STATS 200: Introduction to Statistical Inference

Lecture 29: Course review

Course review

We started in Lecture 1 with a fundamental assumption:

Data is a realization of a random process.

The goal throughout this course has been to use the observed data to draw inferences about the underlying process or probability distribution.

Course review

We started in Lecture 1 with a fundamental assumption:

Data is a realization of a random process.

The goal throughout this course has been to use the observed data to draw inferences about the underlying process or probability distribution.

We discussed:

- Hypothesis testing—deciding whether a particular "null hypothesis" about the underlying distribution is true or false
- Estimation—fitting a parametric model to this distribution and/or estimating a quantity related to the parameters of this model
- Standard errors and confidence intervals—quantifying the uncertainty of these estimates

Hypothesis testing

Goal: Accept or reject a null hypothesis H_0 based on the value of an observed test statistic T.

Question #1: How to choose a test statistic T?

Question #2: How to decide whether H_0 is true/false based on T?

Hypothesis testing

Goal: Accept or reject a null hypothesis H_0 based on the value of an observed test statistic T.

Question #1: How to choose a test statistic T?

Question #2: How to decide whether H_0 is true/false based on T?

In a simple-vs-simple testing problem, there is a "best" answer to Question #1, which is the likelihood ratio statistic

$$L(X_1, \ldots, X_n) = \frac{f_0(X_1, \ldots, X_n)}{f_1(X_1, \ldots, X_n)}.$$

We can equivalently use any monotonic 1-to-1 transformation of this statistic, which is simpler to understand in many examples (e.g. the total count for Bernoulli coin flips).

Hypothesis testing

In problems with composite alternatives, there is oftentimes not a single most powerful test against all of the alternative distributions. We instead discussed popular choices of T for several examples:

Hypothesis testing

In problems with composite alternatives, there is oftentimes not a single most powerful test against all of the alternative distributions. We instead discussed popular choices of T for several examples:

- Testing goodness of fit: Histogram methods (Pearson's chi-squared statistic), QQ-plot methods (one-sample Kolmogorov-Smirnov statistic)

Hypothesis testing

In problems with composite alternatives, there is oftentimes not a single most powerful test against all of the alternative distributions. We instead discussed popular choices of T for several examples:

- Testing goodness of fit: Histogram methods (Pearson's chi-squared statistic), QQ-plot methods (one-sample Kolmogorov-Smirnov statistic)
- Testing if a distribution is centered at 0: One-sample t-statistic, signed-rank statistic, sign statistic

Hypothesis testing

In problems with composite alternatives, there is oftentimes not a single most powerful test against all of the alternative distributions. We instead discussed popular choices of T for several examples:

- Testing goodness of fit: Histogram methods (Pearson's chi-squared statistic), QQ-plot methods (one-sample Kolmogorov-Smirnov statistic)
- Testing if a distribution is centered at 0: One-sample t-statistic, signed-rank statistic, sign statistic
- Testing if two samples have the same distribution or mean: Two-sample t-statistic, rank-sum statistic

Hypothesis testing

In problems with composite alternatives, there is oftentimes not a single most powerful test against all of the alternative distributions. We instead discussed popular choices of T for several examples:

- Testing goodness of fit: Histogram methods (Pearson's chi-squared statistic), QQ-plot methods (one-sample Kolmogorov-Smirnov statistic)
- Testing if a distribution is centered at 0: One-sample t-statistic, signed-rank statistic, sign statistic
- Testing if two samples have the same distribution or mean: Two-sample t-statistic, rank-sum statistic
- Testing if the parameters of a model satisfy additional constraints (i.e. belong to a sub-model): Generalized likelihood ratio statistic

Hypothesis testing

Question #2: How to decide whether H_0 is true/false based on T?

We adopted the frequentist significance testing framework: Control the probability of type I error (falsely rejecting H_0) at a target level α, by considering the distribution of T if H_0 were true.

Hypothesis testing

Question #2: How to decide whether H_0 is true/false based on T?

