ST5215 Advanced Statistical Theory, Lecture 4

HUANG Dongming

National University of Singapore

20 Aug 2020

Review

Last time

- Jensen's inequality, Cauchy-Schwarz inequality, Minkowski's inequality
- Characteristic function and moment generating function
- Independence
- Conditional expectation

Today

- Properties of conditional expectation
- Conditional distribution
- Statistical models

Recap: Dominated convergence theorem

Example (1.8, Interchange of differentiation and integration)

Let $(\Omega, \mathcal{F}, \nu)$ be a measure space and, for any fixed $\theta \in \mathcal{R}$, let $f(\omega, \theta)$ be a Borel function on Ω. Suppose that $\partial f(\omega, \theta)/\partial\theta$ exists a.e. for $\theta \in (a, b) \subset \mathcal{R}$ and that $|\partial f(\omega, \theta)/\partial\theta| \leq g(\omega)$ a.e., where g is an integrable function on Ω. Then for each $\theta \in (a, b), \partial f(\omega, \theta)/\partial\theta$ is integrable and, by Dominated convergence theorem,

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \int f(\omega, \theta) \, \mathrm{d}\nu = \int \frac{\partial f(\omega, \theta)}{\partial\theta} \, \mathrm{d}\nu$$

- LHS means: for any sequence of small numbers $\delta_n \to 0$, $\frac{1}{\delta_n} \left(\int f(\omega, \theta + \delta_n) \, \mathrm{d}\nu - \int f(\omega, \theta) \, \mathrm{d}\nu \right)$ converges to the same limit
- For given $\{\delta_n\}$, define $\varphi_n(\omega) = \frac{1}{\delta_n} \left(f(\omega, \theta + \delta_n) - f(\omega, \theta) \right)$. By mean value theorem and the condition, $|\varphi_n| \leq g(\omega)$ a.e.
- By DCT, $\lim \int \varphi_n \, \mathrm{d}\nu = \int \lim \varphi_n \, \mathrm{d}\nu$

Recap: Conditional expectation

Definition

- Let X be an integrable random variable on $(\Omega, \mathcal{F}, P)$.
- Let $\mathcal{A}$ be a sub-σ-field of $\mathcal{F}$.

The *conditional expectation* of X given $\mathcal{A}$, denoted by $\mathbb{E}(X \mid \mathcal{A})$, is a random variable satisfying the following two conditions:

1. $\mathbb{E}(X \mid \mathcal{A})$ is measurable from $(\Omega, \mathcal{A})$ to $(\mathcal{R}, \mathcal{B})$
2. $\int_C \mathbb{E}(X \mid \mathcal{A}) \, \mathrm{d}P = \int_C X \, \mathrm{d}P$ for any $C \in \mathcal{A}$

Such $\mathbb{E}(X \mid \mathcal{A})$ exists and is unique.

Proof

- W.L.O.G., assume $X \geq 0$; otherwise look at X_+ and X_- separately.
- Define $\lambda(C) = \int_C X \,\mathrm{d}P$ for any $C \in \mathcal{A}$.
- λ is a measure on $(\Omega, \mathcal{A})$ and $\lambda \ll P|_{\mathcal{A}}$
  - $P|_{\mathcal{A}}$ is the *restriction of the measure P on $\mathcal{A}$*, meaning that it has the same image of P but is now only define on $\mathcal{A}$ rather than Ω
- Then by Radon-Nikodym theorem, $\frac{\mathrm{d}\lambda}{\mathrm{d}P|_{\mathcal{A}}}$ exists and is unique, and it satisfies the definition of $\mathbb{E}(X \mid \mathcal{A})$.
- For general X, define $\mathbb{E}(X \mid \mathcal{A}) = \mathbb{E}(X_+ \mid \mathcal{A}) - \mathbb{E}(X_- \mid \mathcal{A})$.

