ECON 501
概率与数理基础自足讲义

Probability, Asymptotics, Estimation, and Testing

Penn State Econometrics Qualifier · Student Study Edition

Rui Zhou

Department of Economics, Penn State University

Public edition · August 2026

中文讲解 · English formal statements · Unofficial student resource

使用说明 · About This Volume

公开版说明 · Public release note

本卷从《经济计量学博士资格考试自足讲义》中拆出 **ECON 501** 部分，覆盖概率测度、分布与矩、条件期望与多元正态、收敛与渐近理论、估计与极大似然，以及经典假设检验。

它是学生整理的非官方学习资料，不代表 Penn State Department of Economics 或任课教师。公式和证明已经过系统复核，但仍可能有疏漏；使用时应以课程讲义、教材与系里发布的原始试卷为准。公开版不包含原讲义中尚未达到同等质量标准的 ECON 510 章节。

建议先读每节的直觉导入，再遮住结论独立复现证明。资格考试往往按中间步骤给分，因此应练习把假设、所用定理、标准化以及极限分布完整写出，而不只背最终公式。

目录

第一部分

概率与数理基础 · Probability & Mathematical Foundations

Chapter 1　测度论概率基础（Measure-Theoretic Probability）

导读 · Roadmap（先读这里，不含公式）

博士计量的第一块地基，是把“概率”从直觉抬升到测度论的语言。为什么非要这么抽象？因为一旦样本空间是连续的（比如 $[0,1]$），你无法给每一个子集都自洽地赋一个概率——存在“不可测集”。出路是：只对一个足够好的集合族（σ-域）谈概率。于是“概率”正式定义为三元组 $(\Omega, \mathcal{F}, \mathrm{P})$ 上满足三条公理的测度，“随机变量”正式定义为一个可测函数。

资格考几乎**年年**在这一章取一道题，最爱考三件事：（1）给你一个古怪的函数，问它是不是随机变量（即是否可测）；（2）“标准机器”——把一个关于 $\mathbb{E}(g(X))$ 的恒等式从指示函数一路推到一般可积函数（换元公式 $\mathbb{E}(g(X)) = \int g\, dP_X$、由密度定义的测度）；（3）用单调收敛 / 控制收敛定理（MCT/DCT）交换极限与积分。这三件事本章逐一拆解，并配上 2025 年资格考原题。

1.1　σ-域与 Borel 域

[优先级 ●●○ 中 Med]　资格考足迹：2022, 2024, 2025

我们先固定一个**样本空间** Ω（所有可能结果的集合）。我们想谈概率的那些子集，叫**事件**。事件的全体必须对“取补”和“可数并”封闭——否则连“A 不发生”或“A_1 或 A_2 或 … 发生”都谈不了。这样的集合族就是 σ-域。

Definition 1.1: σ-域 (σ-field / σ-algebra)

A collection $\mathcal{F}$ of subsets of Ω is a *σ-field* if

(i) $\Omega \in \mathcal{F}$;

(ii) $A \in \mathcal{F} \implies A^c \in \mathcal{F}$ (closed under complements);

(iii) $A_1, A_2, \cdots \in \mathcal{F} \implies \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$ (closed under countable unions).

The pair $(\Omega, \mathcal{F})$ is a *measurable space*; members of $\mathcal{F}$ are *measurable sets* (events).

Remark (由公理立刻能推出的几条).

> 结合 (i)(ii) 得 $\emptyset = \Omega^c \in \mathcal{F}$；结合 (ii)(iii) 与 De Morgan 得对可数交封闭：$\bigcap_i A_i = \left(\bigcup_i A_i^c\right)^c \in \mathcal{F}$。所以 σ-域对“补、可数并、可数交”三种操作全封闭。

给定 Ω 上任意一族子集 $\mathcal{C}$，包含 $\mathcal{C}$ 的最小 σ-域记作 $\sigma(\mathcal{C})$，称为由 $\mathcal{C}$ **生成**的 σ-域（它等于所有包含 $\mathcal{C}$ 的 σ-域之交——交一族 σ-域仍是 σ-域）。最重要的例子是实数上的 Borel 域。

Definition 1.2: Borel σ-field

The *Borel σ-field* on $\mathbb{R}^d$, written $\mathcal{B}(\mathbb{R}^d)$ (or just $\mathcal{B}$), is the σ-field generated by the open sets:

$$\mathcal{B}(\mathbb{R}^d) := \sigma\big(\{\text{open subsets of } \mathbb{R}^d\}\big).$$

Its members are *Borel sets*. Equivalently $\mathcal{B}(\mathbb{R}) = \sigma(\{(-\infty, x] : x \in \mathbb{R}\})$.

