Wrapup

Christof Seiler

Stanford University, Spring 2016, Stats 205

Format Final Project

- You can write your final paper in the format you like: word, latex, R markdown document, etc.
- Number of pages: **12**
- Included in the 12 pages are:
  - Images, plots, tables, code snippets, pseudocode, diagrams, and proofs
- Excluded in the 12 pages are:
  - You should do your statistical analysis using R markdown and attach it to your paper
  - References
- Deadline: **June 3 by 11:59 PM**
- Zip all your files and rename it `YourFirstNameYourLastName` (for group projects add both names to filename)
- Upload your single zip file here: `https://dropitto.me/Stats205` with password `Stats205`

Course Evaluations Now Open

- Axess is now open to complete end-term course evaluations
- You may complete the evaluations until 08:00 AM on Monday, June 13
- You can find it on
  - **Stanford Axess**
  - **in Course and Section Evaluations**
  - **on the Student tab**
- Motivation:
  - If you complete all of your evaluations you will see your grades as soon as they have been submitted by the faculty (at the latest June 9)
  - Otherwise, grades will appear in Axess until June 14

When Is Something Nonparametric?

- Two categories

1. Distribution-free (NOT assumption-free)

- Sample $X_1, \ldots, X_n$ from F
- Statistic $T(X_1, \ldots, X_n)$
- T is distribution-free if
- T has **same distribution** for any F with some restrictions, for example:
  - Sign test: F any continuous distribution
  - Signed-rank test: F any symmetric continuous distribution

When to Use Distribution-Free Tests?

- When the "right" representation of the "center" is the median
- "Right" can mean data with outliers that cannot easily be removed
- Very small sample size (in large samples things become normal)
- With ordinal data (e.g. salary ranges) and ranked data (e.g. sports team ranking)
- In all other cases use parametric test because they are more powerful

When Is Something Nonparametric?

2. Nonparametric models (number of parameters can grow with $n \to \infty$)
   - Nonparametric regression
   - Bayesian nonparametrics
   - Wavelets
   - Graphons

When to Use Nonparametric Models?

- Very large sample size
- Low dimensional setting otherwise curse of dimensionality
- Complex relationship between predictor and response

Regression Fits in Both Categories

- Rank-based linear regression (distribution-free)
- Nonlinear regression (nonparametric model)

