Point Estimation

Andrew Nobel

October, 2024

Overview of Point Estimation

Model: Family $\mathcal{P} = \{f(\cdot|\theta) : \theta \in \Theta\}$ of distributions on sample space $\mathcal{X}$

- Assume parameter space $\Theta \subseteq \mathbb{R}^p$. Our focus is $p = 1, 2$

Task: Given observation $X \sim f(\cdot|\theta) \in \mathcal{P}$, wish to estimate θ

- An estimator is a function $\hat{\theta} : \mathcal{X} \to \Theta$

Evaluation

- Loss function $\ell : \Theta \times \Theta \to \mathbb{R}$. Often squared error $\ell(\theta, \theta') = (\theta - \theta')^2$
- Performance of estimator $\hat{\theta}$ evaluated via risk function $R(\theta, \hat{\theta}) = \mathbb{E}_\theta \, \ell(\theta, \hat{\theta}(X))$

Method of Moments

Method of Moments

Motivation: For many families $\mathcal{P} = \{f(\cdot|\theta) : \theta \in \Theta\}$

- First moment $\mathbb{E}_\theta X$ and second moment $\mathbb{E}_\theta X^2$ are nice functions of θ
- These functions can be inverted to express θ in terms of $\mathbb{E}_\theta X$ and $\mathbb{E}_\theta X^2$

MoM: Given iid observations $X_1, \ldots, X_n \sim f(\cdot|\theta)$

- Express θ in terms of $\mathbb{E}_\theta X$, and $\mathbb{E}_\theta X^2$ if needed
- Replace unknown value $\mathbb{E}_\theta X$ by sample mean $\text{Avg}(X_i) = \frac{1}{n}\sum_{i=1}^n X_i$
- Replace unknown value $\mathbb{E}_\theta X^2$ by sample estimates $\text{Avg}(X_i^2) = \frac{1}{n}\sum_{i=1}^n X_i^2$

Example: Geometric Distribution

Geometric: For $\theta \in (0,1)$ the geometric distribution Geom(θ) has pmf

$$f(x|\theta) = (1-\theta)^{x-1}\theta \text{ for } x = 1, 2, \ldots$$

If $X \sim f(\cdot|\theta)$ a simple calculation shows that $\mathbb{E}_\theta X = \theta^{-1}$. Inverting this relation gives

$$\theta = \frac{1}{\mathbb{E}_\theta X}$$

MoM Estimator: The method of moments estimator for θ is

$$\hat{\theta}_{\text{MoM}}(x) = \frac{1}{\text{Avg}(x_i)}$$

Example: Normal Distribution

Normal: Let $X \sim \mathcal{N}(\mu, \sigma^2)$ with $\theta = (\mu, \sigma^2)$. Then

$$\mathbb{E}_\theta X = \mu \quad \text{and} \quad \mathbb{E}_\theta X^2 = \mathrm{Var}_\theta(X) + (\mathbb{E}_\theta X)^2 = \sigma^2 + \mu^2$$

Inverting these relations gives

$$\mu = \mathbb{E}_\theta X \quad \text{and} \quad \sigma^2 = \mathbb{E}_\theta X^2 - (\mathbb{E}_\theta X)^2$$

MoM Estimator: The method of moments estimators for μ, σ^2 are

$$\hat{\mu}(x) = \mathsf{Avg}(x_i) \quad \text{and} \quad \widehat{\sigma}^2(x) = \mathsf{Avg}(x_i^2) - \mathsf{Avg}(x_i)^2$$

Example: Beta Distribution

Beta: Let $X \sim \text{Beta}(\alpha, \beta)$ and $\theta = (\alpha, \beta)$. Then

$$\mathbb{E}_\theta X = \frac{\alpha}{\alpha+\beta} \qquad \mathbb{E}_\theta X^2 = \frac{(\alpha+1)\alpha}{(\alpha+\beta+1)(\alpha+\beta)}$$

