EN.553.431 Honors Mathematical Statistics

Daniel Yao

Fall 2024

Contents

0.0 Introduction.

Introduction.

0.1 Remark.

The following notes follow the material presented in EN.553.431 Honors Mathematical Statistics taught by Professor Avanti Athreya during the semester of Fall 2024 at The Johns Hopkins University. The content of lectures is presented along with selected homework exercises.

0.2 Notation.

Rice shall refer to 'Mathematical Statistics and Data Analysis' 3rd edition (US) by John A. Rice.

0.3 Notation.

The following abbreviations shall be overserved.

1. The term 'rv' shall denote 'random variable'.
2. The term 'pmf' shall denote 'probability mass function'.
3. The term 'pdf' shall denote 'probability density function'.
4. The term 'cdf' shall denote 'cumulative density function'.

0.4 Notation

A finite population sample shall be as follows. We are given a finite bivariate population of N distinct objects, and associated to each object k is a pair of measurements (x_k, y_k). Suppose our population of measurements is represented by $\{(x_1, y_1), \ldots, (x_N, y_N)\}$. We assume $N > 1$. Let τ_x and τ_y be the population totals of the x- and y-measurements, respectively; let μ_x and μ_y be the population means of the x- and y-measurements, respectively; let σ_x^2 and σ_y^2 denote the population variances of the x- and y-measurements, respectively. Let σ_{xy} denote the population covariance. The population 3rd and 4th moments, $\mu_3(x)$ and $\mu_4(x)$, respectively, of the x-values are

$$\mu_3(x) = \frac{1}{N}\sum_{k=1}^{N} x_k^3, \quad \mu_4(x) = \frac{1}{N}\sum_{k=1}^{N} x_k^4$$

Similarly, the population 3rd and 4th moments are, respectively, $\mu_3(y)$ and $\mu_4(y)$. Let $\sigma_{x^2y^2}$ denote

$$\sigma_{x^2y^2} = \left(\frac{1}{N}\sum_{k=1}^{N} x_k^2 y_k^2\right) - (\sigma_x^2 + \mu_x^2)(\sigma_y^2 + \mu_y^2)$$

Let M_x and M_y represent the population maximum of the x- and y-values, respectively, so that M_x and M_y are defined by

$$M_x = \max\{x_k : 1 \leq k \leq N\}, \quad M_y = \max\{y_k : 1 \leq k \leq N\}$$

Let m_x and m_y denote the population minimum of the x- and y-measurements, respectively, so that

$$m_x = \min\{x_k : 1 \leq k \leq N\}, \quad m_y = \min\{y_k : 1 \leq k \leq N\}$$

All sample sizes n satisfy $n \geq 1$, and in some cases, we specify if $n > 1$ or we give an explicit value for n. In what follows below, $\bar{X}$ denotes the sample mean of the x-measurements in the sample, and $\bar{Y}$ denotes the sample mean of the y-measurements in the sample. The letter E represents expected value; Var represents variance; and Cov represents covariance.

1.0 Introduction.

Finite Population Samples.

1.1 Definition.

For an estimator $\hat{\theta}$ of a parameter θ, the mean squared error is

$$\begin{aligned} \mathrm{MSE}(\hat{\theta}) &= \mathrm{E}\left((\hat{\theta}-\theta)^2\right) \\ &= \mathrm{Var}(\hat{\theta}) + \left(\mathrm{E}(\hat{\theta})-\theta\right)^2 \\ &= \mathrm{Var}(\hat{\theta}) + \mathrm{Bias}(\hat{\theta})^2. \end{aligned}$$

where

$$\mathrm{Bias}(\hat{\theta}) = \mathrm{E}(\hat{\theta}) - \theta.$$

2.0 Introduction.

Confidence Intervals.

2.1 Theorem.

(Central Limit Theorem.) Let $U_1, U_2, \ldots, U_n$ be iid rvs with $\mathrm{E}(U_i) = \mu$ and $\mathrm{Var}(U_i) = \sigma^2$. For

$$\bar{U} = \frac{1}{n}\sum_{i=1}^{n} U_i,$$

we have that for all $t \in \mathbb{R}$,

$$P\left(\left|\frac{\bar{U} - \mu}{\sigma/\sqrt{n}}\right| \leq t\right) \to \Phi(t) \text{ as } n \to \infty.$$

2.2 Definition.

An α-critical value z_α for an rv Z is such that

$$P(Z > z_\alpha) = \alpha.$$

That is, α is the upper-tail probability of Z.

2.3 Example.

For $\bar{X}$ approximately normal, we have that

$$P\left(-z_\alpha \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_\alpha\right) \approx 1 - 2\alpha,$$

so

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \approx 1 - \alpha$$

where the (random) interval is the $1 - \alpha$ confidence inverval for μ.

2.4 Note.

The population standard deviation σ may be unknown, but we may substite the sample standard deviation s in its place.

2.5 Example.

By Chebyshev's Inequality, we have that for a sample of size n,

$$P(|s_n^2 - \sigma^2| \geq \delta) \leq \frac{\mathrm{Var}(s_n^2)}{\delta^2} \to 0 \text{ as } n \to \infty,$$

so

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \approx \mathrm{N}(0, 1)$$

for n large.

2.6 Example.

If we sample without replacement and $n << N$, then

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}} \approx \mathrm{N}(0, 1).$$

2.7 Theorem.

For the sample total

$$T_n = \sum_{i=1}^{n} X_i,$$

the CLT says that

$$P\left(\frac{T_n - n\mu}{\sigma\sqrt{n}} \leq t\right) \to \Phi(t) \text{ as } n \to \infty.$$

2.8 Note.

The sample mean and the sample total are related in that

$$\frac{n}{n}\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{T_n - n\mu}{\sigma\sqrt{n}}.$$

3.0 Introduction.

Taylor Approximations.

