Section 2: Estimators

Ricky Truong (rickytruong@college.harvard.edu),
Emily Xing (exing@college.harvard.edu)

1 Introduction

1.1 Logistics

Welcome! The goal of our weekly section is to review content from lecture, with emphasis on big-picture intuition, mathematical proof, and hands-on practice.

Specifically, we aim for our sections to be interactive, well-paced, inclusive, and fun! Please do not hesitate to reach out if you have any questions/feedback!

1.2 Office Hours

- Mondays, 7:30 - 9:30 PM in Adams D-Hall (Ricky).
- Fridays, 11 AM - 12 PM in Maxwell-Dworkin 2nd Floor (Emily).
- Saturdays, 10:30 – 11:30 AM in Cabot D-Hall (Emily).

2 Big Picture

In statistics, we use estimators to learn about estimands. Though we could use any function of the data, there are two famous estimators we'll see recurring: the *maximum likelihood estimator* and the *method of moments estimator*.

Regardless of which estimator we use, we can evaluate its performance using *loss functions*. Specifically, we usually want estimators that are *consistent* and that have low *MSE*.

3 Mathematics

Idea. Mathematics is the *language* of probability and statistics. So that we're all on the same page, let's review some important concepts.

3.1 Functions

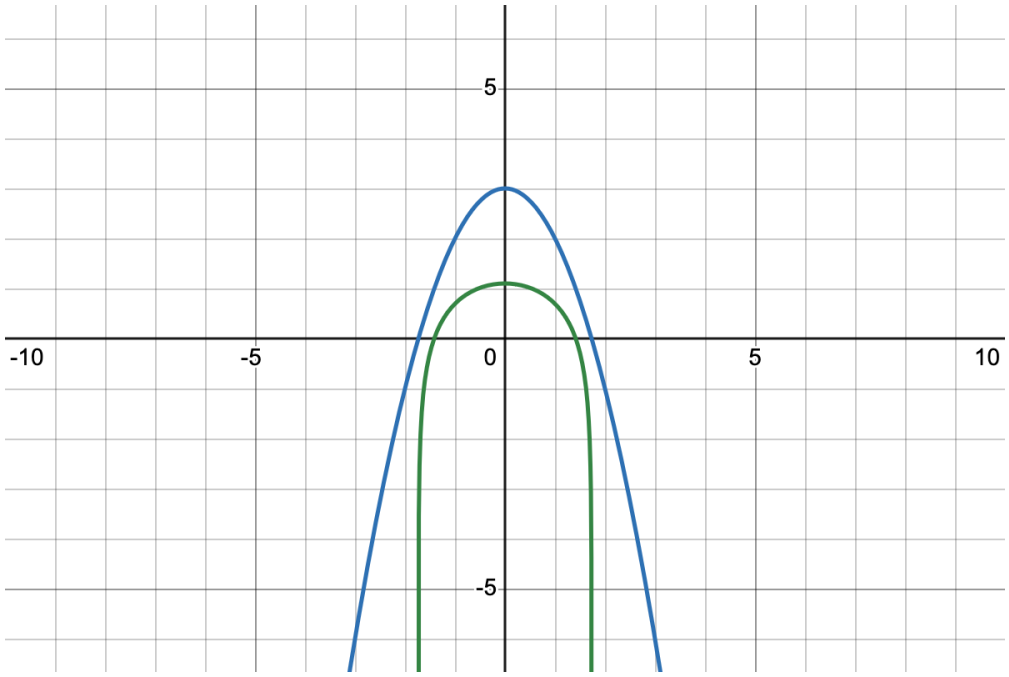

FIGURE 1: The graph for $f(x) = -x^2 + 3$ (in blue) and $\log(f(x)) = \log(-x^2 + 3)$ (in green).

Definition 1 (Maximum of a function)**.** The largest value a function attains within its range.

- In Figure 1, $\max f(x) = 3$.

Definition 2 (Supremum of a function)**.** The smallest value that bounds the function's range from above, which may or may not be in the range itself.[1]

- In Figure 1, $\sup f(x) = 3$.

Definition 3 (Arg max of a function)**.** The input value that achieves the maximum of a function. It is good notation to specify the set of possible input values.

- In Figure 1, $\arg\max_{x \in \mathbb{R}} f(x) = 0$.

3.2 Operations

Definition 4 (Logarithm)**.** $\log_b(a) = c \iff b^c = a$.

