Mathematics/Statistics Bootcamp

Part IV: Basics of Statistical Inference

Fan Bu[1] Kyle Burris[1]

[1]Department of Statistical Science
Duke University

Graduate Orientation, August 2018

Overview

Limiting Theorems

Data Reduction

Point Estimation

Hypothesis Testing

Interval Estimation

Introduction to Bayesian Analysis

Limiting Theorems

The Law of Large Numbers (LLN)

Suppose $\{X_1, X_2, \dots\}$ is a sequence of independently and identically distributed (i.i.d.) random variables with $E[X_i] = \mu$. Let $\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$ be the sample average. Then:

- The **Weak Law**: $\bar{X}_n \xrightarrow{p} \mu$ when $n \to \infty$, that is, for any $\epsilon > 0$,

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| > \epsilon) = 0.$$

- The **Strong Law**: $\bar{X}_n \xrightarrow{a.s.} \mu$ when $n \to \infty$, that is,

$$P\left(\lim_{n\to\infty} \bar{X}_n = \mu\right) = 1.$$

The Central Limit Theorem (CLT)

Suppose $\{X_1, X_2, \ldots\}$ is a sequence of independently and identically distributed (i.i.d.) random variables with $E[X_i] = \mu$ and $Var[X_i] = \sigma^2 < \infty$. Let $\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$ be the sample average, then as $n \to \infty$, the random variable $\sqrt{n}(\bar{X}_n - \mu)$ converges in distribution to $N(0, \sigma^2)$:

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2).$$

Mini-exercises

1. Rewrite the CLT in terms of the sample sum, $S_n = \sum_{i=1}^{n} X_i$.

2. Let $\{X_1, X_2, \ldots, X_n\}$ be a sequence of n independent results from tossing the same fair coin where $X_i = 1$ when the head faces up and $X_i = 0$ otherwise. Let $\bar{X}_n = \frac{\sum_{i=1}^{n} X_i}{n}$. If $n = 100$, estimate the value of

$$P(0.4 < \bar{X}_n < 0.6).$$

Data Reduction

Sufficiency

- **Definition**: A statistic $T(X)$ is a **sufficient statistic for** θ if the conditional distribution of the sample X given the value of $T(X)$ does not depend on θ.
- **Sufficiency Principle**: If $T(X)$ is a sufficient statistic for θ, then any inference about θ should depend on the sample X only through the value $T(X)$. That is, if x and y are two sample points such that $T(X) = T(Y)$, then the inference about θ should be the same whether $X = x$ or $Y = y$ is observed.
- **Factorization Theorem**: Let $f(x|\theta)$ denote the pdf or pmf of a sample X. A statistic $T(X)$ is a sufficient statistic for θ if and only if there exist functions $g(t|\theta)$ and $h(x)$ such that, for all sample points x and all parameter points θ,

$$f(x|\theta) = g(T(x)|\theta)h(x).$$

Sufficiency: An Exercise

Let $X_1, X_2, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$ where σ^2 is known. Show that the sample mean, $T(X) = \bar{X} = (X_1 + X_2 + \cdots + X_n)/n$, is a sufficient statistic for μ.

Note: the joint pdf of the sample **X** is

$$
\begin{aligned}
f(\mathbf{x}|\mu) &= \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp(-(x_i - \mu)^2/(2\sigma^2)) \\
&= (2\pi\sigma^2)^{-1/2} \exp(-\sum_{i=1}^{n} (x_i - \mu)^2/(2\sigma^2)).
\end{aligned}
$$

Sufficient Statistics for the Exponential Family

Let $X_1, X_2, \ldots, X_n$ be i.i.d observations from an exponential family with pdf or pmf of the form

$$f(x|\theta) = h(x)c(\theta)\exp\left(\sum_{j=1}^{k} w(\theta_j)t_j(x)\right),$$

where $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$. Then the statistic

$$T(X) = \left(\sum_{i=1}^{n} t_1(X_i), \sum_{i=1}^{n} t_2(X_i), \ldots, \sum_{i=1}^{n} t_k(X_i)\right)$$

is a sufficient statistic.

Sufficiency: Exercises

1. Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d observations from a Poisson distribution with parameter λ. Find a sufficient statistic for λ.

2. Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d observations from a Gamma distribution with parameters $\theta = (\alpha, \beta)$ (use the shape-rate parametrization). Find a sufficient statistic for θ.