3.1 Theorem.

(Mean Value Theorem.) Suppose that $g : \mathbb{R} \to \mathbb{R}$ is differentiable on (a, b) and that $a < x < y < b$. Then there exists a $\xi \in (x, y)$ such that

$$g'(\xi) = \frac{g(y) - g(x)}{y - x}.$$

3.2 Theorem.

(Taylor's Theorem with Remainder.) Let $g : \mathbb{R} \to \mathbb{R}$ be n times differentiable on (a, b) and let $a < x < y < b$. Then there exists a $\xi \in (x, y)$ such that

$$g(y) = \sum_{k=0}^{n} \frac{g^{(k)}(x)}{k!}(y - x)^k + R_n(y)$$

where

$$R_n(y) = \frac{g^{(n+1)}(\xi)}{(n+1)!}(y - x)^{n+1}.$$

We may bound this remainder by

$$|R_n(y)| \leq \frac{M}{(n+1)!}|y - x|^{n+1}$$

where $M = \max |g^{(n+1)}(\xi)|$ on the interval (x, y).

3.3 Theorem.

For $g : \mathbb{R}^n \to \mathbb{R}$ twice-differentiable on a closed ball B containing x and y, we have that the first-order Taylor polynomial with remainder is

$$g(y) = g(x) + \nabla g(x)^T (y - x) + \frac{1}{2}(y - x)^T H(\xi)(y - x)$$

where

$$\nabla g(x) = \begin{pmatrix} \frac{\partial g}{\partial x_1}(\xi) & \cdots & \frac{\partial g}{\partial x_n}(\xi) \end{pmatrix}^T$$

is the gradient of g at x and

$$H(\xi) = \begin{pmatrix} \frac{\partial^2 g}{\partial x_1^2}(\xi) & \cdots & \frac{\partial^2 g}{\partial x_1 \partial x_n}(\xi) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 g}{\partial x_n \partial x_1}(\xi) & \cdots & \frac{\partial^2 g}{\partial x_n^2}(\xi) \end{pmatrix}$$

is the Hessian of g at ξ.

3.4 Example.

For $g : \mathbb{R}^2 \to \mathbb{R}$, the second order taylor polynomial is

$$g(y) \approx g(x) + \nabla g(x)^T (y - x) + \frac{1}{2}(y - x)^T H(x)(y - x).$$

Written in polynomial form, this is

$$\begin{aligned} g(y_1, y_2) \approx g(x_1, x_2) &+ \frac{\partial g}{\partial x_1}(y_1 - x_1) + \frac{\partial g}{\partial x_2}(y_2 - x_2) \\ &+ \frac{1}{2}\left(\frac{\partial^2 g}{\partial x_1^2}(y_1 - x_1)^2 + 2\frac{\partial^2 g}{\partial x_1 \partial x_2}(y_1 - x_1)(y_2 - x_2) + \frac{\partial^2 g}{\partial x_2^2}(y_2 - x_2)^2 \right) \end{aligned}$$

where the partial derivatives are evaluated at (x_1, x_2).

3.5 Example.

For $h : \mathbb{R}^2 \to \mathbb{R}$, the second-order taylor polynomial about (μ_x, μ_y) is

$$h(x,y) \approx h(\mu_x,\mu_y) + \frac{\partial h}{\partial x}(x-\mu_x) + \frac{\partial h}{\partial y}(y-\mu_y) + \frac{1}{2}\left(\frac{\partial^2 h}{\partial x^2}(x-\mu_x)^2 + 2\frac{\partial^2 h}{\partial x \partial y}(x-\mu_x)(y-\mu_y) + \frac{\partial^2 h}{\partial y^2}(y-\mu_y)^2\right)$$

where the partial derivatives are evaluated at (μ_x, μ_y). Therefore, the expected value of $h(X,Y)$ (under certain conditions) is

$$\mathrm{E}(h(X,Y)) \approx h(\mu_x,\mu_y) + \frac{1}{2}\frac{\partial^2 h}{\partial x^2}\sigma_X^2 + \frac{\partial^2 h}{\partial x \partial y}\sigma_{XY} + \frac{1}{2}\frac{\partial^2 h}{\partial y^2}\sigma_Y^2$$

because the first-order terms vanish when

$$\mathrm{E}(X-\mu_x) = 0.$$

3.6 Example.

For $h : \mathbb{R}^2 \to \mathbb{R}$, the first-order taylor polynomial about (μ_X, μ_Y) is

$$h(X,Y) \approx h(\mu_X,\mu_Y) + \frac{\partial h}{\partial x}(X-\mu_X) + \frac{\partial h}{\partial y}(Y-\mu_Y)$$

where the partial derivatives are evaluated at (μ_X, μ_Y). Therefore, the variance of $h(X,Y)$ (under certain conditions) is

$$\mathrm{Var}(h(X,Y)) \approx \left(\frac{\partial h}{\partial x}\right)^2 \mathrm{Var}(X) + \left(\frac{\partial h}{\partial y}\right)^2 \mathrm{Var}(Y) + 2\frac{\partial h}{\partial x}\frac{\partial h}{\partial y}\mathrm{Cov}(X,Y).$$

To approximate the variance, the second-order terms become small quickly, so a first-order approximation is appropriate.

4.0 Introduction.

Sample Ratio.

4.1 Definition.

For a bivariate population and a sample of size n, the sample ratio is

$$\bar{R} = \frac{\bar{Y}}{\bar{X}}.$$

4.2 Example.

We then have that

$$g(x, y) = \frac{y}{x}$$

with partial derivatives

$$\frac{\partial g}{\partial x} = -\frac{y}{x^2}, \quad \frac{\partial g}{\partial y} = \frac{1}{x}$$
$$\frac{\partial^2 g}{\partial x^2} = \frac{2y}{x^3}, \quad \frac{\partial^2 g}{\partial x \partial y} = -\frac{1}{x^2}, \quad \frac{\partial^2 g}{\partial y^2} = 0.$$

Therefore,

$$\begin{aligned}\bar{R} \approx \frac{\mu_y}{\mu_x} - \frac{\mu_y}{\mu_x^2}(\bar{X} - \mu_x) + \frac{1}{\mu_x}(\bar{Y} - \mu_y) \\ + \frac{\mu_y}{\mu_x^3}(\bar{X} - \mu_x)^2 - \frac{1}{\mu_x^2}(\bar{X} - \mu_x)(\bar{Y} - \mu_y).\end{aligned}$$