直觉：Borel 集是“一切你能用可数次并、交、补从区间造出来的集合”——开集、闭集、单点、可数集、区间……都是 Borel 集。要造一个非 Borel 集得很费劲。考场上但凡碰到“$\{X \le x\}$、$\{X \in B\}$”这类事件，默认 B 是 Borel 集。

Fact 1.3: 有理数集是 Borel 集，且 Lebesgue 测度为零

$\mathbb{Q} \cap [0,1]$ 是可数个单点之并，每个单点 $\{q\} = \bigcap_n (q - \frac{1}{n}, q + \frac{1}{n})$ 是 Borel 集，故 $\mathbb{Q} \cap [0,1] \in \mathcal{B}$。又 $\lambda(\{q\}) = 0$，由可数可加性 $\lambda(\mathbb{Q} \cap [0,1]) = 0$，从而无理数集 $[0,1] \setminus \mathbb{Q}$ 的 Lebesgue 测度为 1。这条事实是 *2025* 年资格考第 *6* 题的钥匙，见本章 *1.3.1* 小节。

1.2 测度与概率空间

[优先级 ●●○ 中 Med] 资格考足迹：2018, 2024

Definition 1.4: 测度 Measure

A *measure* on $(\Omega, \mathcal{F})$ is a function $\mu : \mathcal{F} \to [0, \infty]$ with

(i) $\mu(\emptyset) = 0$;

(ii) (countable additivity) for pairwise disjoint $A_1, A_2, \dots \in \mathcal{F}$, $\mu\big(\bigcup_i A_i\big) = \sum_i \mu(A_i)$.

A *probability measure* additionally satisfies $\mu(\Omega) = 1$; we then write it $\mathrm{P}(\cdot)$ and call $(\Omega, \mathcal{F}, \mathrm{P})$ a *probability space*.

测度就是“给集合量大小”的统一语言：Lebesgue 测度 λ 量长度/面积/体积（$\lambda([a,b]) = b - a$）；计数测度量元素个数；概率测度量“可能性”，并多一条归一化

$\mathrm{P}(\Omega) = 1$。注意值域含 $+\infty$（Lebesgue 测度下 $\lambda(\mathbb{R}) = \infty$），但概率测度恒在 $[0, 1]$。

由两条公理可推出一整套日常性质，证明都是把集合拆成不交并再用可数可加：

Proposition 1.5: 概率测度的基本性质

$\mathrm{P}(A^c) = 1 - \mathrm{P}(A)$；单调性 $A \subseteq B \Rightarrow \mathrm{P}(A) \leq \mathrm{P}(B)$；容斥 $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$；以及连续性 $A_n \uparrow A \Rightarrow \mathrm{P}(A_n) \to \mathrm{P}(A)$。

Remark (为什么是可数可加，而不是有限或不可数).

有限可加太弱（推不出上面的“连续性”，也对付不了“无穷次抛硬币”这类极限事件）；不可数可加又太强会自相矛盾——例如若对 $[0, \frac{1}{2}]$ 中每个单点都要 λ-可加，则 $\lambda([0, \frac{1}{2}]) = \sum_x \lambda(\{x\}) = \sum_x 0 = 0 \neq \frac{1}{2}$。可数可加是恰到好处的折中。

1.3 随机变量 = 可测函数

[优先级 ●●● 高 High] 资格考足迹：2017, 2019, 2022, 2024, 2025

这是全章**最常考**的一节。一个“随机变量”不是随便一个 $\Omega \to \mathbb{R}$ 的函数，而是要保证“X 落在某 Borel 集里”这件事是个事件（属于 $\mathcal{F}$），否则没法谈它的概率。

Definition 1.6: 随机变量 Random variable (measurable function)

Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space. A function $X : \Omega \to \mathbb{R}^d$ is a *random variable* (a *measurable* function) if

$$X^{-1}(B) := \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F} \qquad \text{for every } B \in \mathcal{B}(\mathbb{R}^d).$$

