Lecture 3:

Parametric Inference

Parametric Inference

- The problem of parametric inference is, having a parametric model: $\mathcal{F} = \left\{ f(x;\theta) : \ \theta \in \Theta \right\}$ where $\Theta \subset \mathbb{R}^k$ is the **parameter space** and $\theta = (\theta_1, \ldots, \theta_k)$ is the parameter – to somehow estimate the parameter θ from observarions, x-s.

- Good question is: how do we know the data's distribution is from this parametric model, not any other? Indeed, we rarely know this for sure, this is why nonparametric methods are preferable. But:

    1. Some parametric models provide a good enough approximation (Central Limit Theorem, etc)

    2. Parametric inference provides the background for understanding some nonparametric methods

Parametric Inference

- We will now briefly remember the terminology on 2 examples
- And then cover 2 most important and well-known methods of parametric inference:

    1. The **method of moments**
    2. The **method of maximum likelihood**

Parameter of interest

- We are often only interested in some function $T(\theta)$. For example, if $X \sim \mathcal{N}(\mu, \sigma)$ then the parameter is $\theta = (\mu, \sigma)$. If our goal is to estimate μ, then $\mu = T(\theta)$ is called the **parameter of interest** and σ is called a **nuisance parameter**. The parameter of interest might be quite a complicated function of θ:

- **Example 1**: Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma)$. Suppose that X_i is the outcome of a blood test and suppose we are interested in τ, the fraction of the population whose test score is larger than 1. Then:

$$\tau = 1 - \mathbb{P}\left(Z = \frac{X - \mu}{\sigma} < \frac{1 - \mu}{\sigma}\right) = 1 - \Phi\left(\frac{1 - \mu}{\sigma}\right)$$ so the parameter of interest is $\tau = T(\mu, \sigma) = 1 - \Phi\big((1 - \mu)/\sigma\big)$

Parameter of interest

- **Example 2**: For $X \sim \text{Gamma}(\alpha, \beta)$ – being Gamma-distributed, the pdf is

$$f(x; \alpha, \beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} \exp(-x/\beta), \quad x > 0$$

where $\alpha, \beta > 0$ and $\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1} e^{-y}\, dy$ is the Gamma-function. This distribution is sometimes used to model lifetimes of people, animals, and electronic equipment. Suppose we want to estimate the mean lifetime. Then

$$T(\alpha, \beta) = \mathbb{E}_{\theta}(X) = \alpha\beta$$

The Method of Moments

The Method of Moments

- The idea of this method goes back to Chebyshev and Pearson. The estimators that it gives often aren't optimal, but are easy to compute, and also useful as starting values for iterative methods.

- Let $\theta = (\theta_1, \ldots, \theta_M)$ have M components. Define M **moments**:

$$\alpha_m(\theta) = \mathbb{E}_\theta(X^m) = \int x^m \, dF_\theta(x), \quad 1 \le m \le M, \text{ and the } m\text{-th}$$

sample moment is $\hat{\alpha}_m = \frac{1}{n} \sum_{i=1}^{n} X_i^m$

- Then the **m.o.m. estimator** $\hat{\theta}$ is such that $\alpha_m(\hat{\theta}) = \hat{\alpha}_m$ for all moments (that gives a system of M equations with M unknowns).

The Method of Moments

- **Example 1**: $X_1, \ldots, X_n \sim$ Bernoulli(p). Then $\alpha_1 = \mathbb{E}_p(X) = p$ and $\hat{\alpha}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$ – by equating the two, the estimator $\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

- **Example 2**: $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma)$. Then $\alpha_1 = \mathbb{E}_\theta(X) = \mu$ and $\alpha_2 = \mathbb{E}_\theta(X^2) = \mathbb{V}_\theta(X) + \left(\mathbb{E}_\theta(X)\right)^2 = \sigma^2 + \mu^2$. So:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad \hat{\sigma}^2 + \hat{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

– a system of 2 equations on 2 unknowns. The solution is $\hat{\mu} = \overline{X}_n$ – sample mean and $\hat{\sigma}^2 = n^{-1}\sum (X_i - \overline{X}_n)^2$ – sample s.t.d.