4.3 Example.

(Continued.) The expected value of $\bar{R}$ is

$$\begin{aligned}\mathrm{E}(\bar{R}) &\approx \frac{\mu_y}{\mu_x} + \frac{1}{\mu_x^2}\left(\mathrm{Var}(\bar{X})\frac{\mu_y}{\mu_x} - \mathrm{Cov}(\bar{X}, \bar{Y})\right) \\ &= r + \frac{1}{\mu_x^2}(\mathrm{Var}(\bar{X})r - \mathrm{Cov}(\bar{X}, \bar{Y}))\end{aligned}$$

where

$$r = \frac{\mu_y}{\mu_x}$$

In the case of sampling with replacement, we have that

$$\mathrm{E}(\bar{R}) = \frac{\mu_y}{\mu_x} + \frac{1}{\mu_x^2}\left(\frac{\mu_y}{\mu_x}\frac{\sigma_x^2}{n} - \frac{\sigma_{xy}}{n}\right).$$

In the case of sampling without replacement, we have that

$$\mathrm{E}(\bar{R}) = \frac{\mu_y}{\mu_x} + \frac{1}{\mu_x^2}\left(\frac{\mu_y}{\mu_x}\frac{\sigma_x^2}{n}\frac{N-n}{N-1} - \frac{\sigma_{xy}}{n}\frac{N-n}{N-1}\right).$$

4.4 Example.

(Continued.) The variance of $\bar{R}$ is

$$\mathrm{Var}(\bar{R}) = \frac{\mu_y^2}{\mu_x^4}\sigma_{\bar{x}}^2 + \frac{1}{\mu_x^2}\sigma_{\bar{y}}^2 - \frac{2\mu_y}{\mu_x^3}\sigma_{\bar{x}\bar{y}}.$$

In the case of sampling with replacement, we have that

$$\begin{aligned}\mathrm{Var}(\bar{R}) &= \frac{\mu_y^2}{\mu_x^4}\frac{\sigma_x^2}{n} + \frac{1}{\mu_x^2}\frac{\sigma_y^2}{n} - \frac{2\mu_y}{\mu_x^3}\frac{\sigma_{xy}}{n} \\ &= \frac{1}{\mu_x^2}\frac{1}{n}\left(\frac{\mu_y^2}{\mu_x^2}\sigma_x^2 + \sigma_y^2 - \frac{2\mu_y}{\mu_x}\sigma_{xy}\right).\end{aligned}$$

and in the case of sampling without replacement, we have that

$$\begin{aligned}\mathrm{Var}(\bar{R}) &= \frac{\mu_y^2}{\mu_x^4}\frac{\sigma_x^2}{n}\frac{N-n}{N-1} + \frac{1}{\mu_x^2}\frac{\sigma_y^2}{n}\frac{N-n}{N-1} - \frac{2\mu_y}{\mu_x^3}\frac{\sigma_{xy}}{n}\frac{N-n}{N-1} \\ &= \frac{1}{\mu_x^2}\frac{1}{n}\frac{N-n}{N-1}\left(\frac{\mu_y^2}{\mu_x^2}\sigma_x^2 + \sigma_y^2 - \frac{2\mu_y}{\mu_x}\sigma_{xy}\right).\end{aligned}$$

where n, N are the sample size and population size, respectively.

4.5 Theorem.

We have the following propositions from Rice, Chapter (7), Section (7.4). Consider the case of sampling without replacement. Taking

$$r = \frac{\mu_x}{\mu_y},$$

we can recover Theorem B, that the approximate expectation of $R = \bar{Y}/\bar{X}$ is

$$\mathrm{E}(R) \approx r + \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{\mu_x^2}(r\sigma_x^2 - \rho\sigma_x\sigma_y)$$

where ρ is the correlation

$$\rho = \frac{\sigma_{xy}}{\sigma_x\sigma_y}$$

as well as Corollary B, that the approximate bias of the ratio estimate $\bar{Y}_R = \mu_x R$ of μ_y is

$$\mathrm{E}(\bar{Y}_R) - \mu_y \approx \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{\mu_x}\sigma_x^2.$$

4.6 Theorem.

We have the following propositions from Rice, Chapter (7), Section (7.4). Consider the case of sampling without replacement. Taking

$$r = \frac{\mu_x}{\mu_y},$$

we can recover Theorem A, that the approximate variance of $R = \bar{Y}/\bar{X}$ is

$$\mathrm{Var}(R) \approx \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{\mu_x^2}(r^2\sigma_x^2 + \sigma_y^2 - 2r\sigma_{xy}),$$

and Corollary A, that the the estimated variance of the ratio estimate $\bar{Y}_R = \mu_x R$ of μ_y is

$$\mathrm{Var}(\bar{Y}_R) \approx \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\left(r^2\sigma_y^2 + \sigma_x^2 - 2r\sigma_{xy}\right),$$

and Corollary C, that the variance of $\bar{Y}_R$ can be estimated by

$$s_{\bar{Y}_R}^2 \approx \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)(R^2 s_x^2 + s_y^2 - 2Rs_{xy}).$$

4.7 Theorem.

(Rice, Chapter (7), Exercise (50).) Hartley and Ross (1954) derived the following exact bound on the relative size of the bias and standard error of a ratio estimate:

$$\frac{|\mathrm{E}(R) - r|}{\sigma_R} \leq \frac{\sigma_{\bar{X}}}{\mu_x}.$$

4.8 Proof.

(Continued.) Consider the relation

$$\mathrm{Cov}(R, \bar{X}) = \mathrm{E}(R\bar{X}) - \mathrm{E}(R)\mathrm{E}(\bar{X}).$$

We have that

$$\begin{aligned} E(R\bar{X}) - E(R)E(\bar{X}) &= E(\bar{Y}) - E(R)E(\bar{X}) \\ &= \mu_y - \mu_x E(R) \end{aligned}$$

so by the Cauchy-Schwarz inequality, we have that

$$\begin{aligned} |\mu_y - \mu_x E(R)| &\leq \sigma_{\bar{x}}\sigma_R \\ |\mu_x E(R) - \mu_y| &\leq \sigma_{\bar{x}}\sigma_R \\ |E(R) - r| &\leq \frac{\sigma_{\bar{x}}\sigma_R}{\mu_x} \\ \frac{|E(R) - r|}{\sigma_R} &\leq \frac{\sigma_{\bar{x}}}{\mu_x}. \end{aligned}$$