We adopted the frequentist significance testing framework: Control the probability of type I error (falsely rejecting H_0) at a target level α, by considering the distribution of T if H_0 were true.

To do this for each test, we either derived the null distribution of T exactly (e.g. t-test), appealed to asymptotic approximations (e.g. the χ^2 distribution for the GLRT), or used computer simulation (e.g. permutation two-sample tests).

Estimation

Goal: Fit a parametric probability model to the observed data. For IID data $X_1, \ldots, X_n \sim f(x|\theta)$, we discussed three methods:

- Method of moments: Equate the sample mean of X_i, X_i^2, etc. to the theoretical mean of X, X^2, etc., and solve for θ

Estimation

Goal: Fit a parametric probability model to the observed data. For IID data $X_1, \ldots, X_n \sim f(x|\theta)$, we discussed three methods:

- Method of moments: Equate the sample mean of X_i, X_i^2, etc. to the theoretical mean of X, X^2, etc., and solve for θ
- Maximum likelihood: Pick θ to maximize $\prod_{i=1}^n f(X_i|\theta)$, or equivalently $\sum_{i=1}^n \log f(X_i|\theta)$

Estimation

Goal: Fit a parametric probability model to the observed data. For IID data $X_1, \ldots, X_n \sim f(x|\theta)$, we discussed three methods:

- Method of moments: Equate the sample mean of X_i, X_i^2, etc. to the theoretical mean of X, X^2, etc., and solve for θ
- Maximum likelihood: Pick θ to maximize $\prod_{i=1}^n f(X_i|\theta)$, or equivalently $\sum_{i=1}^n \log f(X_i|\theta)$
- Bayesian inference: Postulate a prior distribution $f_\Theta(\theta)$ for θ, compute the posterior

$$f_{\Theta|X}(\theta|X_1, \ldots, X_n) \propto f_\Theta(\theta) \prod_{i=1}^n f(X_i|\theta),$$

and estimate θ by e.g. the posterior mean

Estimation

Goal: Fit a parametric probability model to the observed data. For IID data $X_1, \ldots, X_n \sim f(x|\theta)$, we discussed three methods:

- Method of moments: Equate the sample mean of X_i, X_i^2, etc. to the theoretical mean of X, X^2, etc., and solve for θ
- Maximum likelihood: Pick θ to maximize $\prod_{i=1}^n f(X_i|\theta)$, or equivalently $\sum_{i=1}^n \log f(X_i|\theta)$
- Bayesian inference: Postulate a prior distribution $f_\Theta(\theta)$ for θ, compute the posterior

$$f_{\Theta|X}(\theta|X_1, \ldots, X_n) \propto f_\Theta(\theta) \prod_{i=1}^n f(X_i|\theta),$$

  and estimate θ by e.g. the posterior mean

We illustrated how the method of maximum likelihood generalizes to regression models with covariates, where the data $Y_1, \ldots, Y_n$ are independent but not identically distributed.

Estimation

We discussed the accuracy of estimators in terms of several finite-n properties:

- The bias $\mathbb{E}_\theta[\hat{\theta}] - \theta$
- The variance $\mathrm{Var}_\theta[\hat{\theta}]$
- The MSE $\mathbb{E}_\theta[(\hat{\theta} - \theta)^2] = \text{bias}^2 + \text{variance}$

Estimation

We discussed the accuracy of estimators in terms of several finite-n properties:

- The bias $\mathbb{E}_\theta[\hat{\theta}] - \theta$
- The variance $\mathrm{Var}_\theta[\hat{\theta}]$
- The MSE $\mathbb{E}_\theta[(\hat{\theta} - \theta)^2] = \text{bias}^2 + \text{variance}$

We also discussed asymptotic properties, as $n \to \infty$ and when the true parameter is θ:

- Consistency: $\hat{\theta} \to \theta$ in probability
- Asymptotic normality: $\sqrt{n}(\hat{\theta} - \theta) \to \mathcal{N}(0, v(\theta))$ in distribution for some variance $v(\theta)$
- Asymptotic efficiency: $\hat{\theta}$ is asymptotically normal, and the limiting variance is $v(\theta) = I(\theta)^{-1}$, where $I(\theta)$ is the Fisher information (of a single observation)

Estimation

The definition of asymptotic efficiency was motivated by the Cramer-Rao lower bound: Under mild smoothness conditions, for any unbiased estimator $\hat{\theta}$ of θ, its variance is at least $\frac{1}{n} I(\theta)^{-1}$.

