QCB 408 / 508 – Notes on Week 4

Yingzi Huang

2020-03-13

Summary

- Likelihood & maximim likelihood estimation(MLE)
  - Sufficient statistic
  - Fisher Information
  - Optimality
  - Delta Method
- Exponential family distribution (EFDs)
  - Natural Single Parameter EFD
  - Calculating Moments
  - Maximum Likelihood
- Frequentist inference from pivital statistics
  - Point estimates
  - Confidence intervals
  - Hypothesis tests (next week)

Likelihood & maximim likelihood estimation(MLE)

Likelihood Function

Suppose random variable $X_1, X_2 \overset{iid}{\sim} F_\theta$, θ parameter should be informative about what we want to know about the population.
$(X_1, X_2, ..., X_n) \sim F_\theta$ is a joint distribution.
There are two levels in a study:

1. observed data: $x_1, x_2, ...x - n$
2. random variables $X_1, X_2, ...X_n$ that model the obtained data

Suppose that we observe $x_1, x_2, \ldots, x_n$ according to the model $X_1, X_2, \ldots, X_n \sim F_\theta$. The joint pdf is $f(\boldsymbol{x}; \theta)$. We view the pdf as being a function of $\boldsymbol{x}$ for a fixed θ.

The **likelihood function** is obtained by reversing the arguments and viewing this as a function of θ for a fixed, observed $\boldsymbol{x}$:

$$L(\theta; \boldsymbol{x}) = f(\boldsymbol{x}; \theta)$$

Log-Likelihood Function

$$\ell(\theta; \boldsymbol{x}) = \log L(\theta; \boldsymbol{x})$$

If the data are i.i.d., we have

$$\begin{aligned}\ell(\theta;\boldsymbol{x}) &= \log L(\theta;\boldsymbol{x}) \\ &= \log(f(\boldsymbol{x};\theta)) \\ &= \log\prod_{i=1}^{n} f(x_i;\theta) \\ &= \sum_{i=1}^{n}\log f(x_i;\theta) \\ &= \sum_{i=1}^{n}\ell(\theta;x_i)\end{aligned}$$

Sufficient Statistic

A statistic $T(\boldsymbol{x})$ is any function of the data. $T(\boldsymbol{x})$ is sufficient if $\boldsymbol{x}|T(\boldsymbol{x})$ does not depend on θ If $f(\boldsymbol{x};\theta) = g(T(\boldsymbol{x}))h(\boldsymbol{x})$, then $T(\boldsymbol{x})$ is sufficient. $L(\theta;\boldsymbol{x}) = g(T(\boldsymbol{x}))h(\boldsymbol{x}) \propto L(\theta;T(\boldsymbol{x}))$

Other topics:

- Minimal sufficient statistics
- complete sufficient statistics
- Ancillary statistics
- Basu's theorem

Maximum Likelihood Estimation

The **maximum likelihood estimate** is the value of θ that maximizes $L(\theta;\boldsymbol{x})$ for an observe data set $\boldsymbol{x}$.

$$\begin{aligned}\hat{\theta}_{\text{MLE}} &= \text{argmax}_\theta\, L(\theta;\boldsymbol{x}) \\ &= \text{argmax}_\theta\, \ell(\theta;\boldsymbol{x}) \\ &= \text{argmax}_\theta\, L(\theta;T(\boldsymbol{x}))\end{aligned}$$

where the last equality holds for sufficient statistics $T(\boldsymbol{x})$.

Example:

$$\begin{aligned}x &\sim \text{Binomial(n, p)} \\ L(p;x) &= \binom{n}{x}p^x(1-p)^{n-x} \\ &\propto p^x(1-p)^{n-x}\end{aligned}$$

$$\ell(p;x) \propto x\log(p) + (n-x)\log(1-p)$$

solve for p when

$$\frac{d}{dp}\ell(p;x) = 0 \Rightarrow \hat{p}_{MLE} = \frac{x}{n}$$

$$\frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \sim \text{Normal}(0,1) \text{ for large n}$$

Suppose we observe real life data $\hat{p} = 0.32$,

$$\begin{aligned}E[\hat{p}] &= p \\ \text{Var}(\hat{p}) &= \frac{p(1-p)}{n} \\ \hat{p} &= \frac{x}{n}\end{aligned}$$

We want to obtain the "sampling distribution" of $\hat{p}$: the distribution of $\hat{p} = \frac{x}{n}$ when the study is repeated. p of the population is unknown, so the distribution of $\hat{x}$.