Remark. For the ratio estimate of the population total

$$T_R = \tau_x R,$$

the squared standard error for T_R is

$$s^2_{T_R} = N^2 \frac{1}{n} \left(\frac{N-n}{N-1} \right) (R^2 s_x^2 + s_y^2 - 2R s_{xy}).$$

Compare that to the standard error for the direct estimate T in part (c), which is

$$s^2_T = N^2 \frac{1}{n} \left(\frac{N-n}{N-1} \right) s_y^2.$$

If R is small or if s_x is small, then

$$\begin{aligned} R^2 s_x^2 + s_y^2 - 2R s_{xy} &< s_y^2 \\ s^2_{T_R} &< s^2_T. \end{aligned}$$

The same argument holds for the variance of the ratio estimate

$$\bar{Y}_R = \mu_x R.$$

This is an example of a biased estimator possessing a smaller variance than the unbiased estimator.

5.0 Introduction.

Real Analysis.

5.1 Definition.

A field is a set F equipped with two operations: addition and multiplication. The field axioms are as follows.

(A) Addition.

(A1) If $x, y \in F$, then $x + y \in F$. (Closure.)

(A2) If $x, y \in F$, then $x + y = y + x$. (Commutativity.)

(A3) If $x, y, z \in F$, then $(x + y) + z = x + (y + z)$. (Associativity.)

(A4) There exists an element $0 \in F$ such that $0 + x = x$ for all $x \in F$. (Identity.)

(A5) To every $x \in F$ there corresponds an element $-x \in F$ such that $x + (-x) = 0$. (Inverse.)

(M) Multiplication.

(M1) If $x, y \in F$, then $xy \in F$. (Closure.)

(M2) If $x, y \in F$, then $xy = yx$. (Commutativity.)

(M3) If $x, y, z \in F$, then $(xy)z = x(yz)$. (Associativity.)

(M4) There exists an element $1 \in F, 1 \neq 0$, such that $1x = x$ for all $x \in F$. (Identity.)

(M5) If $x \in F, \neq 0$, then there corresponds an element $1/x \in F$ such that $x(1/x) = 1$. (Inverse.)

(D) Distribution.

(D1) If $x, y, z \in F$, then $x(y + z) = xy + xz$. (Left distribution.)

5.2 Definition.

An ordered set is a set S equipped with a relation $<$ such that for all $x, y, z \in S$,

(1) If $x, y \in S$, then one and only one of

$$x < y, \quad x = y, \quad y < x$$

is true. (Trichotomy.)

(2) If $x, y, z \in S$ and $x < y$ and $y < z$, then $x < z$. (Transitivity.)

5.3 Definition.

An ordered field is a field F equipped with an order relation $<$ such that for all $x, y, z \in F$,

(1) If $x, y, z \in F$ and $y < z$, then $x + y < x + z$.

(2) If $x, y \in F$ and $x, y > 0$, then $xy > 0$.'

5.4 Remark.

From these axioms, we may derive the familiar properties of $\mathbb{Q}, \mathbb{R}, \mathbb{C}$.

5.5 Definition.

A subset D of an ordered field F is said to be bounded above if there exists an element $M \in F$ such that

$$x \leq M, \quad \forall x \in D.$$

The element M is called an upper bound of D. M is a least upper bound of D if

(1) $\forall x \in D, M \leq x$.

(2) $\forall m < M, \exists x \in D$ s.t. $m < x$.

5.6 Definition.

The least upper bound property states that every nonempty subset D of F that is bounded above has a least upper bound

$$\sup D.$$

5.7 Theorem.

There exists an ordered field $\mathbb{R}$ with the least upper bound property. Moreover, $\mathbb{Q} \subset \mathbb{R}$.

5.8 Definition.

A metric space is a set X equipped with a metric $d : X \times X \to \mathbb{R}$ such that for all $x, y, z \in X$,

(1) $d(x, y) \geq 0$. (Non-negativity.)

(2) $d(x, y) = 0$ if and only if $x = y$. (Positive definiteness.)

(3) $d(x, y) = d(y, x)$. (Symmetry.)

(4) $d(x, y) \leq d(x, z) + d(z, y)$. (Triangle inequality.)

Unless otherwise specified, we assume that the standard metric is

$$d(x, y) = |x - y|.$$

5.9 Definition.

A sequence $\{x_n\}$ in $\mathbb{R}$ is the indexed output of a map

$$\phi : \mathbb{N} \to \mathbb{R}$$

and we denote the sequence as

$$\{a_n : n \in \mathbb{N}\}.$$

or simply as $\{a_n\}$ or even more simply as a_n.

5.10 Definition.

A sequence is Cauchy if

$$\forall \epsilon > 0, \exists N_\epsilon \in \mathbb{N} \text{ s.t. } d(a_n, a_m) < \epsilon, \forall n, m \geq N_\epsilon.$$

5.11 Definition.

A sequence $\{a_n\}$ in $\mathbb{R}$ is convergent if

$$\exists L \in \mathbb{R} \text{ s.t. } \forall \epsilon > 0, \exists N_\epsilon \in \mathbb{N} \text{ s.t. } d(a_n - L) < \epsilon, \forall n \geq N_\epsilon.$$

L is said to be the limit of the sequence $\{a_n\}$.

5.12 Lemma.

A sequence is Cauchy if it is convergent.

5.13 Proof.

(Continued.) Suppose that $\epsilon > 0$. Since

$$a_n \to L,$$

there exists an $N_{\epsilon/2} \in \mathbb{N}$ such that for all $n > N_{\epsilon/2}$,

$$\begin{aligned} d(a_n, L) &< \frac{\epsilon}{2} \\ d(a_m, L) &< \frac{\epsilon}{2} \\ d(a_n, a_m) &< \epsilon \end{aligned}$$

for all $m, n > N_{\epsilon/2}$ by the triangle inequality. Hence, a sequence if convergent if it is Cauchy.

5.14 Definition.

A set D is complete if every Cauchy sequence in D is convergent to a limit $L \in D$.

5.15 Theorem.

$\mathbb{R}$ is complete.