- $\log_b(1) = 0, \log_b(b) = 1, \log_b(0)$ is undefined.
- $\log(xy) = \log(x) + \log(y)$.
- $\log(\frac{x}{y}) = \log(x) - \log(y)$.
- $\log(x^k) = k \log(x)$.

Definition 5 (Derivative)**.** $\frac{d}{dx} f(x) = f'(x)$.

[1]For the course, you can think of max and sup as essentially interchangeable.

- $\frac{d}{dx}(ax+b) = a$ for constants a, b by linearity of differentiation.
- $\frac{d}{dx}(x^n) = nx^{n-1}$ by power rule.
- $\frac{d}{dx}(f(x)g(x)) = f'(x)g(x) + f(x)g'(x)$ by product rule.
- $\frac{d}{dx}f(g(x)) = f'(g(x))g'(x)$ by chain rule.
- $\frac{\partial}{\partial x}(axy + bx + cy + d) = ay + b$ for constants a, b, c, d by partial differentiation.

Definition 6 (Integral)**.** $\int_a^c f(x)dx = F(c) - F(a)$, where $F'(x) = f(x)$.

- $\int_a^c f(x)dx = \int_a^b f(x)dx + \int_b^c f(x)dx$ for constants a, b, c where $a \leq b \leq c$.
- $\int_a^c f(g(x))g'(x)dx = \int_{g(a)}^{g(c)} f(u)du$, where $u = g(x)$ and $du = g'(x)dx$ by u-substitution.

Concept Checker 1. Let $X \sim \text{Unif}(a, b)$ with $0 < a < b$. How can we simplify $P(X > a) = \int_a^\infty f(x)dx$?

Solution

4 Likelihood

Idea. Before we discuss famous estimators, we must define likelihood, which is the foundation for one of the most important estimators we'll be using in the course.

Definition 7 (Likelihood)**.** $\mathcal{L}(\theta; \vec{y}) = f_{\vec{Y}}(\vec{y}; \theta)$, where $f_{\vec{Y}}(\vec{y}; \theta)$ is the joint density of the data.[2] Likelihood is very analogous to probability, but they're not the same.[3]

- In $\mathcal{L}(\theta; \vec{y})$, $\vec{y}$ is viewed as fixed whereas in $f_{\vec{Y}}(\vec{y}; \theta)$, θ is viewed as fixed.
- The extra notation is in the joint density helpful but not necessary. E.g., if $X \sim \text{Expo}(\lambda)$, then the PDF is $f(x) = f_X(x; \lambda) = \lambda e^{-\lambda x}$ since it is a function of x and λ.
- The semicolon can look weird, but if you've taken multivariable calculus, you'll be familiar with functions of multiple variables (e.g., $f(x, y)$). In the same way, likelihood is a function of θ and $\vec{y}$, so we could write $\mathcal{L}(\theta, \vec{y})$, but the semicolon emphasizes the distinction.

[2] $\mathcal{L}(\theta; \vec{Y}) = f_{\vec{Y}}(\vec{Y}; \theta)$ is random whereas $\mathcal{L}(\theta; \vec{y}) = f_{\vec{Y}}(\vec{y}; \theta)$ is fixed.

[3] In the Frequentist paradigm, probability is a long-term frequency, so if the probability a coin lands heads is 0.50, then in the long run, we expect the coin to land heads 50% of the time. On the contrary, likelihood doesn't have this interpretation. Likelihood is really only meaningful relative to other values (i.e., the maximum likelihood).

• We drop multiplicative constants that are not functions of the parameter since they don't affect the argmax. E.g., for $Y \sim \text{Bin}(3, \theta)$, $\mathcal{L}(\theta; y) = \binom{3}{y}\theta^y(1-\theta)^{3-y}$ is equivalent to $\mathcal{L}(\theta; y) = \theta^y(1-\theta)^{1-y}$.

• For likelihood $\mathcal{L}(\theta; \vec{y})$ and reparameterization $\psi = g(\theta)$ (where g is a known function), $\mathcal{L}(\psi; \vec{y}) = \mathcal{L}(\theta; \vec{y})$ by invariance.

• For data $\vec{y} = (y_1, \ldots, y_n)$ from model with parameter θ and reparameterization $\vec{x} = h(\vec{y})$ (where $h : \mathbb{R}^n \to \mathbb{R}^n$ is a known function), $\mathcal{L}(\theta; \vec{x}) = \mathcal{L}(\theta; \vec{y})$ by invariance.

Definition 8 (Log-likelihood)**.** $\ell(\theta; \vec{y}) = \log(\mathcal{L}(\theta; \vec{y}))$.

