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, and Bayesian inference 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
Chi-Yun Wu (TA), chiyunwu@pennmedicine.upenn.edu
Guest Lecturers: Jason Moore, Ben Voight, Marylyn Ritchie, Yoseph Barash

Course Format

Video lectures to be watched asynchronously along with virtual in-class discussion and practical examples and exercises on Tuesdays from 1:30-3:00 pm. Course to be held in the fall semester.

Location

To be held virtually on Bluejeans

Course Outline

Orientation – meet the instructors, syllabus overview [Sept 1] [Laura, Pablo, Chi-Yun]

Part 1. Probability theory

1.1. Probability, random variables, expected values. [Video 1.1 & Sept 8] [Laura]
Definition of probability, relation to set theory, law of addition, conditional probability, law of multiplication.

1.2. Random variables. [Video 1.2 & Sept 8] [Laura]
Discrete random variables, probability distribution, cumulative probability distribution, continuous random variables, probability density function, median, percentiles, multivariate distributions, marginal probabilities.

1.3. Expected values. [Video 1.3 & Sept 8] [Laura]
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. [Video 1.4 & Sept 15] [Laura]
Definition of covariance and correlation.

Examples: Correlation among gene expression levels

Homework assignment 2

1.5. Binomial, Poisson, and exponential distributions. [Video 1.5 & Sept 22] [Laura]
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. Introduction to maximum likelihood.

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

Homework assignment 3

1.6. Normal distribution and the central limit theorem. [Video 1.6 & Sept 29] [Laura]
The central limit theorem, properties of the normal distribution

1.7. Sampling distributions: the chi2 and t distributions. [Video 1.7 & Oct 6] [Laura]
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.

Take home mid-term exam

PART 2. FUNDAMENTALS OF STATISTICAL INFERENCE

2.1. Hypothesis testing, tests of significance. [Video 2.1 & Oct 13, 20] [Pablo]
Null hypothesis, level of significance, rejection region, one-tailed/two-tailed tests, t-test, goodness of fit.

Examples: Differences among populations, differential expression

2.2. Multiple hypothesis testing. [Video 2.2 & Oct 20] [Pablo]
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 4

2.3. Confidence intervals, point estimation. [Video 2.3 & Oct 27] [Pablo]
Concept of sufficient estimator, estimation of the mean, estimation of the variance.

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

Homework assignment 5

2.4. Maximum likelihood inference. [Video 2.4 & Oct 27, Nov 3] [Pablo]
Likelihood function, sampling distribution of the maximum likelihood function, Fisher information.

Homework assignment 6

2.5. Linear regression. [Video 2.5 & Nov 10] [Pablo]
Univariate regression, squared error loss, Gauss-Markov theorem, multiple linear regression, generalized linear models.

Examples: Gene expression levels by genotype.

Homework assignment 7

PART 3. GUEST LECTURES ON STATISTICAL INFERENCE

3.1. Bootstrap and randomization. [Video 3.1 & Nov 17] [J. Moore]
Introduction to the bootstrap method, relation to maximum likelihood

3.2. EM algorithm. [Video 3.2 & Nov 24] [B. Voight]
Introduction to the EM algorithm, example: two-component mixture model.

3.3. Statistical decision theory. [Video 3.3 & Dec 1] [M. Ritchie]
Expected prediction error, model selection and bias-variance tradeoff, curse of dimensionality, subset selection, ridge, lasso.

3.4. Bayesian inference. [Video 3.4 & Dec 8] [Y. Barash]
Prior and posterior probability distributions, conjugate prior, predictive distribution.

Take home final exam

Optional Readings

· M.G. Bulmer, *Principles of Statistics*, Dover Publications, 2nd edition (1967). (A classic. The probability section of the course follows this text closely.)

· 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).

· R. Irizarry, *Introduction to Data Science,* https://rafalab.github.io/dsbook/. This textbook, available free online, is more computational than mathematical but it includes examples in R that may be useful for connecting concept to implementation.

Evaluation

45% homework + 25% mid-term take home exam + 30% final exam. The lowest homework grade will not be taken into account in the evaluation. Class participation is taken into consideration when a grade is on the on the border between two levels.