STA732
Statistical Inference

Final Review

Yuansi Chen

Spring 2022

Duke University

https://www2.stat.duke.edu/courses/Spring22/sta732.01/

Exame logistics

Final Exam

- **Time:** April 27, 7pm-10pm?
- **Location:** Old Chem 025
- **Logistics:**
    - Format:
        - Similar to midterm, but with 5 problems
        - Emphasize a bit more on the second half
    - Closed-book exam
    - No electronic devices
    - I will print two help sheets:
        - Formula sheet with common distributions (like the one in first-year exam)
        - Theorem sheet with 10 theorem copied from textbook (see the post on Ed)

Advice for preparation

- Review homework midterm, sample midterm
- Go back to required readings of each lecture
- Can be helpful to do recitation of
  - Main concepts, definitions
  - Main proof techniques
- Try to solve exercises in Keener/ Lehmann and Casella / Lehmann and Romano

Office hours

- Yuansi: Mon/Wed 3:30-4:45pm, 223B
- Youngsoo: to be annouced on Sakai
- Ed Discussion: we will reply in less than 24 hours

Overview of the main topics

Three main types of statistical inference problems

- Point estimation
- Hypothesis testing
- Interval estimation

Types of optimality in point estimation

With model family $P_\theta, \theta \in \Omega$, loss L and risk R, we can consider an optimal estimator in the sense of

- Uniform minimum risk
- Restrict to a smaller class
    - Uniform minimum variance unbiased (UMVU) (Lec 4-6)
    - Minimum risk equivariant (MRE) (Lec 7-8)
- Global approaches
    - Bayes estimator (Lec 9-11)
    - Minimax estimator (Lec 12-13)
- Large sample: asymptotic efficiency of MLE (Lec 14-15)

Types of optimality in hypothesis testing

With model family $P_\theta, \theta \in \Omega$, null + alternative hypotheses, Neyman-Pearson Paradigm, we can consider an optimal test

- Uniformly most powerful (UMP) (Lec 16-19)
- Restrict to a smaller class
  - Uniformly most powerful unbiased (UMPU) (Lec 20-22)
- Large sample: asymptotic of likelihood ratio tests (LRT) (Lec 23)

Basic toy model families

- Exponential families: Bernoulli, Binomial, Poisson, Exponential, Chi-square, Normal, Gamma, beta, Multinomial, etc.
- Linear model or Generalized linear model

Important concepts (not exhaustive)

- Loss and risk
- Sufficiency, completeness
- Bias, Variance, Fisher information
- Equivariance
- Bayes estimator, empirical Bayes and hierarchical Bayes, shrinkage
- Admissibility
- Minimax estimator, least favorable prior
- Convergence in probability, convergence in distribution, MLE
- Level, power, power function in hypothesis testing
- Neyman-Pearson paradigm
- Simple vs composite, one-side vs two-side tests
- Unbiased test, α-similar
- Conditional test, Neyman structure
- Canonical linear model, general linear model

Important technical details (not exhaustive)

What are the typical ways to

- prove sufficiency?
- prove completeness?
- show independence?
- show an estimator is UMVU?
- find MRE?
- show an estimator is Bayes or admissible or minimax?
- manipulate functions of converging random variables? derive asymptotic distribution?
- derive the form a UMP test?
- prove a test is UMP in single-param?
- prove a test is UMPU in single-param? with nuisance param?
- construct optimal tests in general linear model?
- ...

If you have mastered the above materials, congratulations! Your background in theoretical statistics distinguishes you from 99% PhDs in other fields!

Possible extended readings (in theoretical statistical inference)

- Asymptotics:
  - Asymptotic Statistics, van der Vaart 1998
  - Asymptotics in Statistics, Le Cam 2000
- Large sample non-asymptotics:
  - Empirical Processes in M-Estimation, van de Geer 2000
  - High-Dimensional Statistics: A Non-Asymptotic Viewpoint, Wainwright 2017
- Semi-parameteric models:
  - Efficient and Adaptive Estimation for Semiparametric Models, Bickel, Klassen, Ritov, Wellner 1998

Some research ideas

General question: can we derive an optimal estimator (uniform or minimax) if we restrict to estimators that can be computed in polynomial time (in n and d)?

See workshops at Simons Institute

- Computational Complexity of Statistical Inference Boot Camp, 2021
- Rigorous Evidence for Information-Computation Trade-offs, 2021

Exercices

Chapter 8, Ex 26 in Keener:

Let $X_1, \ldots, X_n$ be i.i.d. Poisson with mean λ, and consider estimating

$$g(\lambda) = P_\lambda\left(X_i = 1\right) = \lambda e^{-\lambda}$$

One natural estimator might be the proportion of ones in the sample:

$$\hat{p}_n = \frac{1}{n} \#\left\{i \le n : X_i = 1\right\}.$$

Another choice would be the maximum likelihood estimator, $g\left(\bar{X}_n\right)$, with $\bar{X}_n$ the sample average.

1. Find the asymptotic relative efficiency of $\hat{p}_n$ with respect to $g\left(\bar{X}_n\right)$.
2. Determine the limiting distribution of

$$n\left[g\left(\bar{X}_n\right) - 1/e\right]$$

when $\lambda = 1$. See Page 24 in Lecture 14

Chapter 14, Ex 10 in Keener:

Consider a regression version of the two-sample problem in which

$$Y_i = \begin{cases} \beta_1 + \beta_2 x_i + \epsilon_i, & i = 1, \dots, n_1 \\ \beta_3 + \beta_4 x_i + \epsilon_i, & i = n_1 + 1, \dots, n_1 + n_2 = n, \end{cases}$$

with $\epsilon_1, \dots, \epsilon_n$ i.i.d. from $N\left(0, \sigma^2\right)$. Derive a $1 - \alpha$ confidence interval for $\beta_4 - \beta_2$, the difference between the two regression slopes.

Chapter 9, Ex 7 in Keener

Let $X_1, X_2, \ldots$ be i.i.d. from a uniform distribution on $(0, 1)$ and let T_n maximize

$$\sum_{i=1}^{n} \frac{\log\left(1 + t^2 X_i\right)}{t}$$

over $t > 0$.

1. Show that $T_n \xrightarrow{p} c$ as $n \to \infty$, identifying the constant c.
2. Find the limiting distribution for $\sqrt{n}\left(T_n - c\right)$ as $n \to \infty$.

Thank you for attending

See you on April 27 in Old Chem 025