Theorem 1.7: 只需对生成元验证（可测性判据）

若 $\mathcal{B} = \sigma(\mathcal{C})$，则 X 可测当且仅当 $X^{-1}(C) \in \mathcal{F}$ 对一切 $C \in \mathcal{C}$ 成立。特别地，$X : \Omega \to \mathbb{R}$ 可测 $\iff \{X \leq x\} \in \mathcal{F}$ 对一切 $x \in \mathbb{R}$ 成立。

这条判据是干活的工具：要证一个函数可测，不必对所有 Borel 集 B 验证 $X^{-1}(B) \in \mathcal{F}$（那是无穷多个），只需对生成元（如所有 $(-\infty, x]$）验证即可，因为“原像保持并、交、补”，可测集族 $\{B : X^{-1}(B) \in \mathcal{F}\}$ 本身是 σ-域，包含生成元就包含整个 $\mathcal{B}$。

Definition 1.8: 诱导分布 / 推前测度 Induced (push-forward) distribution

A random variable X induces a probability measure P_X on $(\mathbb{R}^d, \mathcal{B})$ by

$$P_X(B) := \mathrm{P}\big(X^{-1}(B)\big) = \mathrm{P}(X \in B), \qquad B \in \mathcal{B}.$$

P_X is the *distribution* (law) of X. The CDF is $F_X(x) = P_X((-\infty, x]) =$

$\mathrm{P}(X \le x)$.

1.3.1 Worked example：2025 资格考 Part 501 第 6 题

原题 (Comprehensive Exam 2025, Pinkse, Q6)

Consider $(\Omega, \mathcal{B}, \lambda)$ with $\Omega = [0,1]$, $\mathcal{B}$ the Borel field, and λ Lebesgue measure. Define $\mathbf{x} : \Omega \to \mathbb{R}$ by

$$\mathbf{x}(\omega) = \begin{cases} \sqrt{\omega}, & \omega \text{ rational}, \\ 1 - \omega^2, & \omega \text{ irrational}. \end{cases}$$

Is $\mathbf{x}$ a random variable? Show that it is (or isn't).

Solution.

结论：是。由判据（Theorem 1.7）只需说明 $\{\omega : \mathbf{x}(\omega) \le t\} \in \mathcal{B}$ 对每个 $t \in \mathbb{R}$ 成立。把定义域按有理/无理拆开：

$$\{\mathbf{x} \le t\} \;=\; \underbrace{\big(\mathbb{Q} \cap [0,1] \cap \{\omega : \sqrt{\omega} \le t\}\big)}_{=:A_t} \;\cup\; \underbrace{\big((\,[0,1] \setminus \mathbb{Q}) \cap \{\omega : 1 - \omega^2 \le t\}\big)}_{=:B_t}.$$

A_t **是 Borel 集**：$\{\omega : \sqrt{\omega} \le t\}$ 是一个区间（$\emptyset$、$[0, t^2]$ 之类），与 Borel 集 $\mathbb{Q} \cap [0,1]$ 取交仍是 Borel 集（事实上是可数集，故 Borel）。B_t **是 Borel 集**：$g(\omega) = 1 - \omega^2$ 连续，故 $\{1 - \omega^2 \le t\}$ 是闭集（Borel）；与 Borel 集 $[0,1] \setminus \mathbb{Q}$ 取交仍 Borel。两个 Borel 集之并是 Borel 集，故 $\{\mathbf{x} \le t\} \in \mathcal{B}$。证毕。■

Remark (考点提炼).

> 这类题的"机器"永远一样：(1) 用 $\{X \le t\}$ 判据而非逐 Borel 集；(2) 按定义的分段把事件拆成几块；(3) 每块都是"连续函数的水平集（Borel）"与"可数集 $\mathbb{Q}$（Borel）"的交并，因而 Borel。关键事实：$\mathbb{Q}$ 可数 $\Rightarrow$ Borel。许多同学误以为"在有理点改了值"就破坏可测性——恰恰相反，因为 $\mathbb{Q}$ 可测，分段定义照样可测。顺带一提，由于 $\lambda(\mathbb{Q} \cap [0,1]) = 0$，$\mathbf{x}$ 与 $1 - \omega^2$ 几乎处处相等，故 $\mathbb{E}(\mathbf{x}) = \int_0^1 (1 - \omega^2)\, d\omega = \frac{2}{3}$。