MoM Estimator: The method of moments estimators for α, β are the solutions $\hat{\alpha}(x), \hat{\beta}(x)$ of the equations

$$\text{Avg}(x_i) = \frac{\alpha}{\alpha+\beta}$$

$$\text{Avg}(x_i^2) = \frac{(\alpha+1)\alpha}{(\alpha+\beta+1)(\alpha+\beta)}$$

Maximum Likelihood Estimation

The Likelihood Function

Setting: Family $\mathcal{P} = \{f(\cdot|\theta) : \theta \in \Theta\}$ of distributions on $\mathcal{X}$, with $\Theta \subseteq \mathbb{R}^p$

- Observation $X \sim f(\cdot|\theta) \in \mathcal{P}$ yields data $x \in \mathcal{X}$

Definition: The *likelihood function* $L(\cdot|x) : \Theta \to [0, \infty)$ for the data x is given by

$$L(\theta|x) := f(x|\theta)$$

Idea: Likelihood measures support for different parameters values based on data

- $f(x|\theta_1) \geq f(x|\theta_2)$: data x is more likely under θ_1 than θ_2
- $L(\theta_1|x) \geq L(\theta_2|x)$: more evidence for θ_1 than θ_2 based on data x

Maximum Likelihood Estimation

Idea: Identify the parameter $\theta \in \Theta$ most supported by the data x

Definition: The maximum likelihood estimator $\hat{\theta}_{\text{MLE}} : \mathcal{X} \to \Theta$ is given by

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

Issues

- How to find MLEs
- Theoretical properties

The Log Likelihood

Definition: The *log-likelihood* function $\ell(\theta|x) := \log L(\theta|x)$

Rationale: If $X = X_1, \ldots, X_n \sim f(\cdot|\theta)$ independent then

$$\ell(\theta|x) = \log\left(\prod_{i=1}^{n} f(x_i|\theta)\right) = \sum_{i=1}^{n} \log f(x_i|\theta)$$

Note: As $\log(\cdot)$ is strictly increasing on $(0, \infty)$

$$\operatorname*{argmax}_{\theta\in\Theta} L(\theta|x) = \operatorname*{argmax}_{\theta\in\Theta} \ell(\theta|x)$$

Thus $\hat{\theta}_{\mathsf{MLE}}(x)$ can be found by maximizing $\ell(\theta|x)$ rather than $L(\theta|x)$

Finding MLE: Maximizing the Log Likelihood

Approaches

- Calculus or direct arguments
- Use of optimization software, iterative methods and approximations

Calculus: Check first and second order conditions for local maxima

- Solve equations $\frac{\partial}{\partial \theta_j} \ell(\theta|x) = 0$ for $j = 1, \ldots, p$ to find a candidate θ^*
- Verify that θ^* is a local maximum using second derivatives or direct arguments

Issues: Is θ^* unique, is it a local or a global maximum?

MLE Examples

Bernoulli. Observe $X = X_1, \ldots, X_n$ iid $\sim$ Bern(θ) with $\theta \in [0, 1]$

- $\hat{\theta}_{\text{MLE}}(x) = \overline{x}$
- If $\Theta = [\frac{1}{2}, 1]$ then $\hat{\theta}_{\text{MLE}}(x) = \max\{\frac{1}{2}, \overline{x}\}$

Uniform. Observe $X = X_1, \ldots, X_n$ iid $\sim$ Unif$(0, \theta)$ with $\theta \in (0, \infty)$

- $\hat{\theta}_{\text{MLE}}(x) = \max_i x_i$

Normal. Observe $X = X_1, \ldots, X_n$ iid $\sim \mathcal{N}(\mu, \sigma^2)$ with $\theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$

