Math is Fun & Beautiful! - Measure-theoretic Treatment of Statistics

Sunghee Yun
sunghee.yun@gmail.com

Table of contents

Kinds of fun we can enjoy with math

Notations

- sets of numbers
  - $\mathbf{N}$: set of natural numbers, $\mathbf{Z}$: set of integers, $\mathbf{Q}$: set of rational numbers
  - $\mathbf{R}$: set of real numbers, $\mathbf{R}_+$: set of nonnegative real numbers, $\mathbf{R}_{++}$: set of positive real numbers
- sequences $\langle x_i \rangle$ and like
  - finite $\langle x_i \rangle_{i=1}^n$, infinite $\langle x_i \rangle_{i=1}^\infty$ - use $\langle x_i \rangle$ when unambiguously understood
  - similarly for other operations - $\sum x_i$, $\prod x_i$, $\cup A_i$, $\cap A_i$, $\bigtimes A_i$
  - similarly for integrals - $\int f$ for $\int_{-\infty}^{\infty} f$
- sets
  - $\tilde{A}$: complement of A, $A \sim B$: $A \cap \tilde{B}$, $A \Delta B$: $A \cap \tilde{B} \cup \tilde{A} \cap B$
  - $\mathcal{P}(A)$: set of all subsets of A
- sets in metric vector spaces
  - $\overline{A}$: closure of set A
  - A°: interior of set A
  - $\mathbf{relint}$: relative interior of set A

- $\mathbf{bd}\, A$: boundary of set A

- set algebra
  - $\sigma(\mathcal{A})$: σ-algebra generated by $\mathcal{A}$, $i.e.$, smallest σ-algebra containing $\mathcal{A}$
- norms in $\mathbf{R}^n$
  - $\|x\|_p$ $(p \geq 1)$: p-norm of $x \in \mathbf{R}^n$, $i.e.$, $(|x_1|^p + \cdots + |x_n|^p)^{1/p}$
  - $\|x\|_2$: Euclidean norm
- matrices and vectors
  - a_i: i-th entry of vector a
  - A_{ij}: entry of matrix A at position (i,j), $i.e.$, entry in i-th row and j-th column
  - $\mathbf{Tr}(A)$: trace of $A \in \mathbf{R}^{n \times n}$, $i.e.$, $A_{1,1} + \cdots + A_{n,n}$
- symmetric, positive definite, and positive semi-definite matrices
  - $\mathbf{S}^n \subset \mathbf{R}^{n \times n}$: set of symmetric matrices
  - $\mathbf{S}^n_+ \subset \mathbf{S}^n$: set of positive semi-definite matrices - $A \succeq 0 \Leftrightarrow A \in \mathbf{S}^n_+$
  - $\mathbf{S}^n_{++} \subset \mathbf{S}^n$: set of positive definite matrices - $A \succ 0 \Leftrightarrow A \in \mathbf{S}^n_{++}$
- Python script-like notations (with serious abuse of notations!)

- use $f : \mathbf{R} \to \mathbf{R}$ as if it were $f : \mathbf{R}^n \to \mathbf{R}^n$, *e.g.*,

$$\exp(x) = (\exp(x_1), \ldots, \exp(x_n)) \quad \text{for } x \in \mathbf{R}^n$$

  or

$$\log(x) = (\log(x_1), \ldots, \log(x_n)) \quad \text{for } x \in \mathbf{R}^n_{++}$$

  corresponding to Python code - `numpy.exp(x)` or `numpy.log(x)` - where `x` is instance of `numpy.ndarray`, *i.e.*, `numpy` array
- use $\sum x$ for $\mathbf{1}^T x$ for $x \in \mathbf{R}^n$, *i.e.*

$$\sum x = x_1 + \cdots + x_n$$

  corresponding to Python code - `x.sum()` - where `x` is `numpy` array
- use x/y for $x, y \in \mathbf{R}^n$ for

$$\begin{bmatrix} x_1/y_1 & \cdots & x_n/y_n \end{bmatrix}^T$$

  corresponding to Python code - `x / y` - where `x` and `y` are 1-d `numpy` arrays
  - applies to any two matrices of same dimensions

Some definitions

Definition 1. [infinitely often - i.o.] *statement, P_n, said to happen* infinitely often *or* i.o. *if*

$$(\forall N \in \mathbf{N})\,(\exists n > N)\,(P_n)$$

Definition 2. [almost everywhere - a.e.] *statement, $P(x)$, said to happen* almost everywhere *or* a.e. *or* almost surely *or* a.s. *(depending on context) associated with measure space, $(X, \mathscr{B}, \mu)$ if*

$$\mu\{x|P(x)\} = 1$$

or equivalently

$$\mu\{x| \sim P(x)\} = 0$$

Some conventions

- for some subjects, use following conventions
  - $0 \cdot \infty = \infty \cdot 0 = 0$
  - $(\forall x \in \mathbf{R}_{++})(x \cdot \infty = \infty \cdot x = \infty)$
  - $\infty \cdot \infty = \infty$

Measure-theoretic Treatment of Probabilities

Probability Measure

Measurable functions

– denote *n-dimensional Borel sets* by $\mathscr{R}^n$

- for two measurable spaces, $(\Omega, \mathscr{F})$ and $(\Omega', \mathscr{F}')$, function, $f : \Omega \to \Omega'$ with

$$(\forall A' \in \mathscr{F}') \left(f^{-1}(A') \in \mathscr{F} \right)$$

  said to be *measurable with respect to $\mathscr{F}/\mathscr{F}'$* (thus, measurable functions defined on page **??** and page **??** can be said to be measurable with respect to $\mathcal{B}/\mathscr{R}$)