The Method of Moments

- **Example 1**: $X_1, \ldots, X_n \sim$ Bernoulli(p). Then $\alpha_1 = \mathbb{E}_p(X) = p$ and $\hat{\alpha}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$ – by equating the two, the estimator $\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

- **Example 2**: $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma)$. Then $\alpha_1 = \mathbb{E}_\theta(X) = \mu$ and $\alpha_2 = \mathbb{E}_\theta(X^2) = \mathbb{V}_\theta(X) + \left(\mathbb{E}_\theta(X)\right)^2 = \sigma^2 + \mu^2$. So:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad \hat{\sigma}^2 + \hat{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

– a system of 2 equations on 2 unknowns. The solution is $\hat{\mu} = \overline{X}_n$ – sample mean and $\hat{\sigma}^2 = n^{-1}\sum (X_i - \overline{X}_n)^2$ – sample s.t.d.

The Method of Moments

- **Theorem**: Let $\hat{\theta}_n$ denote the m.o.m. estimator. Under appropriate conditions on the model, the following holds:

    1. The estimate $\hat{\theta}_n$ exists with probability tending to 1.

    2. It is consistent: $\hat{\theta}_n \xrightarrow{P} \theta$

    3. It is asymptotically normal: $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0,\Sigma)$

where $\Sigma = g\,\mathbb{E}_\theta(YY^T)\,g^T, \quad Y = (X, X^2, \ldots, X^M)^T$ and $g_i = \dfrac{\partial \alpha_i^{-1}(\theta)}{\partial \theta}$. This can be used to find the s.e. and confidence intervals. Bootstrap can also be used and is easier.

The Method of Maximum Likelihood

Maximum Likelihood

- Let $X_1, \ldots, X_n$ be IID with PDF $f(x; \theta)$. Then the **likelihood function** is defined by $\mathscr{L}_n(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$. The logarithm of it is called the **log-likelihood** function $\ell_n(\theta) = \log \mathscr{L}_n(\theta)$. So likelihood is just the joint density of the data, but we **treat it as a function of the parameter** θ. Thus, $\mathscr{L}_n : \Theta \to [0, \infty)$. It is **not a density function**: in general, it does not integrate to 1 (w.r.t. θ)

- The **maximum likelihood estimator**, MLE, is the value that maximizes $\mathscr{L}_n(\theta)$ (or its logarithm, which is often easier):

$$\hat{\theta}_n = \arg\max_{\theta} \mathscr{L}_n(\theta) = \arg\max_{\theta} \ell_n(\theta)$$

Maximum Likelihood

- Multiplying $\mathscr{L}_n(\theta)$ by any positive constant (not depending on θ) does not change the MLE. So we shall often drop such constants.

- **Example**: Suppose $X_1, \ldots, X_n \sim$ Bernoulli(p). The pmf is $f(x;\theta) = p^x(1-p)^{1-x}$ for $x = 0,1$. So

$$\mathscr{L}_n(p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^S(1-p)^{n-S} \text{ where } S = \sum_i X_i$$

So $\ell_n(p) = S\log p + (n-S)\log(1-p)$, the MLE $\hat{p}_n = S/n$

- **Example**: Recall MLE for $X \sim \mathcal{N}(\mu, \sigma)$ from last term, then $\hat{\mu} = \overline{X}_n$ and $\hat{\sigma}$ is the sample s.t.d.

