

EN.553.431 Honors Mathematical Statistics

Daniel Yao

Fall 2024

Contents

0	Introduction.	3
0.0	Introduction.	3
0.1	Remark.	3
0.2	Notation.	3
0.3	Notation.	3
0.4	Notation	3
1	Finite Population Samples.	4
1.0	Introduction.	4
1.1	Definition.	4
2	Confidence Intervals.	5
2.0	Introduction.	5
2.1	Theorem.	5
2.2	Definition.	5
2.3	Example.	5
2.4	Note.	5
2.5	Example.	5
2.6	Example.	5
2.7	Theorem.	6
2.8	Note.	6
3	Taylor Approximations.	7
3.0	Introduction.	7
3.1	Theorem.	7
3.2	Theorem.	7
3.3	Theorem.	7
3.4	Example.	7
3.5	Example.	8
3.6	Example.	8
4	Sample Ratio.	9
4.0	Introduction.	9
4.1	Definition.	9
4.2	Example.	9
4.3	Example.	9
4.4	Example.	9
4.5	Theorem.	10
4.6	Theorem.	10
4.7	Theorem.	11
4.8	Proof.	11
5	Real Analysis.	12
5.0	Introduction.	12
5.1	Definition.	12
5.2	Definition.	12
5.3	Definition.	12
5.4	Remark.	12
5.5	Definition.	13
5.6	Definition.	13
5.7	Theorem.	13
5.8	Definition.	13
5.9	Definition.	13

5.10	Definition.	13
5.11	Definition.	13
5.12	Lemma.	14
5.13	Proof.	14
5.14	Definition.	14
5.15	Theorem.	14
5.16	Proof.	14
5.17	Definition.	14
5.18	Lemma.	14
6	Convergence of Random Variables.	15
6.0	Introduction.	15
6.1	Definition.	15
6.2	Definition.	15
6.3	Example.	15
6.4	Lemma.	15
6.5	Definition.	15
6.6	Definition.	15
6.7	Definition.	16
6.8	Example.	16
6.9	Example.	16
6.10	Example.	16
6.11	Definition.	17
6.12	Definition.	17
6.13	Definition.	17
6.14	Definition.	17
6.15	Definition.	17
6.16	Definition.	17
6.17	Note.	18
6.18	Remark.	18
6.19	Definition.	18
6.20	Remark.	18
6.21	Definition.	18
6.22	Definition.	18
6.23	Definition.	18
6.24	Theorem.	19
6.25	Lemma.	19
6.26	Theorem.	19
7	Parametric Estimation.	20
7.0	Introduction.	20

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 $\{(x_1, y_1), \dots, (x_N, y_N)\}$. 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

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

where

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

2.0 Introduction.

Confidence Intervals.

2.1 Theorem.

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

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

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

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

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 interval for μ .

2.4 Note.

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

2.5 Example.

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

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

so

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

for n large.

2.6 Example.

If we sample without replacement and $n \ll N$, then

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

2.7 Theorem.

For the sample total

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

the CLT says that

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

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 $g : \mathbb{R} \rightarrow \mathbb{R}$ 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 $g : \mathbb{R} \rightarrow \mathbb{R}$ 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 $g : \mathbb{R}^n \rightarrow \mathbb{R}$ 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

$$\nabla g(x) = \left(\frac{\partial g}{\partial x_1}(\xi) \quad \cdots \quad \frac{\partial g}{\partial x_n}(\xi) \right)^T$$

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

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

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

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

because the first-order terms vanish when

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

3.6 Example.

For $h : \mathbb{R}^2 \rightarrow \mathbb{R}$, 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

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

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

$$\begin{aligned}\frac{\partial g}{\partial x} &= -\frac{y}{x^2}, & \frac{\partial g}{\partial y} &= \frac{1}{x} \\ \frac{\partial^2 g}{\partial x^2} &= \frac{2y}{x^3}, & \frac{\partial^2 g}{\partial x \partial y} &= -\frac{1}{x^2}, & \frac{\partial^2 g}{\partial y^2} &= 0.\end{aligned}$$

Therefore,

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

4.3 Example.

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

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

where

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

In the case of sampling with replacement, we have that

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

In the case of sampling without replacement, we have that

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

4.4 Example.

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

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

In the case of sampling with replacement, we have that

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

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

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

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

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

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

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

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

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

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

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

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

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

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:

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

4.8 Proof.

(Continued.) Consider the relation

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

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

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

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

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

If R is small or if s_x is small, then

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

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 $x \in F, x \neq 0$, 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 $d : X \times X \rightarrow \mathbb{R}$ 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

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

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

$$a_n \rightarrow L,$$

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 is 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 Definition.

A series is a sequence s_n of partial sums

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

for some sequence $\{a_n\}$.

5.18 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 upper bound, i.e.,

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

We often denote this limit as

$$S = \sup s_n.$$

6.0 Introduction.

Convergence of Random Variables.

6.1 Definition.

A σ -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 $A_1, A_2, \dots \in \mathcal{F}$, 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, \dots, A_n partition Ω , then

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

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

$$\sigma(A_1, A_2, \dots, A_n) = 2^\Omega.$$

6.5 Definition.

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

- (1) $P : \mathcal{F} \rightarrow [0, 1]$. (Non-negativity.)
- (2) If $A_1, A_2, \dots \in \mathcal{F}$ are disjoint, then

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

(Countable additivity.)

6.6 Definition.

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

6.7 Definition.

A real-valued random variable is a function

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

with a $(\mathcal{F}, \mathcal{B}(\mathbb{R}))$ -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} .

6.8 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

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

6.9 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

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

i.e.,

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

6.10 Example.

(Continued.) Recall that

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

Consider the random variable

$$W_1 = \omega_1$$

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

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

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

6.11 Definition.

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

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

6.12 Definition.

Let (Ω, \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 $\omega \in \Omega$,

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

That is, X_n converges pointwise everywhere to X if

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

6.13 Definition.

Let (Ω, \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

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

6.14 Definition.

Let (Ω, \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 $\omega \in \Omega$.

6.15 Definition.

Let (Ω, \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 $\omega \in A^c$.

6.16 Definition.

Let (Ω, \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

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

We denote this type of convergence as

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

6.17 Note.

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

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

6.18 Remark.

Consistency

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

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

6.19 Definition.

Let (Ω, \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

$$E(|X_n - X|^p) \rightarrow 0 \text{ as } n \rightarrow \infty.$$

6.20 Remark.

Consider Markov's Inequality

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

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

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

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

6.21 Definition.

A cumulative distribution function is a function that satisfies

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

$$(C) \ F \text{ is non-decreasing.}$$

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

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

$$(C) \ F \text{ is right-continuous.}$$

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

6.22 Definition.

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

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

6.23 Definition.

A sequence of random variables X_n converges in distribution to a random variable X if for all continuity points t of F ,

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

where F_n is the cdf of X_n and F is the cdf of X .

6.24 Theorem.

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

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

then

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

6.25 Lemma.

If a sequence of random variables 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

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

6.26 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 $c \neq 0.$

7.0 Introduction.

Parametric Estimation.