- when $\Omega = \mathbf{R}^n$ in $(\Omega, \mathscr{F})$, $\mathscr{F}$ is assumed to be $\mathscr{R}^n$, and sometimes drop $\mathscr{R}^n$
  - thus, $e.g.$, we say $f : \Omega \to \mathbf{R}^n$ is measurable with respect to $\mathscr{F}$ (instead of $\mathscr{F}/\mathscr{R}^n$)

- measurable function, $f : \mathbf{R}^n \to \mathbf{R}^m$ ($i.e.$, measurable with respect to $\mathscr{R}^n/\mathscr{R}^m$), called *Borel functions*

- $f : \Omega \to \mathbf{R}^n$ is measurable with respect to $\mathscr{F}/\mathscr{R}^n$ *if and only if* every component, $f_i : \Omega \to \mathbf{R}$, is measurable with respect to $\mathscr{F}/\mathscr{R}$

Probability (measure) spaces

- set function, $P : \mathscr{F} \to [0,1]$, defined on algebra, $\mathscr{F}$, of set Ω, satisfying following properties, called *probability measure* (refer to page **??** for resumblance with measurable spaces)
  - $(\forall A \in \mathscr{F})(0 \leq P(A) \leq 1)$
  - $P(\emptyset) = 0,\ P(\Omega) = 1$
  - $(\forall$ disjoint $\langle A_n \rangle \subset \mathscr{F})(P(\bigcup A_n) = \sum P(A_n))$

- for σ-algebra, $\mathscr{F}$, $(\Omega, \mathscr{F}, P)$, called *probability measure space* or *probability space*

- set $A \in \mathscr{F}$ with $P(A) = 1$, called *a support of P*

Dynkin's π-λ theorem

- class, $\mathcal{P}$, of subsets of Ω closed under finite intersection, called *π-system*, $i.e.$,
  - $(\forall A, B \in \mathcal{P})(A \cap B \in \mathcal{P})$
- class, $\mathcal{L}$, of subsets of Ω containing Ω closed under complements and countable disjoint unions called *λ-system*
  - $\Omega \in \mathcal{L}$
  - $(\forall A \in \mathcal{L})(\tilde{A} \in \mathcal{L})$
  - $(\forall$ disjoint $\langle A_n \rangle)(\bigcup A_n \in \mathcal{L})$
- *class that is both π-system and λ-system is σ-algebra*
- *Dynkin's π-λ theorem* - for π-system, $\mathcal{P}$, and λ-system, $\mathcal{L}$, with $\mathcal{P} \subset \mathcal{L}$,

$$\sigma(\mathcal{P}) \subset \mathcal{L}$$

- for π-system, $\mathscr{P}$, two probability measures, P_1 and P_2, on $\sigma(\mathscr{P})$, agreeing $\mathscr{P}$, agree on $\sigma(\mathscr{P})$

Limits of Events

Theorem 1. [convergence-of-events] *no for sequence of subsets, $\langle A_n \rangle$,*

$$P(\liminf A_n) \leq \liminf P(A_n) \leq \limsup P(A_n) \leq P(\limsup A_n)$$

- *for $\langle A_n \rangle$ converging to A*

$$\lim P(A_n) = P(A)$$

Theorem 2. [independence-of-smallest-sig-alg] *no for sequence of π-systems, $\langle \mathscr{A}_n \rangle$, $\langle \sigma(\mathscr{A}_n) \rangle$ is independent*

Probabilistic independence

– given probability space, $(\Omega, \mathscr{F}, P)$

- $A, B \in \mathscr{F}$ with

$$P(A \cap B) = P(A)P(B)$$

said to be *independent*

- indexed collection, $\langle A_\lambda \rangle$, with

$$(\forall n \in \mathbf{N},\ \text{distinct } \lambda_1, \ldots, \lambda_n \in \Lambda) \left(P\left(\bigcap_{i=1}^n A_{\lambda_i}\right) = \prod_{i=1}^n P(A_{\lambda_i}) \right)$$

said to be *independent*

Independence of classes of events

- indexed collection, $\langle \mathcal{A}_\lambda \rangle$, of classes of events ($i.e.$, subsets) with

$$(\forall A_\lambda \in \mathcal{A}_\lambda)\ (\langle A_\lambda \rangle \text{ are independent})$$

  said to be *independent*

- *for independent indexed collection, $\langle \mathcal{A}_\lambda \rangle$, with every $\mathcal{A}_\lambda$ being π-sytem, $\langle \sigma(\mathcal{A}_\lambda) \rangle$ are independent*

- for independent (countable) collection of events, $\langle \langle A_{ni} \rangle_{i=1}^{\infty} \rangle_{n=1}^{\infty}$, $\langle \mathscr{F}_n \rangle_{n=1}^{\infty}$ with $\mathscr{F}_n = \sigma(\langle A_{ni} \rangle_{i=1}^{\infty})$ are independent

Borel-Cantelli lemmas

- **Lemma 1. [first Borel-Cantelli]** *for sequence of events, $\langle A_n \rangle$, with $\sum P(A_n)$ converging*
$$P(\limsup A_n) = 0$$

- **Lemma 2. [second Borel-Cantelli]** *for independent sequence of events, $\langle A_n \rangle$, with $\sum P(A_n)$ diverging*
$$P(\limsup A_n) = 1$$