However, a **pivotal statistic** does not involve the unknown p.

$$\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \sim \text{Normal}(0, 1)$$

Note that in general for a rv Y it is the case that $(Y - \text{E}[Y])/\sqrt{\text{Var}(Y)}$ has population mean 0 and variance 1.

Properties

When "certain regularity assumptions" are true, the following properties hold for MLEs.

- Consistent
- Equivariant
- Asymptotically Normal
- Asymptotically Efficient (or Optimal)
- Approximate Bayes Estimator

We will assume that the "certain regularity assumptions" are true in the following results.

MLE is "consistent":

An estimator is consistent if it converges in probability to the true parameter value. MLEs are consistent so that as $n \to \infty$,

$$\hat{\theta}_n \xrightarrow{P} \theta$$

where θ is the true value.

Equivariance:

If $\hat{\theta}_n$ is the MLE of θ, then $g\left(\hat{\theta}_n\right)$ is the MLE of $g(\theta)$.

Example:

For the Normal(μ, σ^2) the MLE of μ is $\overline{X}$. Therefore, the MLE of e^{μ} is $e^{\overline{X}}$.

Similarly, for Binomial(n, p), $\hat{p} = \frac{x}{n}$. $n\hat{p}(1-\hat{p})$ is the MLE of $\text{Var}(x) = np(1-p)$.

Fisher Information

The **Fisher Information** of $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} F_\theta$ is:

$$\begin{aligned} I_n(\theta) &= \text{Var}\left(\frac{d}{d\theta} \log f(\boldsymbol{X}; \theta)\right) \\ &= \sum_{i=1}^{n} \text{Var}\left(\frac{d}{d\theta} \log f(X_i; \theta)\right) \\ &= -\,\text{E}\left(\frac{d^2}{d\theta^2} \log f(\boldsymbol{X}; \theta)\right) \\ &= -\sum_{i=1}^{n} \text{E}\left(\frac{d^2}{d\theta^2} \log f(X_i; \theta)\right) \end{aligned}$$

Standard Error

In general, the **standard error** of an estimator is the standard deviation of sampling distribution of an estimate or statistic.

For MLEs, the standard error is $\sqrt{\text{Var}\left(\hat{\theta}_n\right)}$. It has the approximation

$$\text{se}\left(\hat{\theta}_n\right) \approx \frac{1}{\sqrt{I_n(\theta)}}$$

and the standard error estimate of an MLE is

$$\hat{\text{se}}\left(\hat{\theta}_n\right) = \frac{1}{\sqrt{I_n\left(\hat{\theta}_n\right)}}.$$

Asymptotic Normal

MLEs converge in distribution to the Normal distribution. Specifically, as $n \to \infty$,

$$\frac{\hat{\theta}_n - \theta}{\text{se}\left(\hat{\theta}_n\right)} \xrightarrow{D} \text{Normal}(0,1)$$

and

$$\frac{\hat{\theta}_n - \theta}{\hat{\text{se}}\left(\hat{\theta}_n\right)} \xrightarrow{D} \text{Normal}(0,1).$$

Example:

$X \sim \text{Binomial}(n,p)$, $I_n(p) = \frac{n}{p(1-p)}$

Asymptotic Pivotal Statistic:

By the previous result, we now have an approximate (asymptotic) pivotal statistic:

$$Z = \frac{\hat{\theta}_n - \theta}{\hat{\text{se}}\left(\hat{\theta}_n\right)} \xrightarrow{D} \text{Normal}(0,1).$$

This allows us to construct approximate confidence intervals and hypothesis test as in the idealized Normal(μ, σ^2) (with σ^2 known) scenario from the previous sections.

