GCB533: STATISTICS FOR GENOMICS AND BIOMEDICAL INFORMATICS

Description

GCB533 is an introductory course in probability theory and statistical inference for graduate students in Genomics and Computational Biology. The goal of the course is to provide a foundation of basic concepts and tools as well as hands-on practice in their application to problems in genomics. At the completion of the course, students should have an intuitive understanding of basic probability and statistical inference and be prepared to select and execute appropriate statistical approaches in their future research.

The course will be divided into three sections. Part 1 will cover Probability Theory and will introduce key concepts, including probability distributions, density functions, random variables, expected values, correlation and covariance, the central limit theorem, the law of large numbers, and sampling distributions. This portion of the course provides the foundational base for understanding how to describe and quantify variation and covariation.

The second portion of the course will concern Statistical Inference and equips students to choose and execute an appropriate and rigorous means of assessing support for a hypothesis. Part 2 will introduce the fundamentals of Statistical Inference, including hypothesis testing, significance levels, family-wise error rate, false discovery rate, confidence intervals, maximum likelihood, and linear regression. Part 3 will cover more advanced topics in Statistical Inference, including the EM algorithm, bootstrap and randomization, decision theory, Bayesian inference, and Markov Chain Monte Carlo approaches, through a series of guest lectures.

Throughout the course, examples and exercises will utilize genomics problems and data sets to illustrate the application of each concept and approach to contexts relevant to students' dissertations and future research. Students will progressively become familiar with the programming language R throughout exercises.

This course will also contribute to training the students in Scientific Rigor and Reproducibility (SRR) by providing them with the necessary statistical tools for the design of rigorous, accurate, and reproducible experiments and analyses in genomics.

Intended Students

GCB533 is intended for first year students in the Genomics and Computational Biology graduate program. The course assumes experience and familiarity with mathematical concepts and notation through basic calculus but no prior instruction in statistics and probability.

Instructors

Pablo Cámara, pcamara@pennmedicine.upenn.edu
Laura Almasy, almasyl@pennmedicine.upenn.edu

Christopher Adams (TA), adamschr@pennmedicine.upenn.edu
William Bone (TA), wbone@pennmedicine.upenn.edu

Guest Lecturers: Jason Moore, Ben Voight, Marylyn Ritchie, Yoseph Barash, Mary Regina Boland, Ray Bai.

Course Format

Lecture, 3 contact hours per week, 2 days per week (Tuesdays and Thursdays 1:30-3:00 pm). Contact hours include in-class practical examples and exercises. Course to be held in the fall semester.

Location

Stemmler 416

Course Outline

PART 1. PROBABILITY THEORY

1.1. Probability, laws of addition and multiplication. [Aug 27] [Laura Almasy]
Definition of probability, relation to set theory, law of addition, conditional probability, law of multiplication.

1.2. Random variables. [Aug 29] [Laura Almasy]
Discrete random variables, probability distribution, cumulative probability distribution, continuous random variables, probability density function, median, percentiles, multivariate distributions, marginal probabilities.

1.3. Expected values. [Sept 3, 5] [Laura Almasy]
Definition of expected value, properties of the expected value (E(a+bX), E(E+Y), E(XY)), moment generating function, mean, variance, skewness, kurtosis.

Examples: computing genotype probabilities.

Homework assignment 1

1.4. Correlation and covariance. [Sept 10, 12] [Laura Almasy]
Definition of covariance and correlation.

Examples: Correlation among gene expression levels

Homework assignment 2

1.5. Binomial, Poisson, and exponential distributions. [Sept 17, 19] [Laura Almasy]
Probability of x successes in n repetitions of an experiment, properties of the binomial distribution, multinomial distribution, the Poisson distribution as a limit of the binomial distribution, properties of the Poisson distribution, distribution of waiting times, properties of exponential distribution.

Examples: Read counts, distribution of genotypes in a population.

Homework assignment 3

1.6. Normal distribution and the central limit theorem. [Sept 24] [Laura Almasy]
The central limit theorem, properties of the normal distribution

1.7. Sampling distributions: the chi2 and t distributions. [Sept 26] [Laura Almasy]
Definition of sampling distribution, the sampling distribution of the mean, sampling distribution of the variance, properties of the chi2 distribution, properties of the t distribution, limit of large samples.

Midterm review [Oct 1] [Laura Almasy]

Take home mid-term exam

PART 2. FUNDAMENTALS OF STATISTICAL INFERENCE

2.1. Hypothesis testing, tests of significance. [Oct 3, 8] [Pablo Camara]
Null hypothesis, level of significance, rejection region, one-tailed/two-tailed tests, t-test, goodness of fit.

Examples: Differences among populations, differential expression

Homework assignment 4

2.2. Multiple hypothesis testing. [Oct 15] [Pablo Camara]
Type I and type II errors, family-wise error rate, Bonferroni correction, false discovery rate, Benjamini-Hochberg procedure.

Examples: Multiple hypotheses correction in GWAS and differential expression studies

Homework assignment 5

2.3. Confidence intervals, point estimation. [Oct 17, 22] [Pablo Camara]
Concept of sufficient estimator, estimation of the mean, estimation of the variance.

Examples: Gene expression levels revisited, dependency on sample size.

Homework assignment 6

2.4. Maximum likelihood inference. [Oct 24, 29] [Pablo Camara]
Likelihood function, sampling distribution of the maximum likelihood function, Fisher information.

Homework assignment 7

2.5. Linear regression. [Oct 31, Nov 5] [Pablo Camara]
Univariate regression, squared error loss, Gauss-Markov theorem, multiple linear regression, generalized linear models.

Examples: Gene expression levels by genotype.

Homework assignment 8

PART 3. GUEST LECTURES ON STATISTICAL INFERENCE

2.6. Bootstrap and randomization. [Nov 7] [Jason Moore]
Introduction to the bootstrap method, relation to maximum likelihood

2.7. EM algorithm. [Nov 12, 19] [Ben Voight]
Introduction to the EM algorithm, example: two-component mixture model.

2.8. Statistical decision theory. [Nov 14, 21] [Marylyn Ritchie]
Expected prediction error, model selection and bias-variance tradeoff, curse of dimensionality, subset selection, ridge, lasso.

2.9. Bayesian inference. [Nov 26, Dec 5][Yoseph Barash]
Prior and posterior probability distributions, conjugate prior, predictive distribution.

2.10. MCMC approach to posterior sampling. [Dec 3] [Mary Boland / Ray Bai]
Gibbs sampling.

Final exam

Optional Readings

· M.G. Bulmer, *Principles of Statistics*, Dover Publications, 2nd edition (1967).

· T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2nd edition (2016): Chapters 2 and 8.

· D. S. Silva, Data Analysis: A Bayesian Tutorial, Oxford University Press (2012): Bayesian inference.

· A. Gelman, J.B. Carlin, D. B. Dunson, A. Vehtari, D.B. Rubin, *Bayesian Data Analysis*, Chapman & Hall CRC, 3rd edition (2013): Bayesian inference

· S. Holmes, W. Hube, *Modern Statistics for Modern Biology*, Cambridge University Press (2019).

Evaluation

30% homework + 10% participation in class + 25% mid-term take home exam + 35% final exam. The lowest homework grade will not be taken into account in the evaluation.