• $f(x) = \log(x)$ is monotonically increasing (i.e., as the input increases, the output never decreases). In practice, working with log-likelihood is often easier.

• We drop additive constants that are not functions of the parameter since they don't affect the argmax. E.g., for $Y \sim \text{Bin}(3, \theta)$, $\ell(\theta; y) = \log\left(\binom{3}{y}\right) + \log\left(\theta^y\right) + \log\left((1-\theta)^{1-y}\right)$ is equivalent to $\ell(\theta; y) = \log\left(\theta^y\right) + \log\left((1-\theta)^{1-y}\right)$.

5 Famous Estimators

Idea. Thankfully, we don't have to reinvent the wheel. Many statisticians have already defined famous estimators we can use. In particular, we'll often use the *maximum likelihood estimator* or *method of moments estimator*. Many times, they'll actually be equivalent!

Definition 9 (Maximum likelihood estimator)**.** The estimator that maximizes the likelihood (i.e., the most likely value for θ, given the data): $\hat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta; \vec{Y})$.

• Equivalently, $\hat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} \ell(\theta; \vec{Y})$ since $f(x) = \log(x)$ is monotonically increasing. Notice in Figure 1 $\arg\max_{x \in \mathbb{R}} f(x) = \arg\max_{x \in \mathbb{R}} \log(f(x))$. In practice, working with log-likelihood is often easier.

• If $\hat{\theta}_{\text{MLE}}$ is the MLE for θ and g is a known function, then $g(\hat{\theta}_{\text{MLE}})$ is the MLE for $g(\theta)$ by invariance.

• **Strategy**: Solve for $\hat{\theta}$ from $\ell'(\theta; \vec{Y}) = 0$. Verify $\hat{\theta}$ is a local maximum by checking $\ell''(\hat{\theta}; \vec{Y}) < 0$. Then verify $\hat{\theta}$ is a global maximum (e.g., testing concavity or behavior at endpoints of Θ). Alternatively, if θ is a transformation of a distribution with a known MLE, apply invariance.

Concept Checker 2. Let $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \text{Bern}(\theta)$. Provide the justification at each step for

the derivation of the MLE.

$$\begin{aligned}
\mathcal{L}(\theta;\vec{Y}) &= f_{\vec{Y}}(\vec{Y};\theta) && \text{by } _____\\
&= \prod_{i=1}^{n} f_{Y_i}(Y_i;\theta) && \text{by } _____\\
&= \prod_{i=1}^{n} (\theta)^{Y_i}(1-\theta)^{1-Y_i} && \text{by } _____\\
&= (\theta)^{\sum_{i=1}^{n} Y_i}(1-\theta)^{n-\sum_{i=1}^{n} Y_i} && \text{by } _____\\
\ell(\theta;\vec{Y}) &= \log\left((\theta)^{\sum_{i=1}^{n} Y_i}(1-\theta)^{n-\sum_{i=1}^{n} Y_i}\right) && \text{by } _____\\
&= (\sum_{i=1}^{n} Y_i)\log(\theta) + (n - \sum_{i=1}^{n} Y_i)\log(1-\theta) && \text{by } _____\\
\ell'(\theta;\vec{Y}) &= \frac{\sum_{i=1}^{n} Y_i}{\theta} - \frac{n-\sum_{i=1}^{n} Y_i}{1-\theta} && \text{by } _____\\
0 &= \frac{\sum_{i=1}^{n} Y_i}{\theta} - \frac{n-\sum_{i=1}^{n} Y_i}{1-\theta} && \text{by } _____\\
\theta &= \frac{\sum_{i=1}^{n} Y_i}{n} && \text{by } _____\\
\ell''(\theta;\vec{Y}) &= -\frac{\sum_{i=1}^{n} Y_i}{\theta^2} - \frac{n-\sum_{i=1}^{n} Y_i}{(1-\theta)^2} < 0 \;\forall\; \theta \in \Theta && \text{by } _____\\
\hat{\theta}_{\text{MLE}} &= \bar{Y} && \text{by } _____
\end{aligned}$$

Solution

Definition 10 (Method of moments estimator)**.** The estimator where the theoretical moments are replaced with sample moments.

• For i.i.d. data $Y_1, \ldots, Y_n$, $\mu'_k = \mathbb{E}[Y_i^k]$ is the theoretical kth moment whereas $M_k = \frac{1}{n}\sum_{i=1}^n Y_i^k$ is the sample kth moment.[4]

• **Strategy**: Write the parameter in terms of theoretical moments and replace them with their sample analogues.