5.16 Proof.

(Continued.) Suppose that a_n is a Cauchy sequence in $\mathbb{R}$ that is not bounded. Then, for all $\epsilon > 0$, there exists an $N_\epsilon \in \mathbb{N}$ such that

$$d(a_n, a_m) < \epsilon, \forall n, m \geq N_\epsilon.$$

But a_n is not bounded, so for any $\epsilon > 0$ and $m \in \mathbb{N}$, there exists an $n \geq m$ such that

$$d(a_n, a_m) \geq \epsilon$$

because $a_m \pm \epsilon$ would otherwise be a bound for a_n. Hence, we have a contradiction, so $\mathbb{R}$ is complete.

5.17 Lemma.

x such that $x^2 = 2$ is irrational.

5.18 Proof.

(Continued.) Suppose that x such that $x^2 = 2$ is rational, i.e., that $x = p/q$ for some $p, q \in \mathbb{Z}$ coprime. Then,

$$\begin{aligned} \frac{p^2}{q^2} &= 2 \\ p^2 &= 2q^2. \end{aligned}$$

Hence, p^2 is even, which means that p is even. Let $p = 2k$ for some $k \in \mathbb{Z}$. Then,

$$\begin{aligned} 4k^2 &= 2q^2 \\ 2k^2 &= q^2. \end{aligned}$$

Hence, q^2 is even, which means that q is even. But p and q are coprime, so we have a contradiction. Therefore, x such that $x^2 = 2$ is irrational.

5.19 Theorem.

$\mathbb{Q}$ is not complete.

5.20 Proof.

(Continued.) Recall that a field is complete if every Cauchy sequence in the field converges to a limit in the field We shall construct a Cauchy sequence in $\mathbb{Q}$ that does not converge to a limit in $\mathbb{Q}$.

The Sequence. Consider the function $f(x) = x^2 - 2$. Recall Newton's method for finding roots of $y = f(x)$, wherein iterates are defined as

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$

We have that

$$f'(x) = 2x,$$

so

$$\begin{aligned} x_{n+1} &= x_n - \frac{x_n^2 - 2}{2x_n} \\ &= \frac{1}{2}x_n + \frac{1}{x_n}. \end{aligned}$$

Boundedness. Suppose that $x_n = \sqrt{2} + \epsilon$ for some $\epsilon > 0$. We then have that

$$\begin{aligned} x_{n+1} &= \frac{1}{2}(\sqrt{2} + \epsilon) + \frac{1}{\sqrt{2} + \epsilon} \\ &= \frac{(\sqrt{2} + \epsilon)^2/2 + 1}{\sqrt{2} + \epsilon} \\ &= \sqrt{2} + \frac{\epsilon^2}{2(\sqrt{2} + \epsilon)} \\ &> \sqrt{2}. \end{aligned}$$

For $\epsilon > 0$, the inequality always holds, i.e., if $x_n > \sqrt{2}$, then $x_{n+1} > \sqrt{2}$.

Monotonacity. We also have that

$$\begin{aligned} x_{n+1} - x_n &= \sqrt{2} + \frac{\epsilon^2}{2(\sqrt{2} + \epsilon)} - \sqrt{2} - \epsilon \\ &\leq \frac{\epsilon^2}{\epsilon} - \epsilon \\ x_{n+1} - x_n &\leq 0. \end{aligned}$$

For $\epsilon > 0$, the inequality always holds, i.e., if $x_n > \sqrt{2}$, then $x_{n+1} \leq x_n$. In fact, the sequence is strictly decreasing such that $x_{n+1} < x_n$.

Convergence. Furthermore, the sequence x_n is nonempty and bounded below, so it must have a greatest lower bound L. Suppose that x_n does not converge to L. Then,

$$\exists \epsilon > 0 \text{ s.t. } \neg \exists N \in \mathbb{N} \text{ s.t. } |x_n - L| < \epsilon, \quad \forall n \geq N.$$

Recall that L is a lower bound, so the only way for this statement to hold is if $x_n \geq L + \epsilon$ for all $n \geq N$. But then $L + \epsilon$ is a lower bound, which means that L is not the greatest lower bound, hence a contradiction. Therefore, the sequence x_n converges to L.

L cannot be strictly less than $\sqrt{2}$ because $\sqrt{2}$ is a lower bound of x_n. L cannot be strictly greater than $\sqrt{2}$ because we define $x_1 = \sqrt{2} + \epsilon$ for some $\epsilon > 0$ and the sequence is strictly decreasing. Hence, $L = \sqrt{2}$.

Rationality. Suppose that $x_1 = 3/2$. Then, $x_n \in \mathbb{Q}$ for all $n \in \mathbb{N}$. We shall prove this fact by induction.

(1) *Base Case.* $x_1 = 3/2 \in \mathbb{Q}$.

(2) *Inductive Hypothesis.* Suppose that $x_n \in \mathbb{Q}$ for some $n \in \mathbb{N}$, i.e., that $x_n = p/q$ for some $p, q \in \mathbb{Z}$ coprime.

(3) *Inductive Step.* We have that

$$\begin{aligned} x_{n+1} &= \frac{1}{2}x_n + \frac{1}{x_n} \\ &= \frac{1}{2}\frac{p}{q} + \frac{q}{p} \in \mathbb{Q} \end{aligned}$$

because the sums and products of rationals are rational. Therefore, each element $x_n, n \in \mathbb{N}$, for $x_1 = 3/2$, is in $\mathbb{Q}$.

Conclusion. We have hence constructed a sequence x_n, for $x_1 = 3/2$, that is in $\mathbb{Q}$ but converges to a limit $L = \sqrt{2}$ that is not in $\mathbb{Q}$. Therefore, $\mathbb{Q}$ is not complete.

5.21 Definition.

A series is a sequence s_n of partial sums

$$s_n = \sum_{k=1}^{n} a_k.$$

for some sequence $\{a_n\}$.

5.22 Lemma.

Suppose that $a_n \in \mathbb{R}, a_n \geq 0$, i.e., that the series s_n is monotone increasing. Then, if s_n is bounded above, then s_n converges to its least supper bound, i.e.,

$$\lim_{n\to\infty} s_n = \sup s_n.$$

We often denote this limit as

$$S = \sup s_n.$$

6.0 Introduction.

Probability Spaces.