Maximum Likelihood

- **Example**: Let $X_1, \ldots, X_n \sim U(0,\theta)$. Recall that

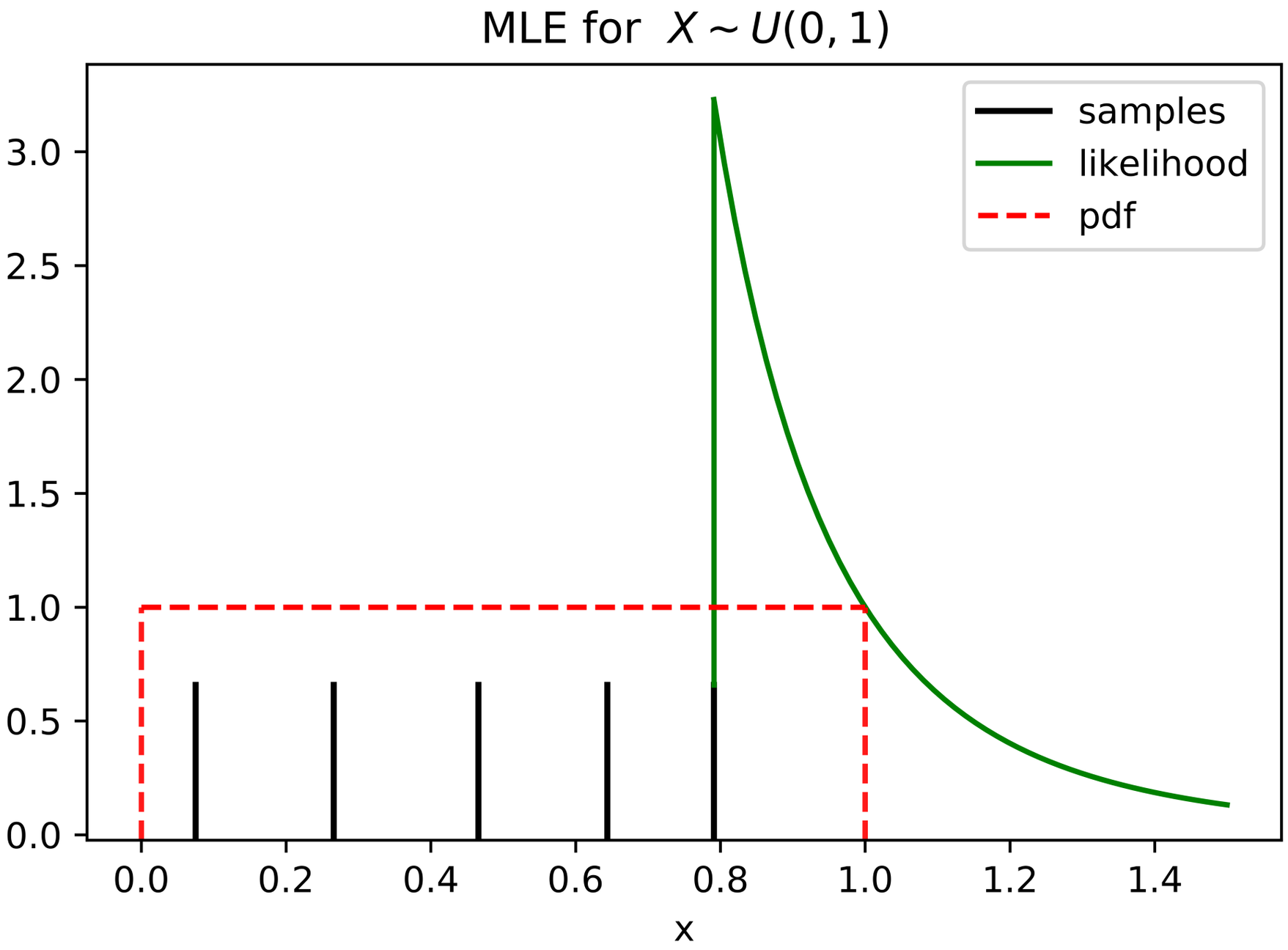
$$f(x;\theta) = \begin{cases} 1/\theta, & 0 \le x \le \theta \\ 0 & \text{otherwise} \end{cases}$$

. Suppose some $X_i > \theta$ – then $f(X_i;\theta) = 0$. So $\mathscr{L}_n(\theta) = 0$ if $X_{\max} = \max(X_1, \ldots, X_n) > \theta$. On the other hand, if $X_{\max} \le \theta$, $\mathscr{L}_n(\theta) = \prod_i f(X_i;\theta) = \theta^{-n}$. So

overall $\mathscr{L}_n(\theta) = \begin{cases} \left(\frac{1}{\theta}\right)^n & X_{\max} \le \theta \\ 0 & \theta \le X_{\max} \end{cases}$ which is strictly decreasing after θ. Thus the MLE is $\hat{\theta}_n = X_{\max}$

Maximum Likelihood

- **Example**: Let $X_1, \ldots, X_n \sim U(0,\theta)$. The MLE is $\hat{\theta}_n = X_{\max}$