Concept Checker 3. What is the sample moment analogue for $\mathbb{E}[Y_1^2]$?

Solution

Concept Checker 4. Let $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \text{Expo}(\lambda)$. Provide the justification at each step for

[4]Sometimes, the kth sample moment can be denoted as $\overline{Y^k}$. E.g., $\overline{Y^2} = \frac{1}{n}\sum_{i=1}^n Y_i^2$. Notice this is different from $(\bar{Y})^k$. E.g., $(\bar{Y})^2 = (\frac{1}{n}\sum_{i=1}^n Y_i)^2$.

the derivation of the MOM estimator.

$$\mathbb{E}[Y_1] = \frac{1}{\lambda} \quad \text{by } \_\_\_\_\_$$
$$\lambda = \frac{1}{\mathbb{E}[Y_1]} \quad \text{by } \_\_\_\_\_$$
$$\hat{\lambda}_{\text{MOM}} = \frac{1}{\overline{Y}} \quad \text{by } \_\_\_\_\_$$

Solution

6 Evaluation

Idea. There's nothing stopping us from using any estimator we want. Ricky's favorite number is 3, so let $\hat{\theta}_{\text{Ricky}} = 3$. Of course, this would be terrible for most cases, so how do we evaluate the quality of estimators? There is no one standard that makes an estimator "good," but often, we look for estimators that are *consistent* and that have low *MSE*.

6.1 Metrics

Definition 11 (Standard error)**.** $\text{SE}[\hat{\theta}] = \text{SD}[\hat{\theta}] = \sqrt{\text{Var}[\hat{\theta}]}$.

Definition 12 (Bias)**.** $\text{Bias}[\hat{\theta}] = \mathbb{E}[\hat{\theta} - \theta]$.[5]

• We say $\hat{\theta}$ is unbiased for θ if $\text{Bias}[\hat{\theta}] = 0$ (i.e., $\mathbb{E}[\hat{\theta}] = \theta$).

6.2 Loss Functions

Definition 13 (Loss)**.** A loss function $L(\theta, \hat{\theta})$ is a function of the estimand θ and the estimator $\hat{\theta}$ that satisfies two properties:

1. $L(\theta, \hat{\theta}) \geq 0$
2. $L(\theta, \hat{\theta}) = 0$ if and only if $\hat{\theta} = \theta$.

Definition 14 (Risk)**.** A risk function $\text{Risk}(\theta, \hat{\theta}) = \mathbb{E}[L(\theta, \hat{\theta})]$ is the expected loss.

Definition 15 (Mean absolute error)**.** $\text{MAE}[\hat{\theta}] = \mathbb{E}[|\hat{\theta} - \theta|]$ is the risk function for the absolute error loss $L(\theta, \hat{\theta}) = |\hat{\theta} - \theta|$.

[5] You may see bias denoted as $\text{Bias}[\hat{\theta}, \theta]$ or $\text{Bias}_\theta[\hat{\theta}]$ to make it explicit this is a function of the estimand as well.

Definition 16 (Mean square error)**.** $\text{MSE}[\hat{\theta}] = \mathbb{E}[(\hat{\theta}-\theta)^2]$ is the risk function for the squared error loss $L(\theta, \hat{\theta}) = (\hat{\theta} - \theta)^2$.

• Equivalently, $\text{MSE}[\hat{\theta}] = \text{Var}[\hat{\theta}] + (\text{Bias}[\hat{\theta}])^2$, which demonstrates bias-variance tradeoff.

Proof. Let $V = \hat{\theta} - \theta$.

$$\begin{aligned}
\text{MSE}[\hat{\theta}] &= \mathbb{E}[(\hat{\theta} - \theta)^2] && \text{by definition of MSE} \\
&= \mathbb{E}[V^2] && \text{by substituting} \\
&= \text{Var}[V] + (\mathbb{E}[V])^2 && \text{by definition of variance} \\
&= \text{Var}[\hat{\theta} - \theta] + (\mathbb{E}[\hat{\theta} - \theta])^2 && \text{by substituting} \\
&= \text{Var}[\hat{\theta}] + (\mathbb{E}[\hat{\theta} - \theta])^2 && \text{by bilinearity} \\
&= \text{Var}[\hat{\theta}] + (\text{Bias}[\hat{\theta}])^2 && \text{by definition of bias}
\end{aligned}$$

□

Definition 17 (Consistency)**.** An estimator $\hat{\theta}$ is consistent for the estimand θ if $\hat{\theta} \xrightarrow{p} \theta$.