Tail events and Kolmogorov's zero-one law

- for sequence of events, $\langle A_n \rangle$

$$\mathscr{T} = \bigcap_{n=1}^{\infty} \sigma\left(\langle A_i \rangle_{i=n}^{\infty}\right)$$

  called *tail σ-algebra associated with $\langle A_n \rangle$*; its lements are called *tail events*

- *Kolmogorov's zero-one law -* for independent sequence of events, $\langle A_n \rangle$ every event in tail σ-algebra has probability measure either 0 or 1

Product probability spaces

- for two measure spaces, $(X, \mathscr{X}, \mu)$ and $(Y, \mathscr{Y}, \nu)$, want to find product measure, π, such that

$$(\forall A \in \mathscr{X}, B \in \mathscr{Y})\,(\pi(A \times B) = \mu(A)\nu(B))$$

  – $e.g.$, if both μ and ν are Lebesgue measure on $\mathbf{R}$, π will be Lebesgue measure on $\mathbf{R}^2$

- $A \times B$ for $A \in \mathscr{X}$ and $B \in \mathscr{Y}$ is *measurable rectangle*

- *σ-algebra generated by measurable rectangles* denoted by

$$\mathscr{X} \times \mathscr{Y}$$

  – thus, *not* Cartesian product in usual sense
  – generally *much larger* than class of measurable rectangles

Sections of measurable subsets and functions

for two measure spaces, $(X, \mathscr{X}, \mu)$ and $(Y, \mathscr{Y}, \nu)$

- sections of measurable subsets
  - $\{y \in Y | (x, y) \in E\}$ is *section of E determined by x*
  - $\{x \in X | (x, y) \in E\}$ is *section of E determined by y*
- sections of measurable functions - for measurable function, f, with respect to $\mathscr{X} \times \mathscr{Y}$
  - $f(x, \cdot)$ is *section of f determined by x*
  - $f(\cdot, y)$ is *section of f determined by y*
- sections of measurable subsets are measurable
  - $(\forall x \in X, E \in \mathscr{X} \times \mathscr{Y}) (\{y \in Y | (x, y) \in E\} \in \mathscr{Y})$
  - $(\forall y \in Y, E \in \mathscr{X} \times \mathscr{Y}) (\{x \in X | (x, y) \in E\} \in \mathscr{X})$
- sections of measurable functions are measurable
  - $f(x, \cdot)$ is measurable with respect to $\mathscr{Y}$ for every $x \in X$
  - $f(\cdot, y)$ is measurable with respect to $\mathscr{X}$ for every $y \in Y$

Product measure

for two σ-finite measure spaces, $(X, \mathscr{X}, \mu)$ and $(Y, \mathscr{Y}, \nu)$

- two functions defined below for every $E \in \mathscr{X} \times \mathscr{Y}$ are σ-finite measures
  - $\pi'(E) = \int_X \nu\{y \in Y | (x, y) \in E\} d\mu$
  - $\pi''(E) = \int_Y \mu\{x \in X | (x, y) \in E\} d\nu$
- for every measurable rectangle, $A \times B$, with $A \in \mathscr{X}$ and $B \in \mathscr{Y}$

$$\pi'(A \times B) = \pi''(A \times B) = \mu(A)\nu(B)$$

(use conventions in page 7 for extended real values)

- indeed, $\pi'(E) = \pi''(E)$ for every $E \in \mathscr{X} \times \mathscr{Y}$; let $\pi = \pi' = \pi''$
- π is
  - called *product measure* and denoted by $\mu \times \nu$
  - σ-finite measure
  - *only* measure such that $\pi(A \times B) = \mu(A)\nu(B)$ for every measurable rectangle

Fubini's theorem

- suppose two σ-finite measure spaces, $(X, \mathscr{X}, \mu)$ and $(Y, \mathscr{Y}, \nu)$ - define
  - $X_0 = \{x \in X | \int_Y |f(x,y)| d\nu < \infty\} \subset X$
  - $Y_0 = \{y \in Y | \int_X |f(x,y)| d\nu < \infty\} \subset Y$
- *Fubini's theorem* - for nonnegative measurable function, f, following are measurable with respect to $\mathscr{X}$ and $\mathscr{Y}$ respectively

$$g(x) = \int_Y f(x,y) d\nu, \quad h(y) = \int_X f(x,y) d\mu$$

and following holds

$$\int_{X \times Y} f(x,y) d\pi = \int_X \left(\int_Y f(x,y) d\nu \right) d\mu = \int_Y \left(\int_X f(x,y) d\mu \right) d\nu$$

- for f, (not necessarily nonnegative) integrable function with respect to π
  - $\mu(X \sim X_0) = 0$, $\nu(Y \sim Y_0) = 0$
  - g and h are finite measurable on X_0 and Y_0 respectively
  - (above) equalities of *double integral* holds

Random Variables

Random variables

– for probability space, $(\Omega, \mathscr{F}, P)$,

- measurable function (with respect to $\mathscr{F}/\mathscr{R}$), $X : \Omega \to \mathbf{R}$, called *random variable*
- measurable function (with respect to $\mathscr{F}/\mathscr{R}^n$), $X : \Omega \to \mathbf{R}^n$, called *random vector*
  – when expressing $X(\omega) = (X_1(\omega), \ldots, X_n(\omega))$, X is measurable *if and only if* every X_i is measurable
  – thus, n-dimensional random vaector is simply n-tuple of random variables
- smallest σ-algebra with respect to which X is measurable, called *σ-algebra generated by X* and denoted by $\sigma(X)$
  – $\sigma(X)$ consists exactly of sets, $\{\omega \in \Omega | X(\omega) \in H\}$, for $H \in \mathscr{R}^n$
  – random variable, Y, is measurable with respect to $\sigma(X)$ *if and only if* exists measurable function, $f : \mathbf{R}^n \to \mathbf{R}$ such that $Y(\omega) = f(X(\omega))$ for all ω, $i.e.$, $Y = f \circ X$