Estimation

The definition of asymptotic efficiency was motivated by the Cramer-Rao lower bound: Under mild smoothness conditions, for any unbiased estimator $\hat{\theta}$ of θ, its variance is at least $\frac{1}{n} I(\theta)^{-1}$.

A major result was that the MLE is asymptotically efficient:

$$\sqrt{n}(\hat{\theta} - \theta) \to \mathcal{N}(0, I(\theta)^{-1})$$

We showed informally that Bayes estimators, asymptotically for large n, approach the MLE—an implication is that Bayes estimators are usually also asymptotically efficient.

On the other hand, method-of-moments estimators are oftentimes not asymptotically efficient and have a larger mean-squared error than the other two procedures for large n.

Standard errors and confidence intervals

We discussed standard error estimates and confidence intervals both when the model is correctly specified and when it is not.

In a correctly specified model, we can derive the variance $v(\theta)$ of the estimate $\hat{\theta}$ in terms of the true parameter θ, and estimate the standard error by the plugin estimate $\sqrt{v(\hat{\theta})}$.

Standard errors and confidence intervals

We discussed standard error estimates and confidence intervals both when the model is correctly specified and when it is not.

In a correctly specified model, we can derive the variance $v(\theta)$ of the estimate $\hat{\theta}$ in terms of the true parameter θ, and estimate the standard error by the plugin estimate $\sqrt{v(\hat{\theta})}$.

For the MLE, asymptotic efficiency implies that $v(\theta) \approx \frac{1}{n} I(\theta)^{-1}$ for large n. In the setting of non-IID data $Y_1, \ldots, Y_n$ in regression models, we used the Fisher information of all n samples, $I_{\mathbf{Y}}(\theta)^{-1}$, in place of $\frac{1}{n} I(\theta)^{-1}$.

For other estimators, we can sometimes derive $v(\theta)$ directly from the form of $\hat{\theta}$, perhaps using asymptotic approximations like the CLT and delta method.

Standard errors and confidence intervals

In an incorrectly specified model, we studied the behavior of the MLE $\hat{\theta}$ and interpreted the parameter θ that it tries to estimate as the probability distribution in the model "closest in KL-divergence" to the true distribution of data.

We derived a more general formula for the variance of $\hat{\theta}$, and showed how this variance may be estimated by a "sandwich" estimator.

Alternatively, we discussed the nonparametric bootstrap as a simulation-based approach for estimating the standard error that is also robust to model misspecification.

Standard errors and confidence intervals

To estimate a function $g(\theta)$, we may use the plugin estimate $g(\hat{\theta})$.

- If $\hat{\theta}$ is asymptotically normal, then $g(\hat{\theta})$ is usually also asymptotically normal, and its asymptotic variance may be derived using the delta method.
- If $\hat{\theta}$ is asymptotically efficient (e.g. the MLE), then $g(\hat{\theta})$ is also asymptotically efficient for $g(\theta)$.

Standard errors and confidence intervals

To estimate a function $g(\theta)$, we may use the plugin estimate $g(\hat{\theta})$.

- If $\hat{\theta}$ is asymptotically normal, then $g(\hat{\theta})$ is usually also asymptotically normal, and its asymptotic variance may be derived using the delta method.
- If $\hat{\theta}$ is asymptotically efficient (e.g. the MLE), then $g(\hat{\theta})$ is also asymptotically efficient for $g(\theta)$.