Example

Let X be an integrable random variable on $(\Omega, \mathcal{F}, P)$. Let $A_1, A_2, \ldots$ be disjoint events such that $\cup A_i = \Omega$ and $P(A_i) > 0$ for all i, and let $a_1, a_2, \ldots$ be distinct real numbers. Define $Y = a_1 I_{A_1} + a_2 I_{A_2} + \cdots$. Then we have

$$\mathbb{E}(X \mid Y) = \sum_{i=1}^{\infty} \frac{\int_{A_i} X \, \mathrm{d}P}{P(A_i)} I_{A_i}$$

- $\sigma(Y) = \sigma(\{A_1, A_2, \ldots\})$
- RHS is measurable on $(\Omega, \sigma(Y))$
- For any $B \in \mathcal{B}$, $Y^{-1}(B) = \underset{i:a_i \in B}{\cup} A_i$.

$$\int_{Y^{-1}(B)} X \, \mathrm{d}P = \sum_{i:a_i \in B} \int_{A_i} X \, \mathrm{d}P$$

$$= \sum_{i=1}^{\infty} \frac{\int_{A_i} X \, \mathrm{d}P}{P(A_i)} \int 1_{A_i \cap Y^{-1}(B)} \, \mathrm{d}P = \int_{Y^{-1}(B)} \left[\sum_{i=1}^{\infty} \frac{\int_{A_i} X \, \mathrm{d}P}{P(A_i)} I_{A_i} \right] \mathrm{d}P$$

Example (Cont.)

Let X be an integrable random variable on $(\Omega, \mathcal{F}, P)$. Let $A_1, A_2, \ldots$ be disjoint events such that $\cup A_i = \Omega$ and $P(A_i) > 0$ for all i, and let $a_1, a_2, \ldots$ be distinct real numbers. Define $Y = a_1 I_{A_1} + a_2 I_{A_2} + \cdots$. Then we have

$$\mathbb{E}(X \mid Y) = \sum_{i=1}^{\infty} \frac{\int_{A_i} X \, \mathrm{d}P}{P(A_i)} I_{A_i}$$

- Define $h : \{a_i\} \mapsto \mathcal{R}$ by $h(a_i) = \frac{\int_{A_i} X \, \mathrm{d}P}{P(A_i)}$.
- $\mathbb{E}(X \mid Y)(\omega) = h(Y(\omega))$
- If $X = I_A$ where $A \in \mathcal{F}$, then $\mathbb{E}(X \mid Y) = \sum_{i=1}^{\infty} \frac{P(A_i \cap A)}{P(A_i)} I_{A_i}$, i.e., $\mathbb{E}(X \mid Y)(\omega) = P(A \mid A_i)$ if $\omega \in A_i$ (i.e., $Y(\omega) = a_i$)

Properties of conditional expectation

All r.v.s. are on the probability space $(\Omega, \mathcal{F}, P)$, and $\mathcal{G}$ is a sub-σ-field of $\mathcal{F}$.

- Linearity: $\mathbb{E}(aX + bY \mid \mathcal{G}) = a\mathbb{E}(X \mid \mathcal{G}) + b\mathbb{E}(X \mid \mathcal{G})$ a.s.
- If $X = c$ a.s. for a constant c, then $\mathbb{E}(X \mid \mathcal{G}) = c$ a.s.
- Monotonicity: if $X \leq Y$, then $\mathbb{E}(X \mid \mathcal{G}) \leq \mathbb{E}(Y \mid \mathcal{G})$ a.s.
- If $\mathbb{E}X^2 < \infty$, then $\{\mathbb{E}(X \mid \mathcal{G})\}^2 \leq \mathbb{E}(X^2 \mid \mathcal{G})$ a.s.
- (Fatou's lemma). If $X_n \geq 0$ for any n, then $\mathbb{E}\left(\liminf_n X_n \mid \mathcal{G}\right) \leq \liminf_n \mathbb{E}\left(X_n \mid \mathcal{G}\right)$ a.s.
- (Dominated convergence theorem). If $|X_n| \leq Y$ for any n and $X_n \to_{\text{a.s.}} X$, then $\mathbb{E}\left(X_n \mid \mathcal{G}\right) \to_{\text{a.s.}} \mathbb{E}(X \mid \mathcal{G})$
- all the integral-inequalities we saw before have conditional versions