Point Estimation

Point Estimation

- A **point estimator** is any function of the sample.
- **Estimator** vs. **Estimate**: The former is a function, while the latter is the realized value of the function (a number) that is obtained when a sample is actually taken.

Maximum Likelihood Estimators

- If $X_1, \ldots, X_n$ are an i.i.d. sample from a population with pdf or pmf $f(\mathbf{x}|\theta_1, \ldots, \theta_k)$, the **likelihood function** is

$$L(\theta|\mathbf{x}) = L(\theta_1, \ldots, \theta_k | x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta_1, \ldots, \theta_k).$$

- For each sample point $\mathbf{x}$, let $\hat{\theta}(\mathbf{x})$ be a parameter value at which $L(\theta|\mathbf{x})$ attains its maximum as a function of θ, with $\mathbf{x}$ held fixed. A **maximum likelihood estimator (MLE)** of the parameter θ based on a sample $\mathbf{X}$ is $\hat{\theta}(\mathbf{X})$.
- If the likelihood function is differentiable (in θ_i), **possible candidates** for the MLE are the values of $(\theta_1, \ldots, \theta_k)$ that satisfy

$$\frac{\partial}{\partial \theta_i} L(\theta|\mathbf{x}) = 0, \quad i = 1, \ldots, k.$$

MLE: Normal Example

Let $X_1, \ldots, X_n$ be i.i.d. $N(\theta, 1)$, and let $L(\theta|\mathbf{x})$ denote the likelihood function. Since maximizing

$$L(\theta|\mathbf{x}) = \frac{1}{(2\pi)^{n/2}} e^{(-1/2)\sum_{i=1}^{n}(x_i-\theta)^2},$$

is equivalent to maximizing $\ln L(\theta|\mathbf{x})$, which reduces to maxizing

$$h(\theta) = (-1/2)\sum_{i=1}^{n}(x_i - \theta)^2,$$

a quadratic function of θ.

Since $\hat{\theta} = \bar{x} = (\sum_{i=1}^{n} x_i)/n$ is the global maxima of $h(\theta)$, it is also the global maxima of $L(\theta|\mathbf{x})$. Therefore $\hat{\theta}$ is the MLE.

MLE: Exercises

1. Let $X_1, \ldots, X_n$ be i.i.d. samples from the uniform distribution $U(0, \theta)$, $\theta > 0$. Find the MLE of θ.

2. Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli(p). Find the MLE of p.

The Invariance Property of MLEs

Theorem

If $\hat{\theta}$ is the MLE of θ, then for any function $\tau(\theta)$ of θ, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.

A mini-exercise: following ex.2 from above, what is the MLE of $\sqrt{p(1-p)}$?

Mean Squared Error (MSE) and Bias

- The **mean squared error (MSE)** of an estimator W of a parameter θ is defined by $E_\theta(W-\theta)^2$.
- The **bias** of a point estimator W of a parameter θ is $\text{Bias}_\theta W = E_\theta W - \theta$, and an estimator is called **unbiased** if $E_\theta W = \theta$ for all θ.
- Relationship between MSE and bias:

$$E_\theta(W-\theta)^2 = \text{Var}_\theta W + (\text{Bias}_\theta W)^2.$$

- If W is an unbiased estimator of θ,

$$E_\theta(W-\theta)^2 = \text{Var}_\theta W.$$

MSE and Bias: An Exercise

Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$, $\bar{X} = (\sum_{i=1}^n X_i)/n$ be the sample mean, and $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$ be the sample variance. Verify that $\bar{X}$ and S^2 are unbiased estimators for μ and σ^2, respectively, and compute their MSEs.
(Note: if $Y \sim \chi_k^2$, then $\mathrm{Var}(Y) = 2k$.)

If we adopt the MLE estimator $\hat{\sigma}^2$ for σ^2 instead, where $\hat{\sigma}^2 = \sum_{i=1}^n (X_i - \bar{X})^2/n$. What is the MSE of $\hat{\sigma}^2$?

Review Exercises: Morning Session

1. Let $X_1, X_2, \ldots, X_n$ be i.i.d. Poisson(λ). Let $Y_n = (\sum_{i=1}^n X_i - \lambda)/\sqrt{n}$. When n is large, Y_n is approximately a normal variable with mean μ and variance σ^2. What are μ and σ^2 ?