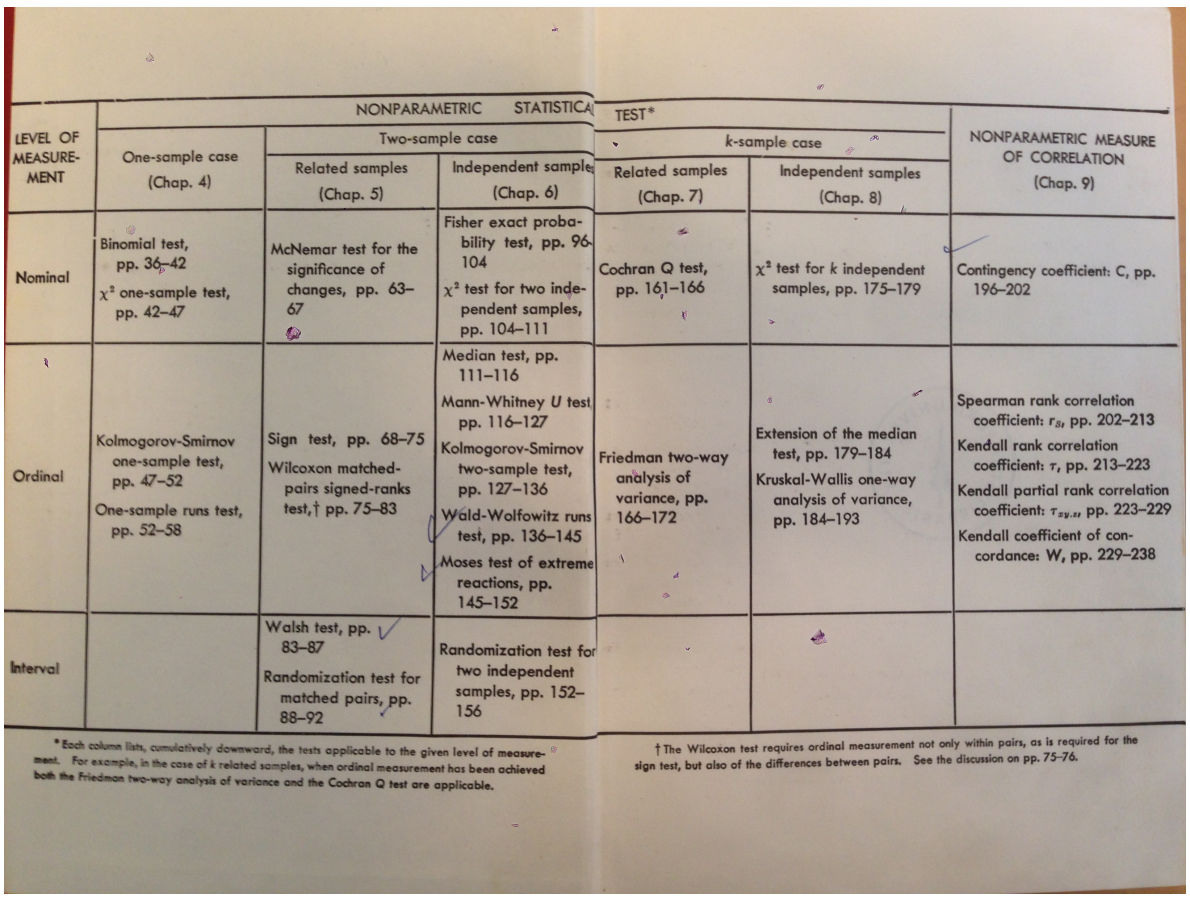
Learning Goals

1. The students will learn to apply methods and explain the statistical assumptions of **Monte Carlo simulations** for analytically intractable problems.

- We used Monte Carlo in most topics
  - permutation tests
  - bootstrap
  - rank-based methods
- As an alternative to asymptotics and complete enumerations

Learning Goals

2. The students will learn to apply methods and explain the statistical assumptions of **rank-based methods** for parameter estimation, confidence intervals, and hypothesis testing.



NONPARAMETRIC STATISTICAL TEST*

LEVEL OF MEASUREMENT	Two-sample case: One-sample case (Chap. 4)	Two-sample case: Related samples (Chap. 5)	Two-sample case: Independent samples (Chap. 6)	k-sample case: Related samples (Chap. 7)	k-sample case: Independent samples (Chap. 8)	NONPARAMETRIC MEASURE OF CORRELATION (Chap. 9)
Nominal	Binomial test, pp. 36–42 χ^2 one-sample test, pp. 42–47	McNemar test for the significance of changes, pp. 63–67	Fisher exact probability test, pp. 96–104 χ^2 test for two independent samples, pp. 104–111	Cochran Q test, pp. 161–166	χ^2 test for k independent samples, pp. 175–179	Contingency coefficient: C, pp. 196–202
Ordinal	Kolmogorov-Smirnov one-sample test, pp. 47–52 One-sample runs test, pp. 52–58	Sign test, pp. 68–75 Wilcoxon matched-pairs signed-ranks test,† pp. 75–83	Median test, pp. 111–116 Mann-Whitney U test, pp. 116–127 Kolmogorov-Smirnov two-sample test, pp. 127–136 Wald-Wolfowitz runs test, pp. 136–145 Moses test of extreme reactions, pp. 145–152	Friedman two-way analysis of variance, pp. 166–172	Extension of the median test, pp. 179–184 Kruskal-Wallis one-way analysis of variance, pp. 184–193	Spearman rank correlation coefficient: r_S, pp. 202–213 Kendall rank correlation coefficient: τ, pp. 213–223 Kendall partial rank correlation coefficient: $\tau_{xy.z}$, pp. 223–229 Kendall coefficient of concordance: W, pp. 229–238
Interval		Walsh test, pp. 83–87 Randomization test for matched pairs, pp. 88–92	Randomization test for two independent samples, pp. 152–156			

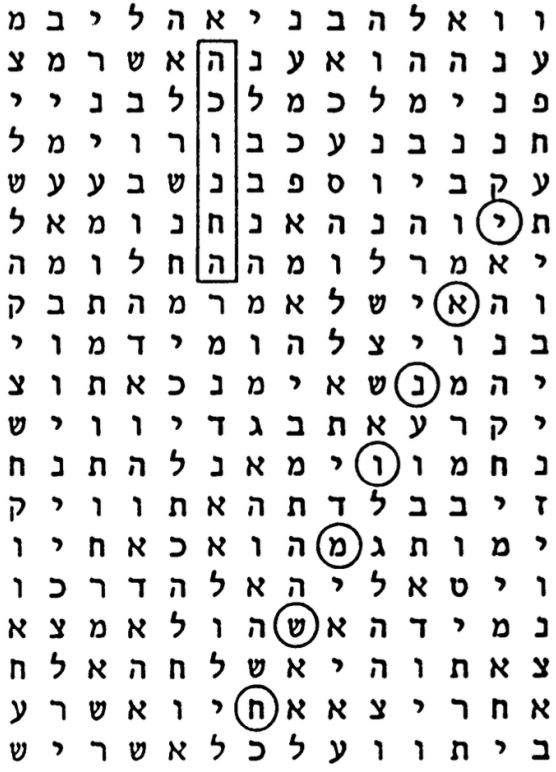
* Each column lists, cumulatively downward, the tests applicable to the given level of measurement. For example, in the case of k related samples, when ordinal measurement has been achieved both the Friedman two-way analysis of variance and the Cochran Q test are applicable.