• Equivalently, $\hat{\theta}$ is consistent if $\lim_{n\to\infty} \text{MSE}[\hat{\theta}] = 0$.

Proof. Suppose $\lim_{n\to\infty} \text{MSE}[\hat{\theta}] = 0$. We want to show this implies $\hat{\theta} \xrightarrow{p} \theta$, which would require $\forall\ \varepsilon > 0$, $\lim_{n\to\infty} P(|\hat{\theta} - \theta| > \varepsilon) = 0$.

First, $\lim_{n\to\infty} \text{MSE}[\hat{\theta}] = 0 \implies \lim_{n\to\infty} \mathbb{E}[(\hat{\theta} - \theta)^2] = 0$ by definition of MSE.

Let $\varepsilon > 0$ be any positive number. Now, $P(|\hat{\theta} - \theta| \geq \varepsilon) = P((\hat{\theta} - \theta)^2 \geq \varepsilon^2) \leq \frac{\mathbb{E}[(\hat{\theta}-\theta)^2]}{\varepsilon^2}$ by Markov's Inequality. Additionally, since the event $|\hat{\theta} - \theta| \geq \varepsilon$ implies the event $|\hat{\theta} - \theta| > \varepsilon$, $P(|\hat{\theta} - \theta| > \varepsilon) \leq P(|\hat{\theta} - \theta| \geq \varepsilon)$. We can rewrite all this as $P(|\hat{\theta} - \theta| > \varepsilon) \leq \frac{\mathbb{E}[(\hat{\theta}-\theta)^2]}{\varepsilon^2}$.

Recall $\lim_{n\to\infty} \frac{\mathbb{E}[(\hat{\theta}-\theta)^2]}{\varepsilon^2} = 0$, and $\lim_{n\to\infty} P(|\hat{\theta} - \theta| > \varepsilon) \leq \lim_{n\to\infty} \frac{\mathbb{E}[(\hat{\theta}-\theta)^2]}{\varepsilon^2}$ by applying the limit to both sides, so $\lim_{n\to\infty} P(|\hat{\theta} - \theta| > \varepsilon) \leq 0$ by substituting. Since probability cannot be negative, $\lim_{n\to\infty} P(|\hat{\theta} - \theta| > \varepsilon) = 0$.

Notice ε was arbitrary. We conclude $\forall\ \varepsilon > 0$, $\lim_{n\to\infty} P(|\hat{\theta} - \theta| > \varepsilon) = 0$, so $\hat{\theta} \xrightarrow{p} \theta$. □

Concept Checker 5. For any unbiased estimator $\hat{\theta}_{\text{UB}}$, what is $\text{MSE}[\hat{\theta}_{\text{UB}}]$ always equal to?

Solution

7 Practice Problems

Problem 1. Let $Y_1, \ldots, Y_n$ be an i.i.d. random sample from a population with PDF $f(y) = \theta y^{\theta-1}$, where $0 < y < 1$ and $\theta > 0$.[6]

(a) Find the likelihood function: $\mathcal{L}(\theta; \vec{Y})$.

[6]Inspired by Problem 1 in "Section 1: An Introduction to Statistics" by Elvin Lo and Akshay Kumar.

(b) Find the log-likelihood function: $\ell(\theta; \vec{Y})$.

(c) Find the maximum likelihood estimator: $\hat{\theta}_{\text{MLE}}$.

(d) Let $\psi = \frac{1}{\theta}$. Find the maximum likelihood estimator: $\hat{\psi}_{\text{MLE}}$.

(e) **Challenge**: Find the method of moments estimator: $\hat{\theta}_{\text{MOM}}$.

Hint: Though this isn't one of our "famous" distributions, we are given enough information through the PDF and support.

Solution

Problem 2. Let $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \text{Gamma}(a, \lambda)$, with a known and λ unknown.

(a) Find the method of moments estimator: $\hat{\lambda}_{\text{MOM}}$.

(b) **Challenge**: Now suppose both a and λ are unknown. Find the method of moments estimators: $\hat{a}_{\text{MOM}}, \hat{\lambda}_{\text{MOM}}$.

Hint: There are two unknown parameters, so write out the first two moments.

Solution

Problem 3. Let $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \text{Bern}(p)$. We already showed $\hat{p}_{\text{MLE}} = \bar{Y}$.

(a) Find the MSE of $\hat{p}_{\text{MLE}}$.

(b) Is $\hat{p}_{\text{MLE}}$ a consistent estimator for p?

Solution