Optimality

The MLE is such that

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{D} \text{Normal}(0, \tau^2)$$

for some τ^2. Suppose that $\tilde{\theta}_n$ is any other estimate so that

$$\sqrt{n}\left(\tilde{\theta}_n - \theta\right) \xrightarrow{D} \text{Normal}(0, \gamma^2).$$

It follows that

$$\frac{\tau^2}{\gamma^2} \leq 1.$$

Delta Method

Suppose that $g()$ is a differentiable function and $g'(\theta) \neq 0$. Note that for some t in a neighborhood of θ, a first-order Taylor expansion tells us that $g(t) \approx g'(\theta)(t-\theta)$. From this we know that

$$\text{Var}\left(g(\hat{\theta}_n)\right) \approx g'(\theta)^2 \,\text{Var}(\hat{\theta}_n)$$

The delta method shows that $\hat{\text{se}}\left(g(\hat{\theta}_n)\right) = |g'(\hat{\theta}_n)|\hat{\text{se}}\left(\hat{\theta}_n\right)$ and

$$\frac{g(\hat{\theta}_n) - g(\theta)}{|g'(\hat{\theta}_n)|\hat{\text{se}}\left(\hat{\theta}_n\right)} \xrightarrow{D} \text{Normal}(0,1).$$

Delta Method Example

Suppose $X \sim \text{Binomial}(n,p)$ which has MLE, $\hat{p} = X/n$. By the equivariance property, the MLE of the per-trial variance $p(1-p)$ is $\hat{p}(1-\hat{p})$. It can be calculated that $\hat{\text{se}}(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n}$.

Let $g(p) = p(1-p)$. Then $g'(p) = 1-2p$. By the delta method,

$$\hat{\text{se}}\left(\hat{p}(1-\hat{p})\right) = |(1-2\hat{p})| \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Summary of MLE Statistics

In all of these scenarios, Z converges in distribution to $\text{Normal}(0,1)$ for large n.

Distribution	MLE	Std Err	Z Statistic
$\text{Binomial}(n,p)$	$\hat{p} = X/n$	$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$\frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$
$\text{Normal}(\mu,\sigma^2)$	$\hat{\mu} = \overline{X}$	$\frac{\hat{\sigma}}{\sqrt{n}}$	$\frac{\hat{\mu}-\mu}{\hat{\sigma}/\sqrt{n}}$
$\text{Poisson}(\lambda)$	$\hat{\lambda} = \overline{X}$	$\sqrt{\frac{\hat{\lambda}}{n}}$	$\frac{\hat{\lambda}-\lambda}{\sqrt{\hat{\lambda}/n}}$

Exponential Family Distributions (EFD)

Exponential family distributions (EFDs) provide a generalized parameterization and form of a very large class of distributions used in inference.

Definition

If X follows an EFD parameterized on the observed sclae by $\boldsymbol{\theta}$, then it has pdf of the form

$$f(x;\boldsymbol{\theta}) = h(x)\exp\left\{\sum_{k=1}^{d}\eta_k(\boldsymbol{\theta})T_k(x) - A(\boldsymbol{\eta})\right\}$$

where $\boldsymbol{\theta}$ is a vector of parameters, $\{T_k(x)\}$ are sufficient statistics, $A(\boldsymbol{\eta})$ is the cumulant generating function

The functions $\eta_k(\boldsymbol{\theta})$ for $k = 1, \ldots, d$ map the usual parameters to the "natural parameters".

$\{T_k(x)\}$ are sufficient statistics for $\{\eta_k\}$ due to the factorization theorem.

$A(\boldsymbol{\eta})$ is sometimes called the "log normalizer" because

$$A(\boldsymbol{\eta}) = \log\int h(x)\exp\left\{\sum_{k=1}^{d}\eta_k(\boldsymbol{\theta})T_k(x)\right\}.$$

Natural Single Parameter EFD

A natural single parameter EFD simplifies to the scenario where $d = 1$ and $T(x) = x$:

$$f(x;\eta) = h(x)\exp\{\eta x - A(\eta)\}$$

Example: Bernoulli

$$\begin{aligned} f(x;p) &= p^x(1-p)^{1-x} \\ &= \exp\{x\log(p) + (1-x)\log(1-p)\} \\ &= \exp\left\{x\log\left(\frac{p}{1-p}\right) + \log(1-p)\right\} \end{aligned}$$