Maximum Likelihood

- **Theorem(s)**: Under certain conditions on the model the MLE:

    1. Is **consistent**: $\hat{\theta}_n \xrightarrow{P} \theta$

    2. Is **equivariant**: If $\hat{\theta}_n$ is MLE of θ, then $g(\hat{\theta}_n)$ is MLE of $g(\theta)$

    3. Is **asymptotically normal** $(\hat{\theta} - \theta)/\hat{\text{se}} \xrightarrow{d} \mathcal{N}(0,1)$ (and $\hat{\text{se}}$ can often be computed analytically

    4. Is **asymptotically optimal** / **efficient** – among all good estimators it has the smallest variance (at least for large samples)

    5. Is approximately the **Bayes estimator** (more on that later)

MLE Consistency

- **Consistency** means MLE converges to the true value. One can formulate the fact that $\hat{\theta}_n \xrightarrow{P} \theta$ in terms of **Kullback-Leibler divergence**:

$$KL(f, g) = \int f(x) \log \frac{f(x)}{g(x)} dx$$ – a measure of "distance" between distributions: $D(f, g) \geq 0$ and $D(f, f) = 0$, but $D(f, g) \neq D(g, f)$.

For parameters $\theta, \psi \in \Theta$ we track $D(f(x; \theta), f(x; \psi))$.

One can prove that with sample size $\to \infty$, maximizing log-likelihood = minimizing $KL(\theta_{\text{true}}, \theta)$. (Seminar material)

MLE Equivariance

- **Equivariance**: If $\hat{\theta}_n$ is MLE of θ, then $\hat{\tau}_n = g(\hat{\theta}_n)$ is MLE of $\tau = g(\theta)$

- **Proof**: Let $h = g^{-1}$ be the inverse of g. Then $\hat{\theta}_n = h(\hat{\tau}_n)$. For any τ, we have $\mathscr{L}(\tau) = \prod_i f(x_i;\, h(\tau)) = \mathscr{L}(\theta)$ where $\theta = h(\tau)$.

  Hence, for any τ, we have $\mathscr{L}_n(\tau) = \mathscr{L}(\theta) \leq \mathscr{L}(\hat{\theta}) = \mathscr{L}_n(\hat{\tau})$

- **Example**: Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu = \theta, 1)$. The MLE for θ is $\hat{\theta}_n = \overline{X}_n$. Let $\tau = e^{\theta}$. Then, the MLE for τ is $\hat{\tau} = e^{\hat{\theta}} = e^{\overline{X}}$.

MLE Asymptotic Normality

- Another important concept, defined in terms of **score function** $s(X;\theta) = \dfrac{\partial \log f(X;\theta)}{\partial \theta}$, is the **Fisher information**:

$$I_n(\theta) = \mathbb{V}_\theta\left(\sum_{i=1}^{n} s(X_i;\theta)\right) = \sum \mathbb{V}$$

It can be shown that $\mathbb{E}_\theta(s(X;\theta)) = 0$, so $\mathbb{V}_\theta(s) = \mathbb{E}_\theta(s^2)$. Denoting $I_1(\theta)$ for 1 observation by $I(\theta)$, for IID X-s we have **Theorem:**

$$I_n(\theta) = nI(\theta), \text{ and, actually, } I(\theta) = -\mathbb{E}_\theta\left(\frac{\partial^2 \log f(X;\theta)}{\partial \theta^2}\right)$$

MLE Asymptotic Normality

- **Theorem** (Asymptotic Normality of MLE): Let $\text{se} = \sqrt{\mathbb{V}(\hat{\theta}_n)}$.

  Under certain regularity conditions,

  1. $\text{se} \approx \sqrt{1/I_n(\theta)}$ and $\dfrac{\hat{\theta}_n - \theta}{\text{se}} \xrightarrow{d} \mathcal{N}(0,1)$
  2. Let $\hat{\text{se}} \approx \sqrt{1/I_n(\hat{\theta})}$ – then $\dfrac{\hat{\theta}_n - \theta}{\hat{\text{se}}} \xrightarrow{d} \mathcal{N}(0,1)$

In a nutshell, the distribution of MLE is $\approx \mathcal{N}(\theta, \hat{\text{se}})$.

We can use that to construct **normal confidence intervals**!