† The Wilcoxon test requires ordinal measurement not only within pairs, as is required for the sign test, but also of the differences between pairs. See the discussion on pp. 75–76.

Source: Siegel (1988)

Learning Goals

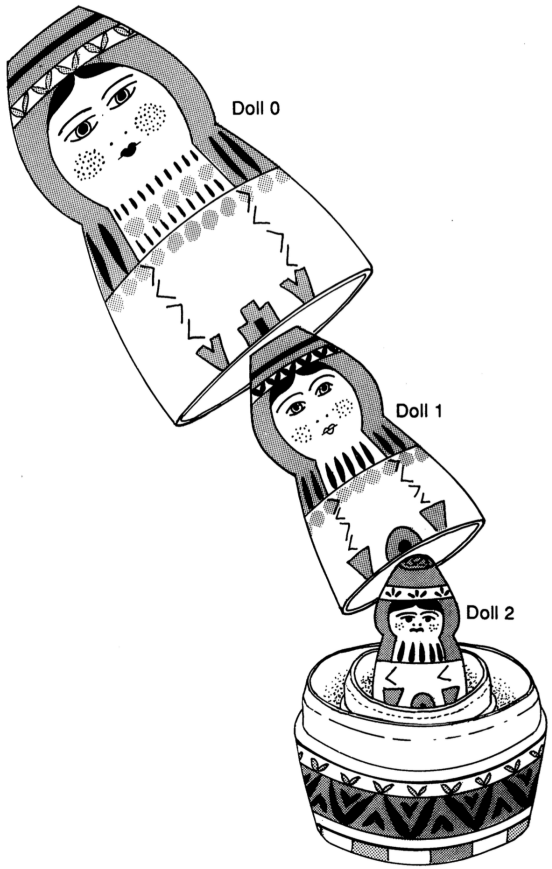
3. The students will learn to apply methods and explain the statistical assumptions of **permutation tests for hypothesis testing**.



Source: Witztum et al. (1994)

Learning Goals

4. The students will learn to apply methods and explain the statistical assumptions of the **bootstrap for confidence intervals**.



Source: Hall (1992)

Learning Goals (Advanced)

5. The students will build-up an **advanced toolbox of methods** that they can use in practical data analysis problems. Tools include:
   - Nonlinear regression,
   - Bayesian nonparametrics,
   - wavelets, and
   - graphons
6. The students will learn to apply various **data visualization tools for data exploration in nonparametric settings**, such as:
   - association plots,
   - mosaic plots,
   - correspondence analysis,
   - median polish, and
   - Tukey additivity plot

Bootstrap

- STATS 208: Introduction to the Bootstrap
  - The bootstrap is a computer-based method for assigning measures of accuracy to statistical estimates. By substituting computation in place of mathematical formulas, it permits the statistical analysis of complicated estimators. Topics: nonparametric assessment of standard errors, biases, and confidence intervals; related resampling methods including the jackknife, cross-validation, and permutation tests. Theory and applications. Prerequisite: course in statistics or probability.

Hypothesis Testing

- STATS 300C: Theory of Statistics
  - The main goal of this course is to expose students to modern ideas in statistical theory. Whereas classical theory is concerned with the behavior of statistical estimates when the number of variables is fixed and the sample size increases, our emphasis is on statistical inference in high-dimensional settings where there may be as many, or more, variables than observations. Our focus is motivated by always newer technologies, which now produce extremely large datasets, often with huge numbers of measurements on each of a comparatively small number of experimental units.

Correspondence Analysis

- STATS 306A: Discrete Data Analysis (Art Owen) but listed as
- STATS 306A: Methods for Applied Statistics
  - Regression modeling extended to categorical data. Logistic regression. Loglinear models. Generalized linear models. Discriminant analysis. Categorical data models from information retrieval and Internet modeling. Prerequisite: 305 or equivalent.

