Section 3: Statistical Inference

September 2015

1 Statistical Inference

1.1 Probability and Statistics

- The basic problem of probability is: Given the distribution of the data, what are the properties (e.g. its expectation) of the outcomes (i.e. the data)?
- The basic problem of statistical inference is the inverse of probability: Given the outcomes, what can we say about the process that generated the data? In other words:

  Given a sample $X_1, ..., X_n \sim F$, what can we say about F?

2 Fundamental Concepts in Statistical Inference

2.1 Point estimation

- *Point estimation:* refers to providing a single "best guess" of some parameter. More formally, let $X_1, \cdots, X_n$ be n *iid* data points from some distribution F. A point estimator $\hat{\theta}_n$ of a parameter θ is some function of $X_1, \cdots, X_n$:
$$\hat{\theta}_n = g(X_1, \cdots, X_n).$$
- Example: Suppose $X_1, ..., X_n \sim F$, we may estimate $\theta = E[X]$ by $\hat{\theta}_n = \sum_{i=1}^n X_i$.

2.2 Confidence sets (Credible set)

A $1-\alpha$ confidence interval for a parameter θ is an interval $C_n = (a, b)$ where $a = a(X_1, ..., X_n)$ and $b = b(X_1, ..., X_n)$ are functions of the data such that $P_\theta(\theta \in C_n) \geq 1 - \alpha$, for all $\theta \in \Theta$. In words, (a, b) traps θ with probability $1 - \alpha$. We call $1 - \alpha$ the coverage of the confidence interval.

2.3 Hypothesis testing

- It is often the case that we wish to use data to make a binary decision about some unknown aspect of nature. For example, we may wish to decide whether or not it is plausible that a parameter takes some particular value.
- A hypothesis is any statement about an unknown aspect of a distribution. In a hypothesis test, we have two hypotheses:
  - H_0, the null hypothesis, and
  - H_1, the alternative hypothesis.
- Often a hypothesis is stated in terms of the value of one or more unknown parameters, in which case it is called a parametric hypothesis. Specifically, suppose we have an unknown parameter θ. Then parametric hypotheses about θ can be written in general as $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$, where Θ_0 and Θ_1 are disjoint, i.e., $\Theta_0 \cap \Theta_1 = \emptyset$.

2.4 Frequentist and Bayesian

- The two dominant approaches in statistical inference are called frequentist inference and Bayesian inference.
- The most basic difference between these two perspective is frequentists assume there is only one true unknown parameter, while the Bayesians are comfortable treating the unknown parameter as random and putting a prior distribution over it.
- We will be focusing on frequentist point estimation in our following discussion.

3 Point Estimation: MLE

3.1 Estimator and Estimates(skipped in class)

Suppose we have an unknown parameter θ and some data $X = (X_1, \cdots, X_n)$. We may with to use the data to estimate θ.

- An estimator $\hat{\theta}$ of an unknown parameter θ is any function of the data that is intended to approximate θ in some sense. Although we typically just write $\hat{\theta}$, it is actually $\hat{\theta}(X)$, a random variable.
- An estimate is the value an estimator takes for a particular set of data values. Thus, the estimator $\hat{\theta}(X)$ would yield the estimate $\hat{\theta}(x)$ if we observe the data $X = x$.
- Any function of the data can be considered an estimator for any parameter. However, it may not be a good estimator.

3.2 Likelihood Function

Definition: Let θ be some unknown parameter, and X be a random variable with density function $f(x;\theta)$. Let $X_1, X_2, \cdots, X_n$ be a set of iid sample of X. The likelihood function for this particular set of data values x is $L(\theta;x) = \prod_{i=1}^{n} f(x;\theta)$

3.2.1 Interpretation of Likelihood Function

- Mathematically, it is the same object with the joint density of the sample. But they should be interpreted differently:
  - For $f(x;\theta)$, we think about fixing a parameter value θ and allowing x to vary.
  - For $L(\theta;x)$, we think about fixing a collection of sample values x and allowing θ to vary.
- In short, a likelihood function is just a conditional probability distribution where the parameter conditioned on can vary.