Probability distributions for random variables

- probability measure on **R**, $\mu = PX^{-1}$, $i.e.$,

$$\mu(A) = P(X \in A) \text{ for } A \in \mathscr{R}$$

  called *distribution* or *law* of random variable, X

- function, F : **R** $\to [0,1]$, defined by

$$F(x) = \mu(-\infty, x] = P(X \leq x)$$

  called *distribution function* or *cumulative distribution function (CDF)* of X

- Borel set, S, with $P(S) = 1$, called *support*
- random variable, its distribution, its distribution function, said to be *discrete* when has *countable* support

Probability distribution of mappings of random variables

- for measurable $g : \mathbf{R} \to \mathbf{R}$,

$$(\forall A \in \mathscr{R}) \left(\mathbf{Prob}\left(g(X) \in A\right) = \mathbf{Prob}\left(X \in g^{-1}(A)\right) = \mu(g^{-1}(A)) \right)$$

  hence, $g(X)$ has distribution of μg^{-1}

Probability density for random variables

- Borel function, $f : \mathbf{R} \to \mathbf{R}_+$, satisfying

$$(\forall A \in \mathscr{R}) \left(\mu(A) = P(X \in A) = \int_A f(x) dx \right)$$

  called *density* or *probability density function (PDF)* of random variable
- above is equivalent to

$$(\forall a < b \in \mathbf{R}) \left(\int_a^b f(x) dx = P(a < X \leq b) = F(b) - F(a) \right)$$

  (refer to statement on page 12)
  - note, though, F does not need to differentiate to f everywhere; only f required to integrate properly
  - if F does differentiate to f and f is continuous, *fundamental theorem of calculus* implies f indeed is density for F

Probability distribution for random vectors

- (similarly to random variables) probability measure on $\mathbf{R}^n$, $\mu = PX^{-1}$, *i.e.*,

$$\mu(A) = P(X \in A) \text{ for } A \in \mathscr{B}^k$$

  called *distribution* or *law* of random vector, X

- function, $F : \mathbf{R}^k \to [0, 1]$, defined by

$$F(x) = \mu S_x = P(X \preceq x)$$

  where

$$S_x = \{\omega \in \Omega | X(\omega) \preceq x\} = \{\omega \in \Omega | X_i(\omega) \leq x_i\}$$

  called *distribution function* or *cumulative distribution function (CDF)* of X

- (similarly to random variables) random vector, its distribution, its distribution function, said to be *discrete* when has *countable* support

Marginal distribution for random vectors

- (similarly to random variables) for measurable $g : \mathbf{R}^n \to \mathbf{R}^m$

$$(\forall A \in \mathscr{R}^m) \left(\mathbf{Prob}\,(g(X) \in A) = \mathbf{Prob}\left(X \in g^{-1}(A)\right) = \mu(g^{-1}(A))\right)$$

  hence, $g(X)$ has distribution of μg^{-1}
- for $g_i : \mathbf{R}^n \to \mathbf{R}$ with $g_i(x) = x_i$

$$(\forall A \in \mathscr{R})\,(\mathbf{Prob}\,(g(X) \in A) = \mathbf{Prob}\,(X_i \in A))$$

- measure, μ_i, defined by $\mu_i(A) = \mathbf{Prob}\,(X_i \in A)$, called *(i-th) marginal distribution of X*
- for μ having density function, $f : \mathbf{R}^n \to \mathbf{R}_+$, density function of marginal distribution is

$$f_i(x) = \int_{\mathscr{R}^{n-1}} f(x_{-i}) d\mu_{-i}$$

  where $x_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ and similarly for $d\mu_{-i}$

Independence of random variables

- random variables, X_1, . . . , X_n, with independent σ-algebras generated by them, said to be *independent*

  (refer to page 15 for independence of collections of subsets)

  – because $\sigma(X_i) = X_i^{-1}(\mathscr{R}) = \{X_i^{-1}(H) | H \in \mathscr{R}\}$, independent *if and only if*

$$(\forall H_1, \ldots, H_n \in \mathscr{R}) \left(P\left(X_1 \in H_1, \ldots, X_n \in H_n\right) = \prod P\left(X_i \in H_i\right) \right)$$

$i.e.$,

$$(\forall H_1, \ldots, H_n \in \mathscr{R}) \left(P\left(\bigcap X_i^{-1}(H_i) \right) = \prod P\left(X_i^{-1}(H_i) \right) \right)$$

Equivalent statements of independence of random variables

- for random variables, $X_1, \ldots, X_n$, having μ and $F : \mathbf{R}^n \to [0,1]$ as their distribution and CDF, with each X_i having μ_i and $F_i : \mathbf{R} \to [0,1]$ as its distribution and CDF, following statements are *equivalent*
  - $X_1, \ldots, X_n$ are independent
  - $(\forall H_1, \ldots, H_n \in \mathscr{R}) \left(P\left(\bigcap X_i^{-1}(H_i)\right) = \prod P\left(X_i^{-1}(H_i)\right)\right)$
  - $(\forall H_1, \ldots, H_n \in \mathscr{R}) \left(P(X_1 \in H_1, \ldots, X_n \in H_n) = \prod P(X_i \in H_i)\right)$
  - $(\forall x \in \mathbf{R}^n) \left(P(X_1 \leq x_1, \ldots, X_n \leq x_n) = \prod P(X_i \leq x_i)\right)$
  - $(\forall x \in \mathbf{R}^n) \left(F(x) = \prod F_i(x_i)\right)$
  - $\mu = \mu_1 \times \cdots \times \mu_n$
  - $(\forall x \in \mathbf{R}^n) \left(f(x) = \prod f_i(x_i)\right)$