6.1 Definition.

A $\sigma-$algebra $\mathcal{F}$ is a collection of subsets of Ω such that

(1) $\emptyset \in \mathcal{F}$.

(2) If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$. (Closure under complement.)

(3) If $A_1, A_2, \ldots \in \mathcal{F}$, then

$$\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}.$$

(Closure under countable union.)

6.2 Definition.

The σ-algebra generated by a collection of subsets $\mathcal{A}$ is the smallest σ-algebra containing $\mathcal{A}$. We denote the σ-algebra generated by $\mathcal{A}$ as

$$\sigma(\mathcal{A}).$$

6.3 Example.

The σ-algebra generated by A is

$$\sigma(A) = \{\emptyset, A, A^c, \Omega\}.$$

6.4 Lemma.

If $A_1, A_2, \ldots, A_n$ partition Ω, then

$$\sigma(A_1, A_2, \ldots, A_n) = \bigcup_{i \in S} A_i$$

for all $S \subseteq \{1, 2, \ldots, n\}$. That is,

$$\sigma(A_1, A_2, \ldots, A_n) = 2^{\Omega}.$$

6.5 Example.

Consider the coin-flip space

$$\Omega = \{\omega = (\omega_1, \omega_2, \ldots) : \omega_i \in \{0, 1\}\},$$

that is, the sample space consisting of countable-length strings of 0 and 1. Denote

$$A_x = \{\omega \in \Omega : \omega_1 = x\}$$
$$A_{xy} = \{\omega \in \Omega : \omega_1 = x, \omega_2 = y\}$$

We claim that the σ-algebra $\mathcal{F}$ generated by the events

$$A_1, \quad A_{11}, \quad A_{01}$$

has cardinality $|\mathcal{F}| = 2^4$. Observe that we can form a partition of Ω by

$$P = \{A_{00}, A_{01}, A_{10}, A_{11}\}.$$

Therefore,

$$\mathcal{F} = 2^P$$

with cardinality $|\mathcal{F}| = 2^4$.

6.6 Definition.

A probability measure P is a function $P : \mathcal{F} \to [0, 1]$ such that

(1) $P : \mathcal{F} \to [0, 1]$. (Non-negativity.)

(2) If $A_1, A_2, \ldots \in \mathcal{F}$ are disjoint, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

(Countable additivity.)

6.7 Definition.

A probability space is a triple $(\Omega, \mathcal{F}, P)$ where the event space F is a σ-algebra of subsets of the sample space Ω and $P : \mathcal{F} \to [0, 1]$ is a probability measure on $\mathcal{F}$.

6.8 Definition.

A real-valued random variable is a function

$$X : \Omega \to \mathbb{R}$$

with a $(F, \mathcal{B}(\mathbb{R}))$-measurability requirement. That is, for all $B \in \mathcal{B}(\mathbb{R})$,

$$X^{-1}(B) \in \mathcal{F}$$

where

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\}$$

is the preimage of B under X and where $\mathcal{B}(\mathbb{R})$ is the Borel σ-algebra on $\mathbb{R}$, i.e., the σ-algebra generated by the open intervals of $\mathbb{R}$.

6.9 Example.

Consider the random variable $X = I_A$, that is, indicator function of the event A. Then, for all $B \in \mathcal{B}(\mathbb{R})$,

(1) If $0 \notin B$ and $1 \notin B$, then $X^{-1}(B) = \emptyset$.

(2) If $0 \notin B$ and $1 \in B$, then $X^{-1}(B) = A$.

(3) If $0 \in B$ and $1 \notin B$, then $X^{-1}(B) = A^c$.

(4) If $0 \in B$ and $1 \in B$, then $X^{-1}(B) = \Omega$.

Therefore, X is a random variable for

$$\mathcal{F} = \{\emptyset, A, A^c, \Omega\}.$$

6.10 Example.

Consider the sample space of two die rolls, that is

$$\Omega = \{(i, j) : i, j \in \{1, 2, 3, 4, 5, 6\}\}.$$

Let $X(\omega) = \omega_1 + \omega_2$ be the sum of the two die rolls. Consider B such that $B \cap \Omega = 3$ with preimage

$$X^{-1}(B) = \{(1, 2), (2, 1)\}.$$

So $\mathcal{F}$ must include $\{(1, 2), (2, 1)\}$. But if A includes $(1, 2)$, then it must also include $(2, 1)$. Therefore,

$$\sigma(X^{-1}(\mathcal{B}(\mathbb{R}))) \neq 2^{\Omega}.$$

In fact, for any $B \in \mathcal{B}(\mathbb{R})$,

$$X^{-1}(B) = X^{-1}(B \cap \text{image}(X)).$$

In particular, $\mathcal{F}$ is the σ-algebra generated by

$$A = \{X^{-1}(\{2\}), X^{-1}(\{3\}), \ldots, X^{-1}(\{12\})\},$$

i.e.,

$$\mathcal{F} = 2^A.$$

6.11 Example.

(Continued.) Recall that

$$\mathcal{F} = 2^A.$$

Consider the random variable

$$W_1 = \omega_1$$

for $\omega \in \Omega$. But the preimage

$$W_1^{-1}(1) = \{(1,1), (1,2), \ldots, (1,6)\},$$

is not in $\mathcal{F}$, so W_1 is not a random variable defined on the specified probability space.

6.12 Definition.

If $X : \Omega \to \mathbb{R}$ is a random variable, then the σ-algebra generated by X is

$$\sigma(X) = \sigma\left(\{X^{-1}(B) : B \in \mathcal{B}(\mathbb{R})\}\right).$$

7.0 Introduction.

Convergence of Random Variables.