- If $\mathcal{G} = \{\emptyset, \Omega\}$ (a trivial σ-field), then $\mathbb{E}(X \mid \mathcal{G}) = \mathbb{E}(X)$
- Tower property: if $\mathcal{H} \subset \mathcal{G}$ is a σ-field, (so that $\mathcal{H} \subset \mathcal{G} \subset \mathcal{F}$), then

$$\mathbb{E}(X \mid \mathcal{H}) = \mathbb{E}\{\mathbb{E}(X \mid \mathcal{G}) \mid \mathcal{H}\}\text{a.s.} \tag{1}$$

  - Let $\mathcal{H}$ be $\{\emptyset, \Omega\}$, then $\mathbb{E}(X) = \mathbb{E}\{\mathbb{E}(X \mid \mathcal{G})\}$.
- If $\sigma(Y) \subset \mathcal{G}$ (i.e., Y is $\mathcal{G}$-measurable) and $\mathbb{E}|XY| < \infty$, then $\mathbb{E}(XY \mid \mathcal{G}) = Y\mathbb{E}(X \mid \mathcal{G})$ a.s.
  - since $\sigma(Y) \subset \mathcal{G}$, information about Y is contained in $\mathcal{G}$, and thus, Y is kind of "known" given the information $\mathcal{G}$.
- If X and Y are independent and $\mathbb{E}|g(X, Y)| < \infty$ for a Borel function g, then $\mathbb{E}[g(X, Y) \mid Y = y] = \mathbb{E}[g(X, y)]$ a.s. P_Y
- If X and Y are independent, $\mathbb{E}(X \mid Y) = \mathbb{E}X$ a.s. P

Recap: Independence

Proposition (1.11 in JS)

Let X be a r.v. with $\mathbb{E}|X| < \infty$ and let Y_i be random k_i-vectors, $i = 1, 2$. Suppose that (X, Y_1) and Y_2 are independent. Then

1. $\mathbb{E}[X \mid (Y_1, Y_2)] = \mathbb{E}(X \mid Y_1)$ *a.s.*
2. $P(A \mid Y_1, Y_2) = P(A \mid Y_1)$ *a.s. for any* $A \in \sigma(X)$

- Suppose Y_1 is nonconstant. Given Y_1, X and Y_2 are conditionally independent iff (2) holds.
- Write "$(X, Y_1) \perp Y_2 \Rightarrow X \perp Y_2 \mid Y_1$"
- Find an example where Y_2 is independent of X and (1) fails to hold
  - Let $X \sim Unif\{-1, 1\}$, and $Y_1 \perp X$ and has the same distribution.
  - Let $Y_2 = XY_1$. Then $Y_2 \perp X$ but not independent of (X, Y_1)
  - $\mathbb{E}[X \mid (Y_1, Y_2)] = Y_1 Y_2$, but $\mathbb{E}[X \mid Y_1] = 0$

Existence of conditional distributions

Theorem

Suppose

- X *is a random n-vector on a probability space* $(\Omega, \mathcal{F}, P)$ *, and*
- Y *is measurable from* $(\Omega, \mathcal{F})$ *to* $(\Lambda, \mathcal{G})$*.*

Then there exists a function $P_{X|Y}(B \mid y)$ *on* $\mathcal{B}^n \times \Lambda$ *such that*

1. $P_{X|Y}(\cdot \mid y)$ *is a probability measure on* $(\mathcal{R}^n, \mathcal{B}^n)$ *for any fixed* $y \in \Lambda$*,*
2. $P_{X|Y}(B \mid y) = P[X \in B \mid Y = y]$ *a.s.* P_Y *for any fixed* $B \in \mathcal{B}^n$*.*

Furthermore, if $E|g(X, Y)| < \infty$ *with a Borel function* g*, then*

$$
\begin{aligned}
E[g(X, Y) \mid Y = y] &= E[g(X, y) \mid Y = y] \qquad (2) \\
&= \int_{\mathcal{R}^n} g(x, y) P_{X|Y}(\,\mathrm{d}x \mid y) \text{ a.s. } P_Y \qquad (3)
\end{aligned}
$$

For a fixed y, $P_{X|Y}(\cdot \mid y)$ is called the conditional distribution of X given $Y = y$, which is also denoted as $P_{X|Y=y}$.