Independence of random variables with separate σ-algebra

– given probability space, $(\Omega, \mathscr{F}, P)$

- random variables, $X_1, \ldots, X_n$, each of which is measurable with respect to each of n independent σ-algebras, $\mathscr{G}_1 \subset \mathscr{F}, \ldots, \mathscr{G}_n \subset \mathscr{F}$ respectively, are independent

Independence of random vectors

- for random vectors, $X_1 : \Omega \to \mathbf{R}^{d_1}, \ldots, X_n : \Omega \to \mathbf{R}^{d_n}$, having μ and $F : \mathbf{R}^{d_1} \times \cdots \times \mathbf{R}^{d_n} \to [0,1]$ as their distribution and CDF, with each X_i having μ_i and $F_i : \mathbf{R}^{d_i} \to [0,1]$ as its distribution and CDF, following statements are *equivalent*
  - $X_1, \ldots, X_n$ are independent
  - $\left(\forall H_1 \in \mathscr{R}^{d_1}, \ldots, H_n \in \mathscr{R}^{d_n}\right) \left(P\left(\bigcap X_i^{-1}(H_i)\right) = \prod P\left(X_i^{-1}(H_i)\right)\right)$
  - $\left(\forall H_1 \in \mathscr{R}^{d_1}, \ldots, H_n \in \mathscr{R}^{d_n}\right) \left(P(X_1 \in H_1, \ldots, X_n \in H_n) = \prod P(X_i \in H_i)\right)$
  - $\left(\forall x_1 \in \mathbf{R}^{d_1}, \ldots, x_n \in \mathbf{R}^{d_n}\right) \left(P(X_1 \preceq x_1, \ldots, X_n \preceq x_n) = \prod P(X_i \preceq x_i)\right)$
  - $\left(\forall x_1 \in \mathbf{R}^{d_1}, \ldots, x_n \in \mathbf{R}^{d_n}\right) \left(F(x_1, \ldots, x_n) = \prod F_i(x_i)\right)$
  - $\mu = \mu_1 \times \cdots \times \mu_n$
  - $\left(\forall x_1 \in \mathbf{R}^{d_1}, \ldots, x_n \in \mathbf{R}^{d_n}\right) \left(f(x_1, \ldots, x_n) = \prod f_i(x_i)\right)$

Independence of infinite collection of random vectors

- infinite collection of random vectors for which every finite subcollection is independent, said to be *independent*

- for independent (countable) collection of random vectors, $\langle\langle X_{ni}\rangle_{i=1}^{\infty}\rangle_{n=1}^{\infty}$, $\langle\mathscr{F}_n\rangle_{n=1}^{\infty}$ with $\mathscr{F}_n = \sigma(\langle X_{ni}\rangle_{i=1}^{\infty})$ are independent

Probability evaluation for two independent random vectors

Theorem 3. [Probability evaluation for two independent random vectors] *for independent random vectors, X and Y, with distributions, μ and ν, in $\mathbf{R}^n$ and $\mathbf{R}^m$ respectively*

$$\left(\forall B \in \mathscr{R}^{n+m}\right)\left(\mathbf{Prob}\left((X,Y)\in B\right) = \int_{\mathbf{R}^n} \mathbf{Prob}\left((x,Y)\in B\right) d\mu_X\right)$$

and

$$\left(\forall A \in \mathscr{R}^{n}, B \in \mathscr{R}^{n+m}\right)\left(\mathbf{Prob}\left(X \in A, (X,Y)\in B\right) = \int_{A} \mathbf{Prob}\left((x,Y)\in B\right) d\mu_X\right)$$

Sequence of random variables

Theorem 4. [squence of random variables] *for sequence of probability measures on $\mathscr{R}$, $\langle \mu_n \rangle$, exists probability space, (X, Ω, P), and sequence of independent random variables in* **R**, $\langle X_n \rangle$, *such that each X_n has μ_n as distribution*

Expected values

Definition 3. [expected values] *for random variable, X, on $(\Omega, \mathscr{F}, P)$, integral of X with respect to measure, P*

$$\mathbf{E}\, X = \int X dP = \int_\Omega X(\omega) dP$$

called expected value of X

- $\mathbf{E}\, X$ is
  - always defined for nonnegative X
  - for general case
    - defined, or
    - X has an expected value if either $\mathbf{E}\, X^+ < \infty$ or $\mathbf{E}\, X^- < \infty$ or both, in which case, $\mathbf{E}\, X = \mathbf{E}\, X^+ - \mathbf{E}\, X^-$
- X is integrable *if and only if* $\mathbf{E}\, |X| < \infty$
- limits
  - if $\langle X_n \rangle$ is dominated by integral random variable or they are uniformly integrable, $\mathbf{E}\, X_n$ converges to $\mathbf{E}\, X$ if X_n converges to X in probability