For any quantity θ, an approximate level $100(1-\alpha)\%$ confidence interval for θ may be obtained from any asymptotically normal estimator $\hat{\theta}$ and an estimate $\hat{\mathrm{se}}$ of its standard error:

$$\hat{\theta} \pm z(\alpha/2)\hat{\mathrm{se}}$$

Most confidence intervals that we constructed in this class were of this form. (The accuracy of such intervals should be checked by simulation if n is small.)

Final exam

- You are allowed one 8.5×11 inch "cheat sheet", front and back. No other outside material is permitted.
- The exam will provide relevant formulas (like the PDFs/PMFs of common distributions), as was done on the midterm.
- Some material is from before the midterm, but the focus is on Units 2 and 3.
- The exam will test conceptual understanding and problem solving. You won't be asked to reproduce detailed mathematical proofs.
- We were informal in our discussion of regularity conditions required for asymptotic efficiency of the MLE, asymptotic χ^2 distribution of the GLRT statistic $-2 \log \Lambda$, etc. For parametric models having differentiable likelihood function and common support, you don't need to check regularity conditions when applying these results.

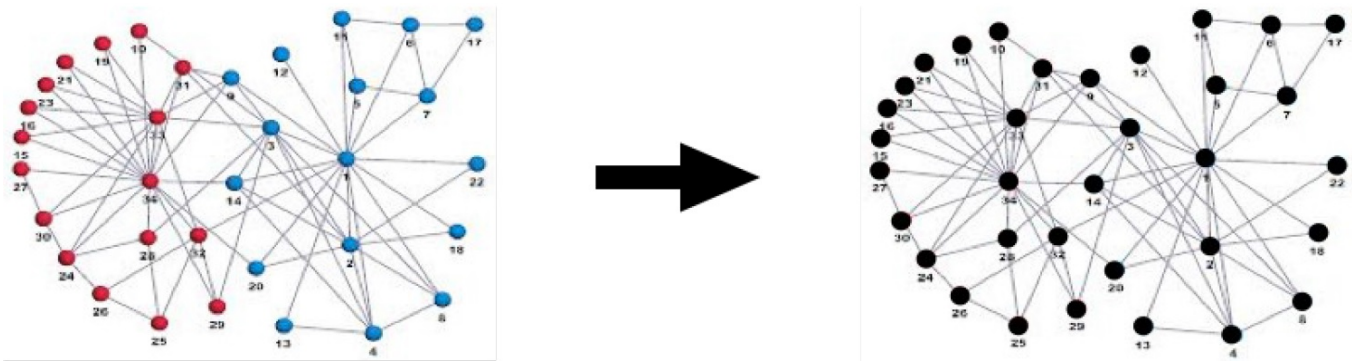
One last example: Beyond the MLE

We live in a time when there is a convergence of ideas and interchange of tools across quantitative disciplines. There is a rich interplay between statistical inference and optimization, algorithms, machine learning, and information theory.

Here is one last example which illustrates how the idea of the MLE, in a seemingly simple problem, connects to interesting questions in a variety of other fields of study.

One last example: Beyond the MLE

- n people, belonging to two distinct communities of equal size $n/2$, are connected in a social network.
- Every pair of people is connected (independently) with probability p if they are in the same community, and with probability q if they are in different communities, where $q < p$.
- We can see the network of connections, but we cannot see which person belongs to which community.



Question: Suppose (for simplicity) that we know the values p and q. How can we infer who belongs to which community?

One last example: Beyond the MLE

Let $S \subset \{1, \ldots, n\}$ denote community 1, and S^c denote community 2.

One last example: Beyond the MLE

Let $S \subset \{1, \ldots, n\}$ denote community 1, and S^c denote community 2.

Let Same be the set of pairs $\{i, j\}$ such that $i, j \in S$ or $i, j \in S^c$, and let $\mathrm{Different}$ be the set of pairs $\{i, j\}$ such that one of i, j belongs to S and the other to S^c.

One last example: Beyond the MLE

Let $S \subset \{1, \ldots, n\}$ denote community 1, and S^c denote community 2.