Conditional p.d.f.

Theorem

Suppose

- X *is a random n-vector,* Y *is a random m-vector*
- (X, Y) *has a p.d.f.* $f(x, y)$ *w.r.t.* $\nu \times \lambda$ *(*ν *on* $\mathcal{B}^n$*,* λ *on* $\mathcal{B}^m$*, both* σ*-finite).*

Let $f_Y(y) = \int f(x, y) d\nu(x)$ *be the marginal p.d.f. of* Y *w.r.t.* λ*, and* $A = \{y \in \mathcal{R}^m : f_Y(y) > 0\}$.
Then for any fixed $y \in A$*, the p.d.f. of* $P_{X|Y=y}$ *w.r.t.* ν *is given by*

$$f_{X|Y}(x \mid y) = \frac{f(x, y)}{f_Y(y)}. \tag{4}$$

Furthermore, if $g(x, y)$ *is a Borel function on* $\mathcal{R}^{n+m}$ *and* $\mathbb{E}|g(X, Y)| < \infty$*, then*

$$\mathbb{E}[g(X, Y) \mid Y] = \int g(x, Y) f_{X|Y}(x \mid Y) \, \mathrm{d}\nu(x), \textit{a.s.} \tag{5}$$

- Let $h(y) = \frac{\int g(x,y) f(x,y) \, \mathrm{d}\nu(x)}{f_Y(y)}$. By Fubini's theorem, $h(y)$ is Borel.
- For any $B \in \mathcal{B}^m$, (W.L.O.G. $B \subset A$)

$$
\begin{aligned}
\int_{Y^{-1}(B)} h(Y) \, \mathrm{d}P &= \int_B h(y) \, \mathrm{d}P_Y \\
\text{(Def. of } h) \qquad &= \int_B \frac{\int g(x,y) f(x,y) \, \mathrm{d}\nu(x)}{f_Y(y)} \, \mathrm{d}P_Y \\
\text{(p.d.f. of } Y \text{ w.r.t. } \lambda) \qquad &= \int_B \frac{\int g(x,y) f(x,y) \, \mathrm{d}\nu(x)}{f_Y(y)} f_Y(y) \, \mathrm{d}\lambda(y) \\
&= \int_B \left(\int g(x,y) f(x,y) \, \mathrm{d}\nu(x) \right) \, \mathrm{d}\lambda(y) \\
\text{(Fubini's theorem)} \qquad &= \int_{\mathcal{R}^n \times B} g(x,y) f(x,y) \, \mathrm{d}(\nu \times \lambda)(x,y) \\
\text{(p.d.f. of } (X,Y) \text{ w.r.t. } \nu \times \lambda) &= \int_{\mathcal{R}^n \times B} g(x,y) \, \mathrm{d}P_{(X,Y)} \\
\text{(Change of variable)} \qquad &= \int_{Y^{-1}(B)} g(X,Y) \, \mathrm{d}P
\end{aligned}
$$

Building joint distributions

Theorem

Let $(\Lambda, \mathcal{G}, P_0)$ be a probability space. Suppose that Q is a function from $\mathcal{B}^n \times \Lambda$ to $\mathcal{R}$ and satisfies

1. *$Q(\cdot, y)$ is a probability measure on $(\mathcal{R}^n, \mathcal{B}^n)$ for any $y \in \Lambda$,*
2. *$Q(B, \cdot)$ is $\mathcal{G}$-measurable for any $B \in \mathcal{B}^n$*

Then there is a unique probability measure P on $(\mathcal{R}^n \times \Lambda, \sigma(\mathcal{B}^n \times \mathcal{G}))$ such that, for $B \in \mathcal{B}^n$ and $C \in \mathcal{G}$

$$P(B \times C) = \int_C Q(B, y) dP_0(y)$$

We can construct a joint distribution in the product space, if given

1. a marginal distribution of Y on a space,
2. a collection of (regular) conditional distributions on another space.

This provides a way of generating dependent random variables.