Markov and Chebyshev's inequalities

Inequality 1. [Markov inequality] *for random variable, X, on $(\Omega, \mathscr{F}, P)$,*

$$\mathbf{Prob}\,(X \geq \alpha) \leq \frac{1}{\alpha}\int_{X\geq\alpha} X\,dP \leq \frac{1}{\alpha}\,\mathbf{E}\,X$$

for nonnegative X, hence

$$\mathbf{Prob}\,(|X| \geq \alpha) \leq \frac{1}{\alpha^n}\int_{|X|\geq\alpha} |X|^n dP \leq \frac{1}{\alpha^n}\,\mathbf{E}\,|X|^n$$

for general X

Inequality 2. [Chebyshev's inequality] *as special case of Markov inequality,*

$$\mathbf{Prob}\,(|X - \mathbf{E}\,X| \geq \alpha) \leq \frac{1}{\alpha^2}\int_{|X-\mathbf{E}\,X|\geq\alpha} (X - \mathbf{E}\,X)^2 dP \leq \frac{1}{\alpha^2}\,\mathbf{Var}\,X$$

for general X

Jensen's, Hölder's, and Lyapunov's inequalities

Inequality 3. [Jensen's inequality] *for random variable,* X*, on* $(\Omega, \mathscr{F}, P)$*, and convex function,* φ

$$\varphi\left(\mathbf{E}\, X\right) \mathbf{Prob}\left(X \geq \alpha\right) \leq \frac{1}{\alpha} \int_{X \geq \alpha} X dP \leq \frac{1}{\alpha} \mathbf{E}\, X$$

Inequality 4. [Holder's inequality] *for two random variables,* X *and* Y*, on* $(\Omega, \mathscr{F}, P)$*, and* $p, q \in (1, \infty)$ *with* $1/p + 1/q = 1$

$$\mathbf{E}\,|XY| \leq \left(\mathbf{E}\,|X|^p\right)^{1/p} \left(\mathbf{E}\,|X|^q\right)^{1/q}$$

Inequality 5. [Lyapunov's inequality] *for random variable,* X*, on* $(\Omega, \mathscr{F}, P)$*, and* $0 < \alpha < \beta$

$$\left(\mathbf{E}\,|X|^\alpha\right)^{1/\alpha} \leq \left(\mathbf{E}\,|X|^\beta\right)^{1/\beta}$$

- note Hölder's inequality implies Lyapunov's inequality

Maximal inequalities

Theorem 5. [Kolmogorov's zero-one lawy] *if* $A \in \mathscr{F} = \bigcap_{n=1}^{\infty} \sigma(X_n, X_{n+1}, \ldots)$ *for independent* $\langle X_n \rangle$,

$$\mathbf{Prob}\,(A) = 0 \vee \mathbf{Prob}\,(A) = 1$$

– define $S_n = \sum X_i$

Inequality 6. [Kolmogorov's maximal inequality] *for independent* $\langle X_i \rangle_{i=1}^{n}$ *with* $\mathbf{E}\, X_i = 0$ *and* $\mathbf{Var}\, X_i < \infty$ *and* $\alpha > 0$

$$\mathbf{Prob}\,(\max S_i \geq \alpha) \leq \frac{1}{\alpha} \mathbf{Var}\, S_n$$

Inequality 7. [Etemadi's maximal inequality] *for independent* $\langle X_i \rangle_{i=1}^{n}$ *and* $\alpha > 0$

$$\mathbf{Prob}\,(\max |S_i| \geq 3\alpha) \leq 3 \max \mathbf{Prob}\,(|S_i| \geq \alpha)$$

Moments

Definition 4. [moments and absolute moments] *for random variable, X, on $(\Omega, \mathscr{F}, P)$, integral of X with respect to measure, P*

$$\mathbf{E}\, X^n = \int x^k d\mu = \int x^k dF(x)$$

called k-th moment *of X or μ or F, and*

$$\mathbf{E}\, |X|^n = \int |x|^k d\mu = \int |x|^k dF(x)$$

called k-th absolute moment *of X or μ or F*

- if $\mathbf{E}\, |X|^n < \infty$, $\mathbf{E}\, |X|^k < \infty$ for $k < n$
- $\mathbf{E}\, X^n$ defined only when $\mathbf{E}\, |X|^n < \infty$

Moment generating functions

Definition 5. [moment generating function] *for random variable, X, on $(\Omega, \mathscr{F}, P)$, M : $\mathbf{C} \to \mathbf{C}$ defined by*

$$M(s) = \mathbf{E}\left(e^{sX}\right) = \int e^{sx} d\mu = \int e^{sx} dF(x)$$

called moment generating function of X

- n-th derivative of M with respect to s is $M^{(n)}(s) = \frac{d^n}{ds^n} F(s) = \mathbf{E}\left(X^n e^{sX}\right) = \int x e^{sx} d\mu$
- thus, n-th derivative of M with respect to s at $s = 0$ is n-th moment of X

$$M^{(n)}(0) = \mathbf{E}\, X^n$$

- for independent random variables, $\langle X_i \rangle_{i=1}^n$, moment generating function of $\sum X_i$

$$\prod M_i(s)$$

Convergence of Random Variables

Convergences of random variables

Definition 6. [convergence with probability 1] *random variables, $\langle X_n \rangle$, with*

$$\mathbf{Prob}\,(\lim X_n = X) = P(\{\omega \in \Omega |\, \lim X_n(\omega) = X(\omega)\}) = 1$$

said to converge to X with probability 1 *and denoted by* $X_n \to X$ a.s.