7.1 Definition.

Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let

$$\{X_n : n \in \mathbb{N}\}$$

be a sequence of rvs on Ω and let X also be an rv on Ω. Then, X_n converges pointwise everywhere to X if for all $\omega \in \Omega$,

$$X_n(\omega) \to X(\omega).$$

That is, X_n converges pointwise everywhere to X if

$$\forall \omega \in \Omega, \forall \epsilon > 0, \exists N_\epsilon \in \mathbb{N} \text{ s.t. } |X_n(\omega) - X(\omega)| < \epsilon, \forall n > N_\epsilon.$$

7.2 Definition.

Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let

$$\{X_n : n \in \mathbb{N}\}$$

be a sequence of rvs on Ω and let X also be an rv on Ω. Then, X_n converges pointwise with probability one to X if there exists a set A with $P(A) = 0$ such that

$$\forall \omega \in A^c, X_n(\omega) \to X(\omega).$$

7.3 Definition.

Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let

$$\{X_n : n \in \mathbb{N}\}$$

be a sequence of rvs on Ω and let X also be an rv on Ω. Then, X_n converges uniformly everywhere to X if

$$\forall \epsilon > 0, \exists N_\epsilon \in \mathbb{N} \text{ s.t. } |X_n(\omega) - X(\omega)| < \epsilon, \forall n > N_\epsilon$$

for all $\omega \in \Omega$.

7.4 Definition.

Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let

$$\{X_n : n \in \mathbb{N}\}$$

be a sequence of rvs on Ω and let X also be an rv on Ω. Then, X_n converges uniformly with probability one to X if there exists a set A with $P(A) = 0$ such that

$$\forall \epsilon > 0, \exists N_\epsilon \in \mathbb{N} \text{ s.t. } |X_n(\omega) - X(\omega)| < \epsilon, \forall n > N_\epsilon$$

for all $\omega \in A^c$.

7.5 Definition.

Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let

$$\{X_n : n \in \mathbb{N}\}$$

be a sequence of rvs on Ω and let X also be an rv on Ω. Then, X_n converges in probability to X if for all $\delta > 0$, the n, δ-problematic set

$$A_{n,\delta} = \{\omega \in \Omega : |X_n(\omega) - X(\omega)| \geq \delta\}$$

satisfies

$$P(A_{n,\delta}) \to 0 \text{ as } n \to \infty.$$

We denote this type of convergence as

$$X_n \xrightarrow{P} X.$$

7.6 Note.

For $\delta_1 < \delta_2$, we have that

$$A_{n,\delta_1} \supseteq A_{n,\delta_2}.$$

7.7 Remark.

Consistentcy

$$P\left(|\bar{X}_n - \mu| \geq \epsilon\right) \to 0 \text{ as } n \to \infty$$

is a statement about convergence in probability to the degenerate random variable $X = \mu$.

7.8 Definition.

Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let

$$\{X_n : n \in \mathbb{N}\}$$

be a sequence of rvs on Ω and let X also be an rv on Ω. Then, X_n converges in L^p to X if

$$\mathrm{E}\left(|X_n - X|^p\right) \to 0 \text{ as } n \to \infty.$$

7.9 Remark.

Consider Markov's Inequality

$$P(|X_n - X| \geq \delta) \leq \frac{\mathrm{E}(|X_n - X|)}{\delta}.$$

Then, for $p > 0$, we have that

$$P(|X_n - X|^p \geq \delta^p) \leq \frac{\mathrm{E}(|X_n - X|^p)}{\delta^p},$$

so if X_n converges in L^p to X, then X_n converges in probability to X.

7.10 Definition.

A cumulative distribution function is a function that satisfies

(C) $F : \mathbb{R} \to [0, 1]$.

(C) F is non-decreasing.

(C) $\lim_{x \to -\infty} F(x) = 0$.

(C) $\lim_{x \to \infty} F(x) = 1$.

(C) F is right-continuous.

If F is continuous, then F is a continuous cumulative distribution function.

7.11 Definition.

A continuity point of a function $f : \mathbb{R} \to \mathbb{R}$ is a point t such that

$$\forall \epsilon > 0, \exists \delta > 0 \text{ s.t. if } |x - t| < \delta \text{ then } |f(x) - f(t)| < \epsilon.$$

7.12 Definition.

A sequence of rvs X_n converges in distribution to an rv X if for all continuity points t of F,

$$\lim_{n \to \infty} F_n(t) = F(t)$$

where F_n is the cdf of X_n and F is the cdf of X. We denote this type of convergence as

$$X_n \xrightarrow{D} X.$$

7.13 Theorem.

Suppose that $g : \mathbb{R} \to \mathbb{R}$ is continuous. Then, if

$$X_n \xrightarrow{D} X,$$

then

$$g(X_n) \xrightarrow{D} g(X).$$

7.14 Lemma.

If a sequence of rvs X_n is defined on a common probability space Ω and

$$X_n \xrightarrow{D} c$$

for some degenerate random variable $c \in \mathbb{R}$, then

$$X_n \xrightarrow{P} c.$$

7.15 Theorem.

(Slutsky's Theorem.) Suppose that

$$X_n \xrightarrow{D} X, \quad Y_n \xrightarrow{D} c$$

for some degenerate random variable $c \in \mathbb{R}$. If X_n, Y_n are defined on a common Ω such that $X + Y$ and XY are well-defined, then

(1) $X_n + Y_n \xrightarrow{D} X + c$.

(2) $X_n Y_n \xrightarrow{D} cX$.

(3) $X_n / Y_n \xrightarrow{D} X/c$ if $c \neq 0$.

8.0 Introduction.

Parametric Estimation.

8.1 Remark.

The motivation for parametric estimation is the following problem. Suppose that we have a collection of iid rvs X_n with a common cdf F_θ for some $\theta \in \Theta \subseteq \mathbb{R}^k$. Suppose that we know the parametric family of F_θ but not the exact value of θ. Then, we would like to estimate θ from the observations of X_n.

8.2 Remark.

We may ask two types of questions about $\hat{\theta}$. If we investigate finite sample properties of $\hat{\theta}$, then we are interested in the properties of $\hat{\theta}$ for a fixed n. Examples of these properties are

(1) Distribution of $\hat{\theta}$

(2) $\mathrm{E}(\hat{\theta})$

(3) $\mathrm{Var}(\hat{\theta})$

If we investigate asymptotic properties of $\hat{\theta}$, then we are interested in the properties of $\hat{\theta}$ as $n \to \infty$. Examples of these properties are

(1) Consistency.

(2) Limiting Distribution.