2. Let $X_1, X_2, \ldots, X_n$ be i.i.d. Exp(β) with pdf $f(x) = \beta e^{-\beta x}(x > 0)$. Show that $\bar{X} = \sum_{i=1}^n X_i/n$ is a sufficient statistic for β, and find the MLE of β.

3. Let X be a sample from $N(0, \sigma^2)$. Find an unbiased estimator of σ^2.

Hypothesis Testing

Intro to Hypothesis Testing

Video tutorial, by *mathtutordvd*

Hypothesis Testing: Likelihood Ratio Tests

The **likelihood ratio test statistic** for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$ is

$$\lambda(\mathbf{x}) = \frac{\sup_{\Theta_0} L(\theta|\mathbf{x})}{\sup_{\Theta} L(\theta|\mathbf{x})}.$$

A **likelihood ratio test (LRT)** is any test that has a rejection region of the form $\{\mathbf{x} : \lambda(\mathbf{x}) \leq c\}$, where c is a number satisfying $0 \leq c \leq 1$.

Suppose $\hat{\theta}$ is an MLE of θ (obtained by the unrestricted maximization of $L(\theta|\mathbf{x})$), and $\hat{\theta}_0$ is the MLE of θ assuming the parameter space is Θ_0 ((obtained by maximizing $L(\theta|\mathbf{x})$) on Θ_0). Then the LRT statistic is

$$\lambda(\mathbf{x}) = \frac{L(\hat{\theta}_0|\mathbf{x})}{L(\hat{\theta}|\mathbf{x})}.$$

LRT and Sufficiency

Theorem

If $T(\mathbf{X})$ *is a sufficient statistic for* θ *and* $\lambda^*(t)$ *and* $\lambda(\mathbf{x})$ *are the LRT statistics based on* T *and* $\mathbf{X}$*, respectively, then* $\lambda^*(T(\mathbf{x})) = \lambda(\mathbf{x})$ *for every* $\mathbf{x}$ *in the sample space.*

An Exercise: Normal LRT

Let $X_1, \ldots, N_n$ be i.i.d. $N(\theta, 1)$. Consider the test $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. Find the LRT statistic and derive the form of the rejection region.

Test Errors and Power Function

- Type I Error and Type II Error:

		Decision	
		Accept H_0	Reject H_0
Truth	H_0	Correct decision	Type I Error
	H_1	Type II Error	Correct decision

- Suppose R denotes the rejection region for a test, then the probability of a Type I Error is $P_\theta(\mathbf{X} \in R | H_0)$, and the probability of a Type II Error is $P_\theta(\mathbf{X} \in R^c | H_1) = 1 - P_\theta(\mathbf{X} \in R | H_1)$.
- The **power function** of a hypothesis test with rejection region R is the function of θ defined by $\beta(\theta) = P_\theta(\mathbf{X} \in R)$.
- For $0 \leq \alpha \leq 1$, a test with power function $\beta(\theta)$ is a **level** α **test** if $\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha$.

An Exercise: Binomial

Let $X \sim \text{Binomial}(5, \theta)$. Consider the test $H_0 : \theta \leq 1/2$ versus $H_1 : \theta > 1/2$. If we adopt the test that rejects H_0 only if $X = 5$ is observed. What is the power function of this test? How small is the probability of Type I Error? For what values of θ is the probability of Type II Error less than $\frac{1}{2}$?

p-values

Definition 1:
A **p-value** $p(\mathbf{X})$ is a test statistic satisfying $0 \leq p(\mathbf{x}) \leq 1$ for every sample point $\mathbf{x}$. A p-value is **valid** if, for every $\theta \in \Theta_0$ and every $0 \leq \alpha \leq 1$,

$$P_\theta(p(\mathbf{X}) \leq \alpha) \leq \alpha.$$

- If we observe $\mathbf{X} = \mathbf{x}$, then for any $\alpha \geq p(\mathbf{x})$, a level α test rejects H_0;
- p-value is essentially a summary statistic of the data. Small values of $p(\mathbf{X})$ give evidence that H_1 is true.

p-values (Cont'd)

Definition 2:
Let $W(\mathbf{X})$ be a test statistic such that large values of W give evidence that H_1 is true. For each sample point $\mathbf{x}$, define

$$p(\mathbf{x}) = \sup_{\theta \in \Theta_0} P_\theta(W(\mathbf{X}) \geq W(\mathbf{x})).$$

Then $p(\mathbf{X})$ is a valid p-value.