Statistical problems

- A typical statistical problem
  - One or a series of random experiments is performed
  - Some data are generated and collected from the experiments
  - Extract information from the data
  - Interpret results and draw conclusions

Example (Measurement problems)

Suppose we want to measure an unknown quantity θ, e.g., weight of some object.

- n measurements $x_1, \ldots, x_n$ are taken in an experiment of measuring θ.
- data are $(x_1, \ldots, x_n)$
- information to extract: some estimator for θ
- draw conclusion: what is the possible range of θ (confidence interval)?

Setup

- We do not consider the problems of planning experiments and collecting data.
- A **population** is a probability space $(\Omega, \mathcal{F}, P)$. For simplicity, we refer to P as the population
- A **sample** is a random element defined on the probability space. The data set is a realization of the sample.
- The size of the data set is called the sample size.

- A population P is *known* iff $P(A)$ is a known value for every event $A \in \mathcal{F}$.
- P is at least partially unknown and we want to deduce some properties of P based on the data.

Example (Measurement problems)

Suppose we want to measure an unknown quantity θ, e.g., weight of some object.

- n measurements $x_1, \ldots, x_n$ are taken in an experiment of measuring θ.
- if no measurement error, then $x_1 = \cdots = x_n = \theta$
- otherwise, x_i are not the same due to measurement errors
- the data set $(x_1, \ldots, x_n)$ is viewed as an outcome of the experiment
- sample size is n
- the sample space is $\Omega = \mathcal{R}^n$, $\mathcal{F} = \mathcal{B}^n$, and P is a probability measure on $\mathcal{R}^n$
- the random element $X = (X_1, \ldots, X_n)$ is a random n-vector defined on $\mathcal{R}^n$, i.e., $X : \mathcal{R}^n \to \mathcal{R}^n$

In applications, it is often reasonable to assume that distributions come from a suitable class of distributions.

- In physics, one requires a mathematical model to describe what are observed
  - $F = ma$, for example
- Models are simplifications or approximation of reality
- Good models approximate the reality well
  - Newton's physics is good for low-speed motion
  - For high-speed motion, we needs special relativity or even general relativity
- In statistics, we use models to approximate the mechanism that generates the observed data
  - "all models are wrong, but some are useful." – George Box.

Statistical models

- A *statistical model* is a set of assumptions on the population P. Mathematically, a statistical model is often expressed as

$$P \in \mathcal{P} = \{Q : Q \textit{ satisfies some conditions}\} \tag{6}$$

Definition (Parametric family and Parametric models)

- A set of probability measures P_θ on $(\Omega, \mathcal{F})$ indexed by a *parameter* $\theta \in \Theta$ is said to be a *parametric family* iff $\Theta \subset \mathcal{R}^d$ for some fixed positive integer d and each P_θ is a known probability measure when θ is known.
- The set Θ is called the *parameter space* and d is called its *dimension*.
- *A parametric model* refers to the assumption that the population P is in a parametric family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$

A family of probability measures is said to be *nonparametric* if it is not parametric. A *nonparametric model* refers to the assumption that the population P is in a given nonparametric family.

Example (Measurement problems)

A statistical model here is a set of *joint* distribution of $X_1, \ldots, X_n$

- We begin with assuming $X_1, \ldots, X_n$ independent and identically distributed (i.i.d., or IID), then $P = P_0^n$, where P_0 is a probability on $(\mathcal{R}, \mathcal{B})$
  - In this case, the product probability measure is determined by P_0, the marginal distribution of X_i.
  - With IID assumption, we usually states the model in terms of P_0.
- We further assume $X_i \sim N(\theta, \sigma^2)$, so $P_0 = N(\theta, \sigma^2)$ with IID assumption.
  - P_0 is partially unknown, since θ and σ^2 are unknown.
  - We want to deduce the values of θ and σ^2 based on the available sample.
  - A statistical model: $P_0 \in \mathcal{P}_1 = \left\{ N\left(\theta, \sigma^2\right) : \theta \in \mathcal{R}, \sigma^2 > 0 \right\}$
- Since we know the weight of an object is positive, it makes more sense to require $\theta > 0$. We can consider a smaller model like $P_0 \in \mathcal{P}_2 = \left\{ N\left(\theta, \sigma^2\right) : \theta > 0, \sigma^2 > 0 \right\}$