3.2.2 What the likelihood function is NOT

The likelihood function is NOT the density function of θ given the data, regardless which perspective one might take:

- In frequentist inference, the unknown parameter θ is not a random variable, so talking about its "distribution" makes no sense.
- Even in Bayesian inference, the likelihood is still the same mathematical object as the pmf or pdf of the data. Hence, it describes probabilities of observing data values given certain parameter values, not the other way around.
- The likelihood does not (in general) sum or integrate to 1 when summing or integrating over θ. In fact, it may sum or integrate to ∞, in which case we cannot even scale it to make it a pdf (or pmf).

3.3 Maximum likelihood estimation

- An estimate $\hat{\theta}(x)$ of θ is called a maximum likelihood estimate (MLE) of θ if it maximizes $L(x;\theta)$ over Θ, the domain for θ. An estimator that takes the value of a maximum likelihood estimate for every possible sample $X = x$ is called a maximum likelihood estimator. (also called MLE).

$$\hat{\theta_{MLE}} = \underset{\theta\in\Theta}{\operatorname{argmax}}\, L(x;\theta)$$

3.3.1 Finding the MLE estimator

- It is often more convenient to work with the logarithm of the likelihood, $l(x;\theta) = L(x;\theta)$.
- Finding the MLE estimator is just a optimization problem. Typically we find all points in Θ where the derivaive $\frac{\partial L(x;\theta)}{\partial \theta}$ is zero.
- In some cases, we can find closed form solution. Otherwise, we need to resort to numerical calculation. When the likelihood function is convex, we are guarantee to find the global optimal.

3.3.2 Some examples of MLE

Distribution	parameter	$E[X]$	MLE
$\mathcal{N}(\mu, \sigma^2)$	μ	μ	$\bar{x}$
$Pois(\lambda)$	λ	λ	$\bar{x}$
$Exp(\lambda)$	λ	λ	$\bar{x}$
$Bern(p)$	p	p	$\bar{x}$
GEM(p)	p	$\frac{1}{p}$	$\frac{1}{\bar{x}}$

3.3.3 MLE estimator and M-projection

- One may recall that most MLE have very simple and intuitive form. This is not a accident but an important property of the exponential family. (One can take cs 228 for more details.)

3.3.4 Existence and Uniqueness

- The maximum likelihood estimator need not be unique or even exist.
- It may be the case that for certain possible samples $X = x$, the likelihood function $L(x;\theta)$ has a non-unique maximum or fails to achieve its maximum altogether.

3.3.5 Invariance to Reparametrization

Theorem Let $\hat{\theta}_{MLE}$ be a maximum likelihood estimator of θ over the parameter space Θ, and let g be a function with domain Θ and image Ξ. Then $\hat{\xi}_{MLE} = g(\hat{\theta}_{MLE})$ is a maximum likelihood estimator of $\xi_{MLE} = g(\theta_{MLE})$ over the parameter space Ξ.

3.3.6 Estimators that Optimize Other Functions

There are situations in which we may want to find an estimator by optimizing some real-valued function other than the likelihood. For example,

- The likelihood function itself may be difficult to work with.
- We may be unsure of some aspect of our model (e.g., we may not know if the observations are normally distributed).
- We may want to favor certain kinds of estimates over others (e.g. sparse estimator).

3.4 M-estimator

An estimator that is found by optimizing some real-valued function other than the likelihood is called an *M-estimator*.

3.4.1 The least square estimator

- Consider the following regression setup: Let $X = (x_{ij})$ be an np matrix of known constants (not random variables, despite the capital letter), and let $Y1, ..., Yn$ be independent random variables with

$$Y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i$$

for each $i \in \{1, 2, \cdots, n\}$ where $\beta = (\beta_1, \cdots, \beta_j) \in R_p$ and ϵ_i being the noise. the least square estimator

$$\hat{\beta}_{LS} = \text{argmin} \, \|Y - X\beta\|_2^2$$

is an M-estimator of β. If we assume Gaussian noise, $\hat{\beta}_{LS} = \hat{\beta}_{MLE}$

3.4.2 The lasso

- In the same regression setup, the lasso estimator

$$\hat{\beta}_{Lasso} = \text{argmin} \, \|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

is another M-estimator of β where $\lambda > 0$ is some fixed constant.

- The addition of the term $\lambda\|\beta\|_1$ does several things:
  - It typically results in an estimate for which some components are exactly zero.
  - $\hat{\beta}_{Lasso}$ is unique in many cases when $\hat{\beta}_{MLE}$ is not. For example, $\hat{\beta}_{Lasso}$ is usually still unique even when $p > n$