- "p-value": the probability of obtaining a sample "more extreme" than the ones observed in the data, assuming H_0 is true.

p-values: An Exercise

A neurologist is testing the effect of a drug on response time by injecting 100 rats with a unit dose of the drug, subjecting each to neurological stimulus, and recording its response time. The neurologist knows that the response time for a rat not injected with the drug follows a normal distribution with a mean response time of 1.2 seconds. The mean of the 100 injected rats' response times is 1.05 seconds with a sample standard deviation of 0.5 seconds.

Do you suggest that the neurologist conclude that the drug has an effect on response time?

Solution to the Exercise

Suppose the mean response time for rats injected with the drug is μ, then we want to test

$$H_0 : \mu = 1.2s \text{ (the drug has no effect)}$$

against

$$H_1 : \mu \neq 1.2s \text{ (the drug has effect)} .$$

Construct the test statistic (here $\bar{X}$ is the sample mean, and S is the sample standard deviation)

$$Z = \frac{\bar{X} - 1.2}{S/\sqrt{100}}.$$

$Z \sim t_{99}$, which is approximately $N(0,1)$. Plug in the observed data, $\bar{x} = 1.05, s = 0.5$, and $z = -3$, so the p-value is approximately $P(|W| \geq |z|) = P(|W| \geq 3) \approx 0.003$ (let $W \sim N(0,1)$).

Interval Estimation

Interval Estimation

- An **interval estimate** of a parameter θ is any pair of functions, $L(x_1, \ldots, x_n)$ and $U(x_1, \ldots, x_n)$, of a sample that satisfy $L(\mathbf{x}) \leq U(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$. The inference $L(\mathbf{x}) \leq \theta \leq U(\mathbf{x})$ is made once $\mathbf{X} = \mathbf{x}$ is observed. The **random interval** $[L(\mathbf{X}), U(\mathbf{X})]$ is called an **interval estimator**.
- The **coverage probability** of an interval estimator $[L(\mathbf{X}), U(\mathbf{X})]$ of a parameter θ is the probability that the random interval $[L(\mathbf{X}), U(\mathbf{X})]$ covers the true parameter, θ. It is denoted by $P(\theta \in [L(\mathbf{X}), U(\mathbf{X})])$, or $P(\theta \in [L(\mathbf{X}), U(\mathbf{X})]|\theta)$.
- The **confidence coefficient** of an interval estimator $[L(\mathbf{X}), U(\mathbf{X})]$ of a parameter θ is the infimum of the coverage probabilities for all values of θ, $\inf_\theta P(\theta \in [L(\mathbf{X}), U(\mathbf{X})])$.

Interval Estimation: Key Points

1. The **interval** is the random quantity, not the parameter;
2. **"Confidence intervals/sets"**: interval estimators with a measure of confidence (a confidence coefficient); eg. a confidence interval/set with confidence coefficient equal to C, is called a "C confidence interval/set".
3. The **coverage probability** is a function of θ, whose true value is unknown, so we can only guarantee the infimum of the coverage probability, the confidence coefficient.

A mini-exercise

Suppose that X is a random sample from a distribution with parameter θ, and $[L(X), U(X)]$ is a 95% confidence interval of θ. If we observe $X = x$, which of the following statements is correct?

A The probability that $\theta \in [L(x), U(x)]$ is 0.95;

B The probability that $\theta \in [L(x), U(x)]$ is either 1 or 0.

Example: Normal Confidence Interval

If $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$ with σ^2 known, then $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ is a standard normal variable ($Z \sim N(0,1)$). Then a confidence interval of μ can be

$$\{\mu : \bar{x} - a\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + a\frac{\sigma}{\sqrt{n}}\},$$

where a is a constant.
If σ^2 is unknown, then $T_{n-1} = (\bar{X} - \mu)/(S/\sqrt{n}) \sim t_{n-1}$ which is independent of μ. Thus, for any given $\alpha \in (0,1)$, a $1 - \alpha$ confidence interval of μ is given by

$$\{\mu : \bar{x} - t_{n-1,(1-\alpha/2)}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{n-1,(1-\alpha/2)}\frac{s}{\sqrt{n}}\},$$

where $t_{df,p}$ is the $p \times 100\%$th quantile of a student-t distribution with df degrees of freedom.