Some terms

- A parametric family $\{P_\theta : \theta \in \Theta\}$ is said to be *identifiable* if and only if $\theta_1 \neq \theta_2$ and $\theta_1, \theta_2 \in \Theta$ imply $P_{\theta_1} \neq P_{\theta_2}$
- Let $\mathcal{P}$ be a family of populations and ν a σ-finite measure on $(\Omega, \mathcal{F})$. If $P \ll \nu$ for all $P \in \mathcal{P}$, then we say $\mathcal{P}$ is dominated by ν,
  - $\mathcal{P}$ can be identified by the family of densities $\{\frac{\mathrm{d}P}{\mathrm{d}\nu} : P \in \mathcal{P}\}$;
  - In statistics, ν is often the Lebesgue measure (for continuous random variables) or the counting measure (for discrete random variables)
- In a given problem, a parametric model is not useful if the dimension of Θ is very large compared with the sample size.

Tutorial

1. (Generalization of Hölder's inequality).
For $0 < p < 1$ and $q = -p/(1-p)$

$$E|XY| \geqslant (E|X|^p)^{1/p} (E|Y|^q)^{1/q}$$

2. (Generalization of Minkowski's inequality).
$\left(E\left(\sum_{j=1}^n |X_j|\right)^r\right)^{1/r} > \sum_{j=1}^n (E|X_j|^r)^{1/r} \quad$ for $0 < r < 1$

3. Let Y be measurable from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$ and Z a function from $(\Omega, \mathcal{F})$ to $\mathcal{R}^k$. If Z is Borel on $(\Omega, \sigma(Y))$, then there is a Borel function h from $(\Lambda, \mathcal{G})$ such that $Z = h \circ Y$

4. Let ϕ_X be a ch.f. of X. Show that $|\phi_X| \leq 1$, and uniformly continuous.

5. Find the ch.f. and m.g.f. for the Cauchy distribution (i.e., P_X has p.d.f. $f(x) = (\pi(1 + x^2))^{-1}$

6. If X_i has the Cauchy distribution $C(0, 1), i = 1, \ldots, k$, then Y/k has the same distribution as X_1.

Ex 1

(Generalization of Hölder's inequality).
For $0 < p < 1$ and $q = -p/(1-p)$

$$E|XY| \geqslant (E|X|^p)^{1/p} (E|Y|^q)^{1/q}$$

Proof: WLOG, assume $E|Y|^q > 0$.

- Put $\tilde{X} = |XY|^p$, $\tilde{Y} = |Y|^{-p}$.
- Let $p' = 1/p$, $q' = 1/(1-p)$. Then $1/p' + 1/q' = 1$.
- Apply the Hölder inequality to $(\tilde{X}, \tilde{Y}, p', q')$,

$$\begin{aligned} E|X|^p = E\tilde{X}\tilde{Y} &\leqslant \left(E\tilde{X}^{p'}\right)^{1/p'} \left(E\tilde{Y}^{q'}\right)^{1/q'} \\ &= \left(E\tilde{X}^{1/p}\right)^{p} \left(E\tilde{Y}^{1/(1-p)}\right)^{1-p} \\ &= (E|XY|)^p (E|Y|^q)^{1-p}. \end{aligned}$$

Ex 2

(Generalization of Minkowski's inequality).

$\left(E\left(\sum_{j=1}^{n}|X_j|\right)^r\right)^{1/r} > \sum_{j=1}^{n}\left(E|X_j|^r\right)^{1/r}$ for $0 < r < 1$

Proof: Suppose $n = 2$ and we write $X = X_1$, $Y = X_2$. WLOG, assume $E(|X| + |Y|)^r] > 0$.

