From Joint Distributions to Regression Models

Understanding regression as modeling the conditional distribution of a response variable given predictors.

5.2 The Linear Model as a Probabilistic Framework

Interpreting regression as a normal distribution whose mean is a linear function of predictors while variance remains constant.

5.3 Conditional Expectations in Regression

The theoretical foundation of regression as modeling E[Y—X], with clear interpretation of parameters.

5.4 Maximum Likelihood and Least Squares Estimation

Statistical justification for parameter estimation through minimizing sum of squared residuals.

5.5 Normal Equations

Mathematical foundation for finding optimal regression parameters through matrix algebra.

5.6 Properties of Regression Estimates

Unbiasedness, consistency, efficiency, and normality of coefficient estimates under standard assumptions.

5.7 Hypothesis Testing for Regression Coefficients

Statistical framework for determining significance of predictors through t-tests.

5.8 Confidence Intervals for Regression Parameters

Quantifying uncertainty in parameter estimates using t-distributions.

5.9 Prediction vs. Confidence Intervals

Distinction between intervals for mean response (confidence) versus individual observations (prediction).

5.10 Geometric Interpretation of Regression

Understanding regression as fitting a line that minimizes squared vertical distances to data points.

5.11 Standardized Coefficients

Comparing predictor importance by adjusting for different scales of measurement.

5.12 Practical Applications in Biological Data Analysis

Using regression for both descriptive modeling and predictive applications in biological contexts.

6 Lecture 6: Multiple Linear Regression

6.1 Extending to Multiple Predictors

Moving from simple to multiple regression by incorporating several explanatory variables simultaneously.

6.2 Geometric Representation of Multiple Regression

Visualizing multiple regression as fitting a hyperplane in multidimensional space.

6.3 Matrix Formulation of Regression

Using linear algebra to express the regression model concisely

6.4 The Design Matrix

Understanding how predictors are organized into the X matrix and its role in parameter estimation.

6.5 Parameter Estimation in Multiple Regression

Generalizing least squares to multiple dimensions through matrix operations.

6.6 Normal Equations in Matrix Form

Finding optimal parameters by solving normal equations in matrix notation, leading to efficient computation.

6.7 Linear Regression in Observation Space

Viewing regression as orthogonal projection of the response vector onto the column space of the design matrix.

6.8 Interpreting Multiple Regression Coefficients

Understanding coefficients as conditional effects that control for other variables in the model.

6.9 Standardized Coefficients

Using standardized variables to compare relative importance of predictors measured on different scales.

6.10 Model Comparison with R^2 and Adjusted R^2

Assessing model fit while accounting for model complexity through the adjusted coefficient of determination.

6.11 Comparing Nested Models

Using likelihood ratio tests to determine whether additional predictors significantly improve model fit.

6.12 ANOVA as a Regression Framework

Understanding how analysis of variance for categorical predictors fits within the linear regression model.

7 Lecture 7: Review week

No extra content on this ”lecture”, just a review of the previous lectures.

8 Lecture 8: Logistic Regression and Generalized Linear Models

8.1 Generalized Linear Models Framework

Extension of linear regression to handle various types of response variables through three components: random component, systematic component, and link function.

8.2 Linear Predictor and Link Functions

Mathematical transformation connecting the linear combination of predictors to the expected response, allowing linear models to be applied to non-linear relationships.

8.3 Exponential Family Distributions

Understanding the theoretical foundation of GLMs with distributions like Normal, Binomial, Poisson, and Gamma, each with canonical link functions.

8.4 Poisson Regression for Count Data

Modeling discrete count outcomes using the log link, allowing for multiplicative effects and naturally handling increasing variance with the mean.

8.5 Classification vs. Regression Problems

The fundamental distinction between modeling continuous outcomes (regression) and categorical outcomes (classification).

8.6 Binary Response and Logit Link

Using the logistic function to transform unbounded linear predictors to bounded probabilities between 0 and 1.