Definition 7. [convergence in probability] *random variables, $\langle X_n \rangle$, with*

$$(\forall \epsilon > 0)\,(\lim \mathbf{Prob}\,(|X_n - X| > \epsilon) = 0)$$

said to converge to X in probability

Definition 8. [weak convergence] *distribution functions, $\langle F_n \rangle$, with*

$$(\forall x \text{ in domain of } F)\,(\lim F_n(x) = F(x))$$

said to converge weakly to distribution function, F, *and denoted by* $F_n \Rightarrow F$

Definition 9. [converge in distribution] *When $F_n \Rightarrow F$, associated random variables, $\langle X_n \rangle$, said to converge in distribution to X, associated with F, and denoted by $X_n \Rightarrow X$*

Definition 10. [weak convergence of measures] *for measures on $(\mathbf{R}, \mathscr{R})$, $\langle \mu_n \rangle$, associated with distribution functions, $\langle F_n \rangle$, respectively, and measure on $(\mathbf{R}, \mathscr{R})$, μ, associated with distribution function, F, we denote*

$$\mu_n \Rightarrow \mu$$

if

$$(\forall A = (-\infty, x] \textit{ with } x \in \mathbf{R}) \left(\lim \mu_n(A) = \mu(A) \right)$$

- indeed, if above equation holds for $A = (-\infty, x)$, it holds for many other subsets

Relations of different types of convergences of random variables

Proposition 1. [relations of convergence of random variables] *convergence with probability* 1 *implies convergence in probability, which implies* $X_n \Rightarrow X$, $i.e.$

$$\begin{aligned} & X_n \to X \text{ a.s., } i.e., X_n \text{ converge to } X \text{ with probability } 1 \\ \Rightarrow \quad & X_n \text{ converge to } X \text{ in probability} \\ \Rightarrow \quad & X_n \Rightarrow X, \; i.e., X_n \text{ converge to } X \text{ in distribution}, \end{aligned}$$

Necessary and sufficient conditions for convergence of probability

X_n converge in probability

if and only if

$$(\forall \epsilon > 0)\left(\mathbf{Prob}\left(|X_n - X| > \epsilon \text{ i.o}\right) = \mathbf{Prob}\left(\limsup |X_n - X| > \epsilon\right) = 0\right)$$

if and only if

$$\left(\forall \text{ subsequence } \left\langle X_{n_k} \right\rangle\right) \left(\exists \text{ its subsequence } \left\langle X_{n_{k_l}} \right\rangle \text{ converging to } f \text{ with probability } 1\right)$$

Necessary and sufficient conditions for convergence in distribution

$$X_n \Rightarrow X, \text{ i.e., } X_n \text{ converge in distribution}$$

if and only if

$$F_n \Rightarrow F, \text{ i.e., } F_n \text{ converge weakly}$$

if and only if

$$(\forall A = (-\infty, x] \text{ with } x \in \mathbf{R}) \left(\lim \mu_n(A) = \mu(A)\right)$$

if and only if

$$(\forall x \text{ with } \mathbf{Prob}\,(X = x) = 0) \left(\lim \mathbf{Prob}\,(X_n \leq x) = \mathbf{Prob}\,(X \leq x)\right)$$

Strong law of large numbers

– define $S_n = \sum_{i=1}^n X_i$

Theorem 6. [strong law of large numbers] *for sequence of independent and identically distributed (i.i.d.) random variables with finite mean,* $\langle X_n \rangle$

$$\frac{1}{n} S_n \to \mathbf{E}\, X_1$$

with probability 1

- strong law of large numbers also called *Kolmogorov's law*

Corollary 1. [strong law of large numbers] *for sequence of independent and identically distributed (i.i.d.) random variables with* $\mathbf{E}\, X_1^- < \infty$ *and* $\mathbf{E}\, X_1^+ = \infty$ *(hence,* $\mathbf{E}\, X = \infty$*)*

$$\frac{1}{n} S_n \to \infty$$

with probability 1

Weak law of large numbers

– define $S_n = \sum_{i=1}^n X_i$

Theorem 7. [weak law of large numbers] *for sequence of independent and identically distributed (i.i.d.) random variables with finite mean, $\langle X_n \rangle$*

$$\frac{1}{n} S_n \to \mathbf{E}\, X_1$$

in probability

- because convergence with probability 1 implies convergence in probability (Proposition 1), strong law of large numbers implies weak law of large numbers

Normal distributions

– assume probability space, $(\Omega, \mathscr{F}, P)$

Definition 11. [normal distributions] *Random variable,* $X : \Omega \to \mathbf{R}$, *with*

$$(A \in \mathscr{R}) \left(\mathbf{Prob}\left(X \in A\right) = \frac{1}{\sqrt{2\pi}\sigma} \int_A e^{-(x-c)^2/2} d\mu \right)$$

where $\mu = PX^{-1}$ *for some* $\sigma > 0$ *and* $c \in \mathbf{R}$, *called* normal distribution *and denoted by* $X \sim \mathcal{N}(c, \sigma^2)$

– note $\mathbf{E}\, X = c$ and $\mathbf{Var}\, X = \sigma^2$

– called *standard normal distribution* when $c = 0$ and $\sigma = 1$

Multivariate normal distributions

– assume probability space, $(\Omega, \mathscr{F}, P)$

Definition 12. [multivariate normal distributions] *Random variable,* $X : \Omega \to \mathbf{R}^n$*, with*

$$(A \in \mathscr{R}^n) \left(\mathbf{Prob}\,(X \in A) = \frac{1}{\sqrt{(2\pi)^n}\sqrt{\det \Sigma}} \int_A e^{-(x-c)^T \Sigma^{-1} (x-c)/2} d\mu \right)$$

where $\mu = PX^{-1}$ *for some* $\Sigma \succ 0 \in \mathbf{S}^n_{++}$ *and* $c \in \mathbf{R}^n$*, called* (n-dimensional) normal distribution*, and denoted by* $X \sim \mathcal{N}(c, \Sigma)$