- $\hat{\theta}_{\text{MLE}}(x) = (\overline{x}, s^2)$

Bayesian Point Estimation

Review of Bayesian Inference

Ingredients

- Family $\mathcal{P} = \{f(\cdot|\theta) : \theta \in \Theta\}$ of distributions on $\mathcal{X}$
- Prior density $\pi(\theta)$ on parameter space Θ
- Joint densitiy $f(x, \theta) = f(x|\theta)\pi(\theta)$, marginal density $m(x) = \int f(x, \theta)d\theta$
- Observation model: First draw θ from π, then draw X from $f(\cdot|\theta)$

Idea: Prior $\pi(\theta)$ reflects information about parameters before the experiment. Once data x is obtained, prior updated using Bayes formula to obtain *posterior*

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{m(x)}$$

Key: All inferences about θ are based on the posterior density

Bayesian Point Estimation

Additional Ingredients

- Loss function $\ell : \Theta \times \Theta \to \mathbb{R}$
- Estimator $h : \mathcal{X} \to \Theta$
- Risk $R(\alpha, h) = \mathbb{E}_\alpha \, \ell(\alpha, h(X))$
- Bayes risk $R_\pi(h) = \int_\Theta R(\alpha, h)\, \pi(\alpha)\, d\alpha$

Recall: The Bayes estimator $\hat{\theta}_\pi$ is the estimator h minimizing the Bayes risk $R_\pi(h)$

$$\hat{\theta}_\pi \;=\; \operatorname*{argmin}_h R_\pi(h)$$

Task: Determine form of $\hat{\theta}_\pi(x)$ based on the posterior $\pi(\theta|x)$ and the loss function ℓ

Bayesian Point Estimation

Perspective: Fix a loss function $\ell : \theta \times \theta \to \mathbb{R}$ and an estimator $h : \mathcal{X} \to \theta$

- Parameter θ is random with distribution π
- Pair (θ, X) is random with joint distribution $f(\alpha, x) = \pi(\alpha) f(x|\alpha)$
- Risk function $R(\alpha, h) = \mathbb{E}[\ell(\theta, h(X))|\theta = \alpha]$

Upshot: The Bayes risk $R_\pi(h) = \mathbb{E}\ell(\theta, h(X))$ where the expectation is over the joint distribution of (θ, X), and the Bayes rule can be written as

$$\hat{\theta}_\pi = \operatorname*{argmin}_h \mathbb{E}\ell(\theta, h(X))$$

By conditioning on the observed value of X we find that

$$\hat{\theta}_\pi(x) = \operatorname*{argmin}_\alpha \mathbb{E}[\ell(\theta, \alpha)|X = x]$$

Bayesian Estimators for Standard Loss Functions

Fact: Let $\mathcal{P} = \{f(x|\theta) : \theta \in \Theta\}$ with $\Theta \subseteq \mathbb{R}$ and prior $\pi(\theta)$ be given

1. Squared loss $\ell(\theta, \theta') = (\theta - \theta')^2$. Bayes rule $\hat{\theta}_\pi$ is the posterior mean $\mathbb{E}(\theta \,|\, X = x)$

$$\hat{\theta}_\pi(x) \;=\; \int_\Theta \theta \, \pi(\theta|x) \, d\theta$$

2. Absolute loss $\ell(\theta, \theta') = |\theta - \theta'|$. Bayes rule $\hat{\theta}_\pi$ is posterior median $\text{Med}(\theta|X = x)$

$$\hat{\theta}_\pi(x) \;=\; M \text{ such that } \int_{-\infty}^{M} \pi(\theta|x) \, d\theta = \frac{1}{2}$$

3. Zero-one loss $\ell(\theta, \theta') = \mathbb{I}(\theta \neq \theta')$. Bayes rule $\hat{\theta}_\pi$ is posterior mode $\text{Mode}(\theta|X = x)$

$$\hat{\theta}_\pi(x) \;=\; \operatorname*{argmax}_{\theta \in \Theta} \pi(\theta|x)$$