- $E(|X| + |Y|)^r = E(|X| + |Y|)^{r-1}(|X| + |Y|) =$ $E\left[(|X| + |Y|)^{r-1}|X|\right] + E\left[(|X| + |Y|)^{r-1}|Y|\right]$
- Apply Ex 1 to $(|X|, (|X| + |Y|)^{r-1}, r, r/(r-1))$, we have

$$E\left[(|X| + |Y|)^{r-1}|X|\right] \geq (E|X|^r)^{1/r}[E(|X| + |Y|)^r]^{(r-1)/r}$$

- So $E(|X| + |Y|)^r \geq \left[(E|X|^r)^{1/r} + (E|Y|^r)^{1/r}\right][E(|X| + |Y|)^r]^{(r-1)/r}$
- Divide both side by $[E(|X| + |Y|)^r]^{(r-1)/r}$, we have

$$[E(|X| + |Y|)^r]^{1/r} \geq (E|X|^r)^{1/r} + (E|Y|^r)^{1/r} \tag{7}$$

For general n, use induction and the last inequality to $\sum_{i=1}^{n-1}|X_i|$ and $|X_n|$

Ex 3

Let Y be measurable from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$ and Z a function from $(\Omega, \mathcal{F})$ to $\mathcal{R}^k$. If Z is Borel on $(\Omega, \sigma(Y))$, then there is a Borel function h from $(\Lambda, \mathcal{G})$ such that $Z = h \circ Y$

Proof: First, suppose Z is a simple function on $(\Omega, \sigma(Y))$, i.e., $Z = \sum_{i=1}^n z_i I_{A_i}$, where c_i's are real numbers, and A_i's are disjoint and in $Y^{-1}\mathcal{G}$.

- We can assume $A_i = Y^{-1} C_i$ where $C_i \in \mathcal{G}$. Note that C_i's are not necessarily disjoint (if $\emptyset \neq C_i \cap C_j \subset Y(\Omega)^c$)
- Let $B_1 = C_1$ and $B_i = C_i \setminus (\cup_{k=1}^{i-1} C_k)$, $i \geq 2$. Then B_i's are disjoint and in $\mathcal{G}$
- We can check that $A_i = Y^{-1} B_i$.
- Define $h = \sum_{i=1}^n z_i I_{B_i}$. It is a simple function on $(\Lambda, \mathcal{G})$ and for any $\omega \in \Omega$,

$$Z(\omega) = \sum_{i=1}^n z_i I_{A_i}(\omega) = \sum_{i=1}^n z_i I_{Y^{-1}B_i}(\omega) = \sum_{i=1}^n z_i I_{B_i}(Y(\omega)) \quad (8)$$

$$= h(Y(\omega)) \quad (9)$$

Ex 3 (Cont.)

Let Y be measurable from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$ and Z a function from $(\Omega, \mathcal{F})$ to $\mathcal{R}^k$. If Z is Borel on $(\Omega, \sigma(Y))$, then there is a Borel function h from $(\Lambda, \mathcal{G})$ such that $Z = h \circ Y$

Proof: Second, suppose Z is a general Borel function on $(\Omega, \sigma(Y))$.

- We can find a sequence of simple functions ϕ_n on $(\Omega, \sigma(Y))$ such that $\lim \phi_n = Z$.
- The first part shows that we can find a sequence of simple functions h_n from $(\Lambda, \mathcal{G})$ such that $\phi_n = h_n \circ Y$.
- Let $A = \{y \in \Lambda : \lim h_n(y) \text{ exists }\}$.
- Define $h(y) = \lim h_n(y)$ for $y \in A$ and $h(y) = 0$ for $y \notin A$. By Proposition 1.4 in JS, h is $\mathcal{G}$-measurable.
- For any $\omega \in \Omega$, we have $Z(\omega) = \lim \phi_n(\omega) = \lim h_n(Y(\omega))$, which implies that $Y(\omega) \in A$ and RHS $= h(Y(\omega))$.

Ex 4

Let ϕ_X be a ch.f. of X. Show that $|\phi_X| \leq 1$, and uniformly continuous.
Proof: Part 1: By Cauchy-Schwartz inequality,
$(E\cos(t^\top X))^2 \leq E\cos^2(t^\top X)$ and $(E\sin(t^\top X))^2 \leq E\sin^2(t^\top X)$, so

$$\begin{aligned}
|\phi_X(t)|^2 &= (E\cos(t^\top X))^2 + (E\sin(t^\top X))^2 \\
&\leq E\cos^2(t^\top X) + E\sin^2(t^\top X) = 1
\end{aligned}$$

Ex 4 (Cont.)