Introduction to Bayesian Analysis

Video Tutorial

Introduction to Bayesian Data Analysis, by *asmusab*

Exercise:
Python: `https://goo.gl//ceShN5`
R: `https://goo.gl//cxfnYK`

Bayesian Analysis: the Basics

Two quantities of interest:

1. $y \in \mathcal{Y}$: the data ($\mathcal{Y}$: the sample space), a subset of members of the population of interest;
2. $\theta \in \Theta$: the parameter (Θ: the parameter space), expressing the population characteristics.

Three distributions:

1. For each numerical value $\theta \in \Theta$, the **prior distribution** $p(\theta)$ describes our belief that θ represents the true population characteristics;
2. For each $\theta \in \Theta$ and $y \in \mathcal{Y}$, the **sampling model** $p(y|\theta)$ describes our belief that y would be the outcome of the study if we knew θ to be true;
3. For each numerical value of $\theta \in \Theta$, the **posterior distribution** $p(\theta|y)$ describes our belief that θ is the true value, having observed dataset y.

Posterior Distribution

The posterior distribution is obtained from the prior distribution and sampling model via **Bayes' rule**:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int_\Theta p(y|\tilde{\theta})p(\tilde{\theta})d\tilde{\theta}}.$$

The Bayes' rule tells us how our beliefs should change after seeing new information.
In practice, however, since evaluating $\int_\Theta p(y|\tilde{\theta})p(\tilde{\theta})d\tilde{\theta}$ is often intractable, the posterior is instead obtained by

$$p(\theta|y) \propto p(y|\theta)p(\theta),$$

and the form of the right hand side can help us determine $p(\theta|y)$.

A Binomial Example

A survey is carried out to study the support rate θ $(0 < \theta < 1)$ of a policy. 100 people are surveyed, and a binary response Y_i is obtained from each person i $(i = 1, 2, \ldots, 100)$, $Y_i \sim \text{Bernoulli}(\theta)$ (that is, $Y = \sum_{i=1}^{100} Y_i \sim \text{Binomial}(100, \theta)$).

Before the survey, we believe that $\theta \sim \text{Beta}(5, 5)$, while the result of the survey is $Y = 60$. We'd like to obtain the posterior distribution of θ given the survey outcome.

A Binomial Example (Cont'd)

The prior distribution is $\theta \sim \text{Beta}(5,5)$, that is

$$p(\theta) = \frac{\theta^{5-1}(1-\theta)^{5-1}}{B(5,5)} \propto \theta^{5-1}(1-\theta)^{5-1}.$$

The sampling distribution is $Y \sim \text{Binomial}(100, \theta)$, that is, for each $\theta \in (0,1)$ and $y = 0, 1, \ldots, 100$,

$$P(Y = y|\theta) = \binom{100}{y} \theta^{y}(1-\theta)^{100-y}.$$

Using Bayes' rule, the posterior distribution of θ given that $Y = 60$ is

$$\begin{aligned} p(\theta|Y=60) &\propto p(Y=60|\theta)p(\theta) \\ &= \theta^{60}(1-\theta)^{100-60}\theta^{5-1}(1-\theta)^{5-1} \\ &= \theta^{65-1}(1-\theta)^{45-1}, \end{aligned}$$

which has the form of the p.d.f. of a Beta(65, 45) distribution. Thus, we have $\theta|Y = 60 \sim \text{Beta}(65, 45)$.

Review Exercise: Bayesian Analysis

Two tennis players, Serena and Venus, have played against each other 13 times in the past decade, with Serena winning 9 times. Assume that the outcome of a match between them is a binary variable Y ($Y = 1$ when Serena wins, $Y = 0$ when Venus does) that follows Bernoulli(θ) where $0 < \theta < 1$ is an unknown parameter, and we are interested in **estimating θ, Serena's winning rate against Venus**. We further assume that the outcomes of tennis matches, $Y_1, \ldots, Y_{13}$, are independent.

1. What is the maximum likelihood estimate of θ?
2. Suppose we ask a tennis expert, John, for prior information. John believes that Serena's winning rate is either 50% or 75%, and that these values are equally likely. Given the data, which value of θ do you think is more likely?
3. Another expert, Martina, suggests that we adopt a Beta$(9, 8)$ prior for θ upon analyzing match outcomes more than 10 years ago. What is the posterior distribution of θ given the match outcomes in the past decade? What is the posterior mean of θ?