Let Same be the set of pairs $\{i, j\}$ such that $i, j \in S$ or $i, j \in S^c$, and let $\mathrm{Different}$ be the set of pairs $\{i, j\}$ such that one of i, j belongs to S and the other to S^c.

Our observed data are the connections in this network:

$$A_{ij} = \begin{cases} 1 \text{ if } \{i, j\} \text{ are connected in the network} \\ 0 \text{ otherwise} \end{cases}$$

Under our model, A_{ij} are independent Bernoulli random variables with $A_{ij} \sim \text{Bernoulli}(p)$ if $\{i, j\} \in \mathrm{Same}$ and $A_{ij} \sim \text{Bernoulli}(q)$ if $\{i, j\} \in \mathrm{Different}$.

One last example: Beyond the MLE

The likelihood function is

$$\begin{aligned}\text{lik}(S) &= \prod_{\{i,j\}\in\text{Same}} p^{A_{ij}}(1-p)^{1-A_{ij}} \prod_{\{i,j\}\in\text{Different}} q^{A_{ij}}(1-q)^{1-A_{ij}} \\ &= \prod_{\{i,j\}\in\text{Same}} (1-p)\left(\frac{p}{1-p}\right)^{A_{ij}} \prod_{\{i,j\}\in\text{Different}} (1-q)\left(\frac{q}{1-q}\right)^{A_{ij}}\end{aligned}$$

One last example: Beyond the MLE

The likelihood function is

$$\begin{aligned}\text{lik}(S) &= \prod_{\{i,j\}\in\text{Same}} p^{A_{ij}}(1-p)^{1-A_{ij}} \prod_{\{i,j\}\in\text{Different}} q^{A_{ij}}(1-q)^{1-A_{ij}} \\ &= \prod_{\{i,j\}\in\text{Same}} (1-p)\left(\frac{p}{1-p}\right)^{A_{ij}} \prod_{\{i,j\}\in\text{Different}} (1-q)\left(\frac{q}{1-q}\right)^{A_{ij}}\end{aligned}$$

Each observed connection (where $A_{ij} = 1$) contributes a factor of $p/(1-p)$ to the likelihood if $\{i,j\} \in \text{Same}$, and a factor of $q/(1-q)$ if $\{i,j\} \in \text{Different}$.

Since $p > q$, the MLE $\hat{S}$ is given by the partition of the people into two communities that minimizes the number of observed connections between communities (or, equivalently, maximizes the number of connections within communities).

One last example: Beyond the MLE

More formally, the MLE $\hat{S}$ solves the optimization problem

$$\begin{aligned} &\text{minimize} \sum_{i \in S} \sum_{j \in S^c} A_{ij} \\ &\text{subject to } |S| = |S^c| = n/2 \end{aligned}$$

This is a well-known problem in computer science, called the **minimum graph bisection** problem.

One last example: Beyond the MLE

More formally, the MLE $\hat{S}$ solves the optimization problem

$$\begin{aligned} &\text{minimize} \sum_{i \in S} \sum_{j \in S^c} A_{ij} \\ &\text{subject to } |S| = |S^c| = n/2 \end{aligned}$$

This is a well-known problem in computer science, called the **minimum graph bisection** problem.

Unfortunately, this problem is known to be NP-complete—it is widely believed that no computationally efficient computer algorithm can compute this MLE (for all possible realizations of the network).

One last example: Beyond the MLE

This leads to a number of interesting questions:

- Can we approximately solve this optimization problem, and prove that our answer is not too far off?
- Are there other algorithms (not directly based on this optimization) that can yield a good estimate of S?
- What is a lower bound for the error (expected fraction of people assigned to the incorrect community) achievable by any estimator?
- How robust are our algorithms to the modeling assumptions, and how well do they generalize to settings with more than two communities?

These questions have attracted the attention of people working in statistics, mathematics, computer science, statistical physics, and optimization, and they remain an active area of research today.