$\eta(p) = \log\left(\frac{p}{1-p}\right)$

$T(x) = x$

$A(\eta) = \log\left(1 + e^{\eta}\right)$

Example: Normal

$$\begin{aligned} f(x;\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \\ &= \frac{1}{\sqrt{2\pi}}\exp\left\{\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \log(\sigma) - \frac{\mu^2}{2\sigma^2}\right\} \end{aligned}$$

$\boldsymbol{\eta}(\mu,\sigma^2) = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)^T$

$\boldsymbol{T}(x) = \left(x, x^2\right)^T$

$A(\boldsymbol{\eta}) = \log(\sigma) + \frac{\mu^2}{2\sigma^2} = -\frac{1}{2}\log(-2\eta_2) - \frac{\eta_1^2}{4\eta_2}$

Calculating Moments

$$\frac{d}{d\eta_k} A(\boldsymbol{\eta}) = \mathrm{E}[T_k(X)]$$

$$\frac{d^2}{d\eta_k^2} A(\boldsymbol{\eta}) = \mathrm{Var}[T_k(X)]$$

Example: Normal

For $X \sim \mathrm{Normal}(\mu, \sigma^2)$,

$$\mathrm{E}[X] = \frac{d}{d\eta_1} A(\boldsymbol{\eta}) = -\frac{\eta_1}{2\eta_2} = \mu,$$

$$\mathrm{Var}(X) = \frac{d^2}{d\eta_1^2} A(\boldsymbol{\eta}) = -\frac{1}{2\eta_2} = \sigma^2.$$

Maximum Likelihood

Suppose $X_1, X_2, \ldots, X_n$ are iid from some EFD. Then,

$$\ell(\boldsymbol{\eta}; \boldsymbol{x}) = \sum_{i=1}^{n} \left[\log h(x_i) + \sum_{k=1}^{d} \eta_k(\boldsymbol{\theta}) T_k(x_i) - A(\boldsymbol{\eta}) \right]$$

$$\frac{d}{d\eta_k} \ell(\boldsymbol{\eta}; \boldsymbol{x}) = \sum_{i=1}^{n} T_k(x_i) - n \frac{d}{d\eta_k} A(\boldsymbol{\eta})$$

Setting the second equation to 0, it follows that the MLE of η_k is the solution to

$$\frac{1}{n} \sum_{i=1}^{n} T_k(x_i) = \frac{d}{d\eta_k} A(\boldsymbol{\eta}).$$

Statistical Inference

We have observed data that is modeled by a probability generation process. The probability distribution has parameters informative about the population. **Statistical inference** reverse engineers this forward to estimate parameters and provide measures of uncertainty about the estimate.

- A **parameter** is a number that describes a population
  - It is often a fixed number and we usually do not know its value
- A **statistic** is a number calculated from a sample of data
  - A statistic is used to estimate a parameter
- The **sampling distribution** of a statistic is the probability disribution of the statistic under repeated realizations of the data from the assumed data generating probability distribution.

The sampling distribution connects a calculated statistic to the population (probability model).

Inference Goals and Strategies

Data collected in such a way that there exists a reasonable probability model for this process that involves parameters informative about the population.
so we have data $x_1, x_2, ...x_n$ and model $X_1, X_2, ..., X_n \sim \mathrm{F}_\theta$

Common Goals:

1. Form point estimates the parameter θ
2. Confidence interval of θ
   - Quantify uncertainty on the estimates
3. Hypotheses test on the parameters
   - assesses specific value(s) of θ

1. Point Estimation

See example MLE $\hat{\theta_n}$

2. Confidence Intervals (CI)

Once we have a point estimate of a parameter, we would like to know its uncertainty. We interpret this measure of uncertainty in terms of hypothetical repetitions of the sampling scheme we used to collect the original data set.

for MLEs:

Confidence intervals take the form

$$(\hat{\theta} - C_\ell, \hat{\theta} + C_u)$$

where

$$\Pr(\hat{\theta} - C_\ell \leq \theta \leq \hat{\theta} + C_u; \theta)$$

forms the "level" or coverage probablity.