8.7 The Logistic Regression Model

Probabilistic framework for binary classification, connecting predictor variables to the probability of class membership through the logit link.

8.8 Parameter Interpretation in Logistic Regression

Interpreting coefficients in terms of odds ratios and log-odds for meaningful biological insights.

8.9 Maximum Likelihood Estimation in GLMs

Computational methods for fitting GLM parameters when closed-form solutions aren't available.

8.10 From Probabilities to Binary Decisions

The use of thresholds to convert probabilistic predictions into categorical classifications.

8.11 Confusion Matrix and Performance Metrics

Evaluating classification performance through accuracy, sensitivity, specificity, and precision.

8.12 ROC Curves and AUC

Threshold-independent evaluation of classifier performance across all possible operating points.

9 Lecture 9: Regularization and Maximum a Posteriori Estimation

9.1 The James-Stein Phenomenon

Theoretical discovery showing how biased estimators can outperform unbiased ones when estimating multiple parameters simultaneously.

9.2 Underdetermined Systems in Regression

Challenges when the number of parameters approaches or exceeds the number of observations, leading to infinite solutions.

9.3 Overfitting and Poor Generalization

The problem of models capturing noise in training data rather than underlying patterns, and methods to detect it.

9.4 Cross-Validation Methods

Techniques to assess model generalization performance by repeatedly partitioning data into training and validation sets.

9.5 The Bias-Variance Trade-off

Theoretical foundation explaining why introducing some bias through regularization can improve overall prediction performance.

9.6 Ridge Regression (L2 Regularization)

Shrinking regression coefficients toward zero by adding a penalty term proportional to the sum of squared coefficients.

9.7 Regularization Parameter Selection

Methods to determine the optimal strength of regularization using cross-validation and the regularization path.

9.8 Bayesian Perspective on Regularization

Viewing regularization as incorporating prior beliefs about parameter values through Bayes' rule.

9.9 Maximum a Posteriori (MAP) Estimation

A bridge between frequentist and Bayesian approaches that maximizes the posterior probability.

9.10 Gaussian Priors and Ridge Regression

The equivalence between ridge regression and MAP estimation with zero-centered Gaussian priors.

9.11 Lasso Regression (L1 Regularization)

Feature selection through a penalty term proportional to the absolute values of coefficients, producing sparse solutions.

9.12 Laplace Priors and Lasso

The Bayesian interpretation of lasso regression as MAP estimation with Laplace (double-exponential) priors.

10 Lecture 10: Multivariate Normal and Principal Component Analysis

10.1 The Multivariate Normal Distribution

Extending normal distributions to multiple dimensions with mean vector and covariance matrix parameters.

10.2 Covariance Matrix Structure

Encoding relationships between variables through variance (diagonal) and covariance (off-diagonal) elements.

10.3 Geometric Interpretation of Multivariate Normal

Understanding level sets as ellipsoids whose orientation and axis lengths are determined by the covariance matrix.

10.4 Sample Mean Vector and Covariance Matrix

Estimating multivariate normal parameters from data and challenges in high-dimensional settings.

10.5 Eigendecomposition of the Covariance Matrix

Finding principal directions of variation through eigenvalues and eigenvectors.

10.6 Singular Value Decomposition (SVD)

Alternative matrix factorization providing computational advantages for high-dimensional data.

10.7 Principal Component Analysis (PCA)

Dimensionality reduction technique identifying orthogonal axes of maximum variance.

10.8 Geometric Interpretation of PCA

Viewing PCA as a rotation of the coordinate system to align with directions of maximum variation.

10.9 Dimension Reduction Through PCA

Approximating high-dimensional data with lower-dimensional representations while preserving maximum variance.

10.10 The Scree Plot and Variance Explained

Methods to determine how many principal components to retain based on eigenvalue patterns.