Let ϕ_X be a ch.f. of X. Show that $|\phi_X| \leq 1$, and uniformly continuous.
Proof: Part 2:

- We need a result: for any $x \in \mathcal{R}$, $|e^{ix} - 1| \leq \min(|x|, 2)$
- For any $\epsilon > 0$, choose $M > 0$ s.t. $P(\|X\| > M) < \epsilon/4$.
- For any $t, s \in \mathcal{R}^d$, s.t. $\|t - s\| \leq \epsilon/(2M)$, we have

$$
\begin{aligned}
|\phi_X(t) - \phi_X(s)| &= |E[e^{is^\top X}\left(e^{i(t-s)^\top X} - 1\right)]| \\
&\leq E\left|e^{i(t-s)^\top X} - 1\right| \\
&\leq 2P(|X| > M) + E\left(I_{\{\|X\| \leq M\}}\left|e^{i(t-s)^\top X} - 1\right|\right) \\
&< \epsilon/2 + E\left(I_{\{\|X\| \leq M\}}\|t - s\|\|X\|\right) \\
&\leq \epsilon.
\end{aligned}
$$

Ex 5

Find the ch.f. and m.g.f. for the Cauchy distribution (i.e., P_X has p.d.f. $f(x) = (\pi(1+x^2))^{-1}$
Proof: We need a theorem

Theorem

Let $(\Omega, \mathcal{F}, \nu)$ be a measure space. Let A_k be an increasing sequence of measurable sets, whose limit is A. Let f be a Borel function. If for each k, fI_{A_k} is integrable, and $\lim \int_{A_k} |f| \, \mathrm{d}\nu < \infty$, then fI_A is integrable and

$$\lim \int_{A_k} f \, \mathrm{d}\nu = \int_A f \, \mathrm{d}\nu. \tag{10}$$

This result, together with the connection of Riemann integral and Lebesgue integral, allows us to compute the Lebesgue integral as we did in undergraduate calculus.

Ex 5 (Cont.)

Find the ch.f. and m.g.f. for the Cauchy distribution (i.e., P_X has p.d.f. $f(x) = (\pi(1+x^2))^{-1}$

Proof:

- For any $t \in \mathcal{R}$, $sin(tx)/(1+x^2)$ is an odd function of x, so for X being a Cauchy r.v.,

$$\varphi_X(t) = \int_{\mathcal{R}} \frac{\cos(tx)}{\pi(1+x^2)} \, \mathrm{d}m$$

- By the theorem and the fact that the Riemann integral $\int_{-n}^{n} \frac{|\cos(tx)|}{\pi(1+x^2)} \, \mathrm{d}x$ converges, we have

$$\int_{\mathcal{R}} \frac{\cos(tx)}{\pi(1+x^2)} \, \mathrm{d}m = \int_{-\infty}^{\infty} \frac{\cos(tx)}{\pi(1+x^2)} \, \mathrm{d}x = e^{-|t|}. \tag{11}$$

Ex 6

If X_i has the Cauchy distribution $C(0,1), i = 1, \ldots, k$, then Y/k has the same distribution as X_1.

Proof: We need to find the ch.f. for Y: For any $t \in \mathcal{R}$,

$$E[\exp(itY)] = E[\exp(it\frac{\sum_{j=1}^{k} X_j}{k})] \tag{12}$$

$$= E[\exp(\sum_{j=1}^{k} i\frac{t}{k}X_j)] \tag{13}$$

$$= E[\prod_{j=1}^{k} \exp(i\frac{t}{k}X_j)] \tag{14}$$

$$= \prod_{j=1}^{k} E[\exp(i\frac{t}{k}X_j)] \tag{15}$$

$$= \prod_{j=1}^{k} \exp(-|\frac{t}{k}|) = \exp(-|t|). \tag{16}$$