Conjugate Families of Priors

Given: Family $\mathcal{P} = \{f(x|\theta) : \theta \in \Theta\}$

Definition: A family $\Pi = \{\pi\}$ of prior distributions on Θ is said to be *conjugate* for $\mathcal{P}$ if

$$\pi \in \Pi \text{ and } f(\cdot|x) \in \mathcal{P} \text{ imply } \pi(\theta|x) \propto f(\cdot|x)\pi(\theta) \in \Pi$$

Examples of conjugate $(\Pi, \mathcal{P})$ pairs

- Beta, Binomial
- Dirichlet, Multinomial
- Gamma, Poisson
- Normal, Normal mean (known variance)
- Inverse Gamma, Normal variance (known mean)

Estimation of Binomial Probability with Beta Prior

Problem: Observe $X \sim \text{Bin}(n, \theta)$. Prior $\pi = \text{Beta}(\alpha, \beta)$ on $\Theta = (0, 1)$

- Given data x, posterior distribution is

$$\pi(\theta|x) = \text{Beta}(x + \alpha, n - x + \beta)$$

- Under squared loss, Bayes rule is posterior mean

$$\hat{\theta}_\pi(x) \;=\; \mathbb{E}(\theta|x) \;=\; \frac{x + \alpha}{n + \alpha + \beta}$$

- Interpretation: Simple calculation shows that

$$\text{posterior mean} \;=\; \lambda \cdot \text{sample mean} \;+ (1 - \lambda) \cdot \text{prior mean}$$

where $\lambda = n/(n + \alpha + \beta)$ tends to 1 as n increases

Bayesian Estimation: Normal Mean, Known Variance

Problem: Observe $X \sim \mathcal{N}(\theta, \sigma^2)$ with σ^2 known. Prior $\pi = \mathcal{N}(\alpha, \tau^2)$ on $\Theta = \mathbb{R}$

- Given data x posterior distribution is $\pi(\theta|x) = \mathcal{N}(a, b^2)$ where

$$a = \frac{\tau^2}{\tau^2 + \sigma^2} x + \frac{\sigma^2}{\tau^2 + \sigma^2} \alpha \qquad b^2 = \frac{\sigma^2 \tau^2}{\tau^2 + \sigma^2}$$

- Posterior mean, median, and mode of θ are all equal to a. So the Bayes rules under squared, absolute, and zero-one loss coincide and are given by

$$\hat{\theta}_\pi(x) = \frac{\tau^2}{\tau^2 + \sigma^2} x + \frac{\sigma^2}{\tau^2 + \sigma^2} \alpha$$

- If prior variance τ^2 is very large then $\pi(\theta|x) \approx \mathcal{N}(x, \sigma^2)$ and $\hat{\theta}_\pi(x) \approx x$

- If prior variance τ^2 is very small then $\pi(\theta|x) \approx \mathcal{N}(\alpha, 0)$ and $\hat{\theta}_\pi(x) \approx \alpha$

Rao-Blackwell Theorem

Conditioning on a Sufficient Statistic

Setting: Family $\mathcal{P} = \{f(\cdot|\theta) : \theta \in \Theta\}$ of distributions on $\mathcal{X}$

- Point estimator $\tilde{\theta} : \mathcal{X} \to \Theta$
- Sufficient statistic $T : \mathcal{X} \to \mathcal{Y}$ for $\mathcal{P}$

Idea: The sufficient statistic $T(x)$ captures information in x relevant to estimation of θ, so we might consider point estimates that are functions of $T(x)$.

Conditioning: From the given estimator $\tilde{\theta}$ we can obtain a new estimator $\hat{\theta}$ that depends on $T(x)$ by conditioning on T

$$\hat{\theta}(x) = \mathbb{E}_\theta \left[\tilde{\theta}(X) \,|\, T(X) = T(x) \right]$$

Questions: Is $\hat{\theta}$ a legitimate estimator? How does it compare with $\tilde{\theta}$?