Bayesian Nonparametrics

- STATS 270: Bayesian Statistics I (STATS 370)
  - This is the first of a two course sequence on modern Bayesian statistics. Topics covered include: real world examples of large scale Bayesian analysis; basic tools (models, conjugate priors and their mixtures); Bayesian estimates, tests and credible intervals; foundations (axioms, exchangeability, likelihood principle); Bayesian computations (Gibbs sampler, data augmentation, etc.); prior specification. Prerequisites: statistics and probability at the level of Stats 300A, Stats 305, and Stats 310.

Bayesian Nonparametrics

- STATS 271: Bayesian Statistics II (STATS 371)
  - This is the second of a two course sequence on modern Bayesian statistics. Topics covered include: Asymptotic properties of Bayesian procedures and consistency (Doobs theorem, frequentists consistency, counter examples); connections between Bayesian methods and classical methods (the complete class theorem); generalization of exchangeability; general versions of the Bayes theorem in the undominated case; non parametric Bayesian methods (Dirichelet and Polya tree priors). Throughout general theory will be illustrated with classical examples. Prerequisites: Stats 270/370.

ANOVA

- STATS 203: Introduction to Regression Models and Analysis of Variance
- Modeling and interpretation of observational and experimental data using linear and nonlinear regression methods. Model building and selection methods. Multivariable analysis. Fixed and random effects models. Experimental design. Prerequisites: 200.

Survival Analysis

- STATS 331: Survival Analysis
  - The course introduces basic concepts, theoretical basis and statistical methods associated with survival data. Topics include censoring, Kaplan-Meier estimation, logrank test, proportional hazards regression, accelerated failure time model, multivariate failure time analysis and competing risks. The traditional counting process/martingale methods as well as modern empirical process methods will be covered. Prerequisite: Understanding of basic probability theory and statistical inference methods.

Wavelets

- STATS 322: Function Estimation in White Noise
  - Gaussian white noise model sequence space form. Hyperrectangles, quadratic convexity, and Pinsker's theorem. Minimax estimation on Lp balls and Besov spaces. Role of wavelets and unconditional bases. Linear and threshold estimators. Oracle inequalities. Optimal recovery and universal thresholding. Stein's unbiased risk estimator and threshold choice. Complexity penalized model selection. Connecting fast wavelet algorithms and theory. Beyond orthogonal bases.

Ranked-Set Sampling

- STATS 263: Design of Experiments (STATS 363)
  - Experiments vs observation. Confounding. Randomization. ANOVA. Blocking. Latin squares. Factorials and fractional factorials. Split plot. Response surfaces. Mixture designs. Optimal design. Central composite. Box-Behnken. Taguchi methods. Computer experiments and space filling designs. Prerequisites: probability at STATS 116 level or higher, and at least one course in linear models.

Graphons

- STATS 300: Advanced Topics in Statistics: Stochastic Block Models and Latent Variable Models
  - Main topic: statistical inference of latent variable models (including SBM), using EM-like algorithms. The critical step is the determination of the conditional distribution of the latent variables given the observed data, which is doable for mixture models and hidden Markov models. For more complex models such as the stochastic block model (SBM: popular in sociology, physics, biology, etc.) variational approximations can be used to derive a generalized version of EM algorithm. This approach can be extended to Bayesian inference (variational Bayes EM algorithm). If time permits, change-point detection models will be introduced. Topics will be illustrated with examples from genomics.

Background

- PHIL 166: Probability: Ten Great Ideas About Chance (PHIL 266, STATS 167, STATS 267)
  - Foundational approaches to thinking about chance in matters such as gambling, the law, and everyday affairs. Topics include: chance and decisions; the mathematics of chance; frequencies, symmetry, and chance; Bayes great idea; chance and psychology; misuses of chance; and harnessing chance. Emphasis is on the philosophical underpinnings and problems. Prerequisite: exposure to probability or a first course in statistics at the level of STATS 60 or 116.

Biostatistics Seminar

- STATS 260C: Workshop in Biostatistics (HRP 260C)
- Applications of statistical techniques to current problems in medical science. To receive credit for one or two units, a student must attend every workshop. To receive two units, in addition to attending every workshop, the student is required to write an acceptable one page summary of two of the workshops, with choices made by the student.

References

- Hall (1992). The Bootstrap and Edgeworth Expansion
- Witztum, Rips, and Rosenberg (1994). Equidistant Letter Sequences in the Book of Genesis
- Siegel (1988). Nonparametric Statistics for the Behavioral Sciences