8.3 Definition.

For an rv X with pdf f, the support of f is the set

$$\mathrm{supp}(f) = \{x \in \mathbb{R} : f(x) > 0\}.$$

For an rv X with a pmf p, the support of p is the set

$$\mathrm{supp}(p) = \{x \in \mathbb{R} : p(x) > 0\}.$$

8.4 Example.

The support of f_θ can depend on θ. Consider $X_n \sim \mathrm{unif}(0, \theta)$ for $\theta > 0$. Then, the support of f_θ is

$$\mathrm{supp}(f_\theta) = [0, \theta].$$

8.5 Definition.

Suppose that we have a collection of random variables X_n iid with a common cdf F_θ for some $\theta \in \Theta \subseteq \mathbb{R}^k$. We may write the parameter θ as the solution to the system of equations

$$\mu_r(\theta) = \mathrm{E}_\theta(X^r)$$

for $r = 1, 2, \ldots, k$ where $\mathrm{E}(X^r)$ is the rth moment of X. We denote the solution to this system of equations as the inverse function

$$\theta = g(\mu_1, \mu_2, \ldots, \mu_k).$$

Suppose that we do not know the exact value of θ but we observe a collection of iid rvs X_n with a common cdf F_θ. Then, we would like to estimate θ from the observations of X_n. In particular,

$$\hat{\theta}_{MOM} = g(\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_k)$$

where $\hat{\mu}_r$ is the rth sample moment of X_n. We call $\hat{\theta}_{MOM}$ the method of moments estimator of θ.

8.6 Example.

If $X_n \sim \mathrm{N}(\theta_1, \theta_2)$ iid, then

$$\begin{aligned}\mu_1 &= \theta_1 \\ \mu_2 &= \theta_1^2 + \theta_2,\end{aligned}$$

so

$$\begin{aligned}\hat{\theta}_{1_{MOM}} &= \hat{\mu}_1 \\ \hat{\theta}_{2_{MOM}} &= \hat{\mu}_2 - \hat{\mu}_1^2.\end{aligned}$$

Note that this method of moments estimator is simply the sample mean and sample variance, which is expected for a distribution $X \sim \mathrm{N}(\mu, \sigma^2)$. Note that $\hat{\theta}_{1_{MOM}} = \bar{X}$ is unbiased for θ_1 but that $\hat{\theta}_{2_{MOM}} = \hat{\sigma}^2$ is biased for θ_2.

8.7 Example.

Sometimes, the distribution of $\hat{\theta}$ can be determined exactly. For $X_n \sim \mathrm{Exp}(\lambda)$ iid, we have that

$$\hat{\lambda}_{MOM} = \frac{1}{\bar{X}}.$$

But we know that

$$\bar{X} \sim \mathrm{Gamma}(n, n\lambda).$$

8.8 Example.

(Continued.) We may also determine the asymptotic properties of $\hat{\lambda}_{MOM} = 1/\bar{X}$. We have that

$$P\left(\frac{1}{\bar{X}} \leq t\right) = P(\bar{X} \geq \frac{1}{t}).$$

Since $f : t \to 1/t$ is continuous for $t > 0$ and

$$\bar{X} \xrightarrow{P} \frac{1}{\lambda},$$

then

$$\hat{\lambda}_{MOM} \xrightarrow{P} \lambda.$$

8.9 Definition.

For the likelihood function

$$f(x_1, x_2, \ldots, x_n|\theta),$$

the maximum likelihood estimator of θ given the observations $X_i = x_i$ is the value of θ that maximizes the likelihood function. That is,

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} f(x_1, x_2, \ldots, x_n|\theta).$$

8.10 Example.

Suppose that X_i are iid. Then, the arg max of the likelihood function occurs at

$$\begin{aligned}\arg\max_{\theta \in \Theta} f(x_1, x_2, \ldots, x_n|\theta) &= \arg\max_{\theta \in \Theta} \log f(x_1, x_2, \ldots, x_n|\theta) \\ &= \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \log f(x_i|\theta)\end{aligned}$$

since $f : t \to \log t$ is monotone increasing. Under certain conditions, we maximize the likelihood function by setting the gradient of the log-likelihood function equal to zero. That is, setting

$$\nabla \sum_{i=1}^{n} \log f(x_i|\theta) = \left(\frac{\partial}{\partial \theta_j} \sum_{i=1}^{n} \log f(x_i|\theta) \right)$$

for $j = 1, 2, \ldots k$ equal to zero and solving for θ.

8.11 Example.

For $X_i \sim \text{Exp}(\lambda)$ iid for $\lambda > 0$, the likelihood function is

$$f(x_1, x_2, \ldots, x_n|\lambda) = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i \right).$$

Then, the log-likelihood function is

$$\log f(x_1, x_2, \ldots, x_n|\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i,$$

which is maximized at

$$0 = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$$
$$\lambda = n \left(\sum_{i=1}^{n} x_i \right)^{-1}.$$

Therefore,

$$\hat{\lambda}_{MLE} = \frac{1}{\bar{X}}.$$

Note that this agrees with the method of moments estimator for λ.

8.12 Example.

If $X_n \sim \text{unif}(0, \theta)$ iid, then

$$\mu_1 = \frac{\theta}{2},$$

so

$$\hat{\theta}_{MOM} = 2\hat{\mu}_1.$$

But some observations X_i may be greater than $\hat{\theta}_{MOM}$, which means that those observations are beyond the support of the estimated f_θ.

8.13 Example.

(Continued.) The likelihood function of $X_n \sim \text{unif}(0, \theta)$ iid is

$$f(x_1, x_2, \ldots, x_n|\theta) = \theta^{-n} I_{x_i \in [0,\theta] \forall i}.$$

Since the support of f_θ depends on θ, then $f(x|\theta)$ is not differentiable in θ, so we cannot set the derivative equal to zero. But we observe that if $\theta \leq \max\{x_i\}$, i.e., if θ is feasible, then

$$f(x_1, x_2, \ldots, x_n|\theta) = \theta^{-n},$$

which is maximized when θ is small. Therefore, the maximum likelihood estimator of θ is the smallest feasible θ, i.e.,

$$\hat{\theta}_{MLE} = \max\{x_i\}.$$

Note that $\hat{\theta}_{MLE}$ is always feasible unlike $\hat{\theta}_{MOM}$.