Rao-Blackwell Theorem

Theorem: Let $\hat{\theta}(x) = \mathbb{E}_\theta\left[\tilde{\theta}(X)\,|\,T(X) = T(x)\right]$ be defined by conditioning on T

1. If T is sufficient then $\hat{\theta}(x)$ is independent of θ and is a legitimate estimator
2. The estimators $\hat{\theta}$ and $\tilde{\theta}$ have the same mean behavior, $\mathbb{E}_\theta\,\hat{\theta}(X) = \mathbb{E}_\theta\,\tilde{\theta}(X)$
3. If for each $\theta \in \Theta$ the loss $\ell(\theta, \cdot) : \Theta \to \mathbb{R}$ is convex then $R(\theta, \hat{\theta}) \leq R(\theta, \tilde{\theta})$ for all $\theta \in \Theta$. Thus $\hat{\theta}$ is at least as good as $\tilde{\theta}$

Corollary

1. If $\tilde{\theta}$ is unbiased for θ, then $\hat{\theta}$ is unbiased for θ
2. The theorem applies to the squared loss $\ell(\theta, \theta') = (\theta - \theta')^2$. In particular

$$\mathrm{Var}_\theta(\hat{\theta}) \leq \mathrm{Var}_\theta(\tilde{\theta})$$

Lehmann-Scheffe Theorem

Complete Statistics

Definition: A statistic T is *complete* for $\mathcal{P} = \{f(\cdot|\theta) : \theta \in \Theta\}$ if for any function h,

$$\mathbb{E}_\theta\, h(T(X)) = 0 \text{ for all } \theta \quad \text{implies} \quad \mathbb{P}_\theta\left(h(T(X)) = 0\right) = 1 \text{ for all } \theta$$

Note that reverse implication is always true

Example: If $X_1, \ldots, X_n$ are i.i.d. $\sim$ Unif$(0, \theta)$ then $T(x) = x_{(n)}$ is complete

Fact: If $\mathcal{P}$ is a nice exponential family with sufficient statistic T, then T is complete

Lehmann-Scheffe Theorem

Let $\mathcal{P} = \{f(\cdot|\theta) : \theta \in \Theta\}$ be a family of distributions on a sample space $\mathcal{X}$. Under mild conditions $\mathcal{P}$ has a minimal sufficient statistic

Fact: If T is complete and sufficient for $\mathcal{P}$ then it is minimal sufficient for $\mathcal{P}$

Theorem: Let T be complete and sufficient for $\mathcal{P}$. Let $\hat{\theta}$ be any estimator such that

- $\hat{\theta}(x) = g(T(x))$ ($\hat{\theta}$ is a function of T)
- $\mathbb{E}_\theta\, \hat{\theta}(X) = \tau(\theta)$ ($\hat{\theta}$ is an unbiased estimator of $\tau(\theta)$)

If $\tilde{\theta}$ is any estimator s.t. $\mathbb{E}_\theta\, \tilde{\theta}(X) = \tau(\theta)$ then $\mathrm{Var}_\theta(\hat{\theta}) \leq \mathrm{Var}_\theta(\tilde{\theta})$ for every θ

Upshot: An estimator that is a function of a complete statistic has minimum variance among all unbiased estimators with the same mean function

Applications of Minimum Variance Theorem

Ex 1. Let $X_1, \ldots, X_n$ be i.i.d. $\sim$ Unif$(0, \theta)$. Then

$$\hat{\theta}(x) \;=\; \left(\frac{n+1}{n}\right) x_{(n)}$$

is a minimum variance unbiased estimator of θ

Ex 2. Let $X_1, \ldots, X_n$ be i.i.d. $\sim$ Bern(θ). Then

$$\hat{\theta}(x) \;=\; \frac{T(x)(T(x)-1)}{n(n-1)},$$

with $T(x) = \sum_{i=1}^{n} x_i$, is a minimum variance unbiased estimator of θ^2