Approximate 95% CI for MLEs:

$$\begin{aligned}
0.95 &\approx \Pr(-1.96 \leq \frac{\hat{\theta} - \theta}{\hat{se}(\hat{\theta})} \leq 1.96) \\
&= \Pr(-1.96\hat{se}(\hat{\theta}) \leq \hat{\theta} - \theta \leq 1.96\hat{se}(\hat{\theta})) \\
&= \Pr(\theta - 1.96\hat{se}(\hat{\theta}) \leq \hat{\theta} \leq \theta + 1.96\hat{se}(\hat{\theta}) \\
&= \Pr(-1.96\hat{se}(\hat{\theta}) \leq \theta - \hat{\theta} \leq 1.96\hat{se}(\hat{\theta}) \\
&= \Pr(\hat{\theta} - 1.96\hat{se}(\hat{\theta}) \leq \theta \leq \hat{\theta} + 1.96\hat{se}(\hat{\theta})
\end{aligned}$$

So 95% approx. CI is

$$(\hat{\theta} - 1.96\hat{se}(\hat{\theta}), \hat{\theta} + 1.96\hat{se}(\hat{\theta})$$

$(1-\alpha)$-Level CIs

If $Z \sim$ Normal(0,1), then $\Pr(-|z_{\alpha/2}| \leq Z \leq |z_{\alpha/2}|) = 1 - \alpha$.

Repeating the steps from the 95% CI case, we get the following is a $(1-\alpha)$-Level CI for $\hat{\theta}$:

$$\left(\hat{\theta} - |z_{\alpha/2}|\hat{se}(\hat{\theta}), \hat{\theta} + |z_{\alpha/2}|\hat{se}(\hat{\theta})\right)$$

z_α is the α-percentile of Normal(0,1).

One-Sided CIs

The CIs we have considered so far are "two-sided". Sometimes we are also interested in "one-sided" CIs.

If $Z \sim$ Normal(0,1), then $1 - \alpha = \Pr(Z \geq -|z_\alpha|)$ and $1 - \alpha = \Pr(Z \leq |z_\alpha|)$. We can use this fact along with the earlier derivations to show that the following are valid CIs:

$$(1-\alpha)\text{-level upper: } \left(-\infty, \hat{\theta} + |z_\alpha|\hat{se}(\hat{\theta})\right)$$

$$(1-\alpha)\text{-level lower: } \left(\hat{\theta} - |z_\alpha|\hat{se}(\hat{\theta}), \infty\right)$$

Session Information

```
> sessionInfo()
R version 3.6.1 (2019-07-05)
Platform: x86_64-apple-darwin15.6.0 (64-bit)
Running under: macOS Catalina 10.15.3

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] gapminder_0.3.0 forcats_0.4.0   stringr_1.4.0   dplyr_0.8.3
 [5] purrr_0.3.2     readr_1.3.1     tidyr_1.0.0     tibble_2.1.3
 [9] ggplot2_3.2.1   tidyverse_1.2.1 knitr_1.24

loaded via a namespace (and not attached):
 [1] Rcpp_1.0.2       cellranger_1.1.0 pillar_1.4.2     compiler_3.6.1
 [5] tools_3.6.1      zeallot_0.1.0    digest_0.6.20    lubridate_1.7.4
 [9] jsonlite_1.6     evaluate_0.14    lifecycle_0.1.0  nlme_3.1-140
[13] gtable_0.3.0     lattice_0.20-38  pkgconfig_2.0.3  rlang_0.4.0
[17] cli_1.1.0        rstudioapi_0.10  yaml_2.2.0       haven_2.1.1
[21] xfun_0.9         withr_2.1.2      xml2_1.2.2       httr_1.4.1
[25] hms_0.5.1        generics_0.0.2   vctrs_0.2.0      grid_3.6.1
[29] tidyselect_0.2.5 glue_1.3.1       R6_2.4.0         readxl_1.3.1
[33] rmarkdown_1.15   modelr_0.1.5     magrittr_1.5     backports_1.1.4
[37] scales_1.0.0     htmltools_0.3.6  rvest_0.3.4      assertthat_0.2.1
[41] colorspace_1.4-1 stringi_1.4.3    lazyeval_0.2.2   munsell_0.5.0
[45] broom_0.5.2      crayon_1.3.4
```