– note that $\mathbf{E}\, X = c$ and covariance matrix is Σ

Lindeberg-Lévy theorem

– define $S_n = \sum^n X_i$

Theorem 8. [Lindeberg-Levy theorem] *for independent random variables, $\langle X_n \rangle$, having same distribution with expected value, c, and same variance, $\sigma^2 < \infty$, $(S_n - nc)/\sigma\sqrt{n}$ converges to standard normal distribution in distribution, $i.e.$,*

$$\frac{S_n - nc}{\sigma\sqrt{n}} \Rightarrow N$$

where N is standard normal distribution

– Theorem 8 implies

$$S_n/n \Rightarrow c$$

Limit theorems in $\mathbf{R}^n$

Theorem 9. [equivalent statements to weak convergence] *each of following statements are equivalent to weak convergence of measures, $\langle \mu_n \rangle$, to μ, on measurable space, $(\mathbf{R}^k, \mathscr{R}^k)$*

- $\lim \int f d\mu_n = \int f d\mu$ *for every bounded continuous f*
- $\limsup \mu_n(C) \leq \mu(C)$ *for every closed C*
- $\liminf \mu_n(G) \geq \mu(G)$ *for every open G*
- $\lim \mu_n(A) = \mu(A)$ *for every μ-continuity A*

Theorem 10. [convergence in distribution of random vector] *for random vectors, $\langle X_n \rangle$, and random vector, Y, of k-dimension, $X_n \Rightarrow Y$, $i.e.$, X_n converge to Y in distribution* if and only if

$$\left(\forall z \in \mathbf{R}^k\right) \left(z^T X_n \Rightarrow z^T Y\right)$$

Central limit theorem

– assume probability space, $(\Omega, \mathscr{F}, P)$ and define $\sum^n X_i = S_n$

Theorem 11. [central limit theorem] *for random variables, $\langle X_n \rangle$, having same distributions with $\mathbf{E}\, X_n = c \in \mathbf{R}^k$ and positive definite covariance matrix, $\Sigma \succ 0 \in \mathcal{S}_k$, $i.e.$, $\mathbf{E}(X_n - c)(X_n - c)^T = \Sigma$, where $\Sigma_{ii} < \infty$ (hence $\Sigma \prec M I_n$ for some $M \in \mathbf{R}_{++}$ due to Cauchy-Schwarz inequality),*

$$(S_n - nc)/\sqrt{n} \text{ converges in distribution to } Y$$

where $Y \sim \mathcal{N}(0, \Sigma)$

(proof can be found in Proof 1)

Convergence of random series

- for independent $\langle X_n \rangle$, probability of $\sum X_n$ converging is either 0 or 1
- below characterize two cases in terms of distributions of individual X_n – XXX: diagram

Theorem 12. [convergence with probability 1 for random series] *for independent $\langle X_n \rangle$ with $\mathbf{E}\, X_n = 0$ and $\mathbf{Var}\, X_n < \infty$*

$$\sum X_n \text{ converges with probability } 1$$

Theorem 13. [convergence conditions for random series] *for independent $\langle X_n \rangle$, $\sum X_n$ converges with probability 1* if and only if *they converges in probability*

– define trucated version of X_n by $X_n^{(c)}$, $i.e.$, $X_n I_{|X_n| \leq c}$

Theorem 14. [convergence conditions for truncated random series] *for independent* $\langle X_n \rangle$,

$$\sum X_n \text{ converge with probability } 1$$

$$\text{if all of } \sum \mathbf{Prob}\left(|X_n| > c\right), \sum \mathbf{E}(X_n^{(c)}), \sum \mathbf{Var}(X_n^{(c)}) \text{ converge for some } c > 0$$

Selected Proofs

Selected proofs

- **Proof 1.** *(Proof for "central limit theorem" on page 54)*
  Let $Z_n(t) = t^T(X_n - c)$ for $t \in \mathbf{R}^k$ and $Z(t) = t^T Y$. Then $\langle Z_n(t) \rangle$ are independent random variables having same distribution with $\mathbf{E}\, Z_n(t) = t^T(\mathbf{E}\, X_n - c) = 0$ and

  $$\mathbf{Var}\, Z_n(t) = \mathbf{E}\, Z_n(t)^2 = t^T\, \mathbf{E}(X_n - c)(X_n - c)^T t = t^T \Sigma t$$

  Then by Theorem 8 $\sum^n Z_i(t)/\sqrt{n t^T \Sigma t}$ converges in distribution to standard normal random variable. Because $\mathbf{E}\, Z(t) = 0$ and $\mathbf{Var}\, Z(t) = t^T\, \mathbf{E}\, Y Y^T z = t^T \Sigma t$, for $t \neq 0$, $Z(t)/\sqrt{t^T \Sigma t}$ is standard normal random variable. Therefore $\sum^n Z_i(t)/\sqrt{n t^T \Sigma t}$ converges in distribution to $Z/\sqrt{t^T \Sigma t}$ for every $t \neq 0$, thus, $\sum^n Z_i(t)/\sqrt{n} = t^T(\sum^n X_i - nc)/\sqrt{n}$ converges in distribution to $Z(t) = t^T Y$ for every $t \in \mathbf{R}$. Then Theorem 10 implies $(S_n - nc)/\sqrt{n}$ converges in distribution to Y. ∎

Index