10.11 Loading Analysis and Component Interpretation

Techniques to interpret the biological meaning of principal components through variable contributions.

10.12 Biplots and Visualization

Simultaneous visualization of observations and variables in principal component space.

11 Lecture 11: Dimensionality Reduction and Clustering

11.1 PCA for Data Visualization

Using principal components to create informative visualizations of high-dimensional data.

11.2 Data Shape Characterization

Inferring the structure of high-dimensional data through patterns in eigenvalue distributions.

11.3 Clustering as Discrete Data Modeling

Identifying natural groupings in data by partitioning observations into distinct clusters.

11.4 Clustering as Optimization

Formal definition of clustering as finding the partition that optimizes an objective function measuring cluster quality.

11.5 Distance and Similarity Measures

Various metrics for quantifying relationships between observations, including Euclidean, Manhattan, correlation-based, and cosine distances.

11.6 Partition-Based vs. Hierarchical Clustering

Contrasting approaches that either create flat partitions or build nested hierarchies of clusters.

11.7 K-Means Algorithm

Iterative procedure alternating between assigning points to nearest centroids and updating centroids.

11.8 Voronoi Tessellation in K-Means

Geometric interpretation of K-means as partitioning the feature space into regions.

11.9 Determining Optimal Cluster Numbers

Methods like silhouette analysis to objectively select the appropriate number of clusters.

11.10 Agglomerative Hierarchical Clustering

Bottom-up approach that progressively merges similar clusters to build a hierarchical structure.

11.11 Dendrogram Interpretation

Reading tree-like visualizations of hierarchical clustering results and extracting different partitions.

11.12 Linkage Methods

Different approaches to define inter-cluster distances in hierarchical clustering, including single, complete, average, and Ward's methods.

12 Lecture 12: Permutation Tests, the Jackknife, and the Bootstrap

12.1 Computational Statistics

Introduction to the computational revolution in statistics, driven by exponential increases in computational power. Discussion of how computational approaches allow tackling problems through simulation and algorithms, handling violations of standard assumptions, and working with complex data structures.

12.2 Permutation Tests

Permutation tests as a distribution-free approach to hypothesis testing.

12.3 Use Cases of Permutation Tests

Applications of permutation tests in scenarios such as small sample sizes, skewed data, outliers, and differing variances between groups. Relaxation of assumptions required by traditional tests.

12.4 The Jackknife

Introduction to the jackknife method as an early computational approach to statistics. Systematic leave-one-out procedure to understand the stability of a statistic. Steps include calculating jackknife replicates, estimating standard errors, and bias correction.

12.5 Estimating Standard Errors with the Jackknife

Using jackknife replicates to compute standard errors, measuring the fluctuation of a statistic when individual observations are removed. Ensures consistency with analytical formulas for standard errors.

12.6 Bias Estimation with the Jackknife

Estimating bias of a statistic using jackknife replicates and constructing bias-corrected estimates. Applications in ratio estimators, variance components, and other statistics with systematic bias.

12.7 Limitations of the Jackknife

Discussion of limitations, including non-smooth statistics, insufficient exploration of sampling distributions, and edge effects for extreme values. Motivates the development of more comprehensive resampling approaches.

12.8 The Bootstrap

Introduction to the bootstrap method as a paradigm shift in computational statistics. Resampling with replacement to simulate the sampling process and approximate the sampling distribution of a statistic.

12.9 Bootstrap Algorithm

Steps for generating bootstrap samples, calculating statistics for each sample, and using the empirical distribution to estimate standard errors, confidence intervals, and bias.

12.10 Bootstrap Confidence Intervals

Construction of confidence intervals using the percentile method, preserving range restrictions and accounting for asymmetry in the sampling distribution.

12.11 Parametric Bootstrap

Extension of the bootstrap method to small sample sizes by fitting a parametric model to the observed data and generating synthetic datasets from the fitted model. Applications in scenarios where nonparametric bootstrap may not capture population variability.