1.4 积分与"标准机器"：换元公式

[优先级 ●●● 高 High] 资格考足迹：2017, 2018, 2019, 2024

期望就是 Lebesgue 积分 $\mathbb{E}(g(X)) = \int_\Omega g(X(\omega))\, d\mathrm{P}(\omega)$。资格考反复考的一个证明套路是**"标准机器" (standard machine)**：要证某个关于积分的恒等式对一切可

积 g 成立，按四步推进——先对指示函数，再线性延拓到简单函数，再用 MCT 上升到非负可测函数，最后拆 $g = g^+ - g^-$ 到一般可积函数。

Theorem 1.9: 换元 / 推前公式 (Law of the unconscious statistician)

[REPRODUCE —memorize this proof]

Let $X : \Omega \to \mathbb{R}^d$ have distribution P_X, and let $g : \mathbb{R}^d \to \mathbb{R}$ be measurable with $\mathbb{E}(|g(X)|) < \infty$. Then

$$\mathbb{E}(g(X)) \;=\; \int_\Omega g(X(\omega))\, d\mathrm{P}(\omega) \;=\; \int_{\mathbb{R}^d} g(x)\, dP_X(x).$$

Proof for Theorem

Step 1 (indicators). For $g = \mathbb{1}_B$, $B \in \mathcal{B}$: $\int_\Omega \mathbb{1}_B(X)\, d\mathrm{P} = \mathrm{P}(X \in B) = P_X(B) = \int \mathbb{1}_B\, dP_X$ by the very definition of P_X.

Step 2 (simple functions). For $g = \sum_{k=1}^m c_k \mathbb{1}_{B_k}$, both sides are linear, so equality follows from Step 1.

Step 3 (nonnegative g). Take simple $g_m \uparrow g$ pointwise (always possible for measurable $g \geq 0$). By the Monotone Convergence Theorem applied on both sides, $\int g_m(X)\, d\mathrm{P} \to \int g(X)\, d\mathrm{P}$ and $\int g_m\, dP_X \to \int g\, dP_X$; Step 2 gives equality at each m, hence in the limit.

Step 4 (general g). Write $g = g^+ - g^-$ with $g^\pm \geq 0$. Integrability $\mathbb{E}(|g(X)|) < \infty$ makes both pieces finite; apply Step 3 to each and subtract. ■

Remark (为什么这套机器值得专门背).

"indicators → simple → nonnegative → general"这**四步**是测度论里最通用的证明骨架，资格考至少考过三种外衣：(a) 换元公式（上）；(b) 由密度定义的测度 $\nu(A) = \int_A f\, d\mu$ 确为测度，且 $\int g\, d\nu = \int gf\, d\mu$（2018 Q3、2024 Q1，见下）；(c) 条件期望的各种恒等式。**记住这四步，半道证明分就到手了。**

Theorem 1.10: 由密度定义的测度 (2018 Q3 / 2024 Q1)

[STRUCTURE —know the steps, fudge details]

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and $f : \Omega \to [0, \infty]$ measurable. Define $\nu(A) := \int_A f\, d\mu$ for $A \in \mathcal{F}$. Then (a) ν is a measure on $(\Omega, \mathcal{F})$, and (b) for any measurable $g \geq 0$, $\int g\, d\nu = \int gf\, d\mu$. We call $f = \dfrac{d\nu}{d\mu}$ the *Radon–Nikodym derivative* of ν w.r.t. μ.

Proof for Theorem

(a) $\nu(\emptyset) = \int_\emptyset f\,d\mu = 0$ and $\nu \geq 0$. Countable additivity: for disjoint A_i, $\nu(\bigcup_i A_i) = \int \mathbb{1}_{\bigcup A_i} f\,d\mu = \int \left(\sum_i \mathbb{1}_{A_i}\right) f\,d\mu = \sum_i \int \mathbb{1}_{A_i} f\,d\mu = \sum_i \nu(A_i)$, where swapping $\sum$ and $\int$ is the Monotone Convergence Theorem applied to the nonnegative partial sums. **(b)** Run the standard machine of Theorem 1.9: it holds for $g = \mathbb{1}_A$ (that is the definition of $\nu(A)$), extends to simple g by linearity, to nonnegative g by MCT. ■

Theorem 1.11: 单调收敛 (MCT) 与控制收敛 (DCT)

[STRUCTURE —know the steps, fudge details]

Let $f_n \to f$ pointwise (a.e.). **MCT:** if $0 \leq f_1 \leq f_2 \leq \cdots$, then $\int f_n\,d\mu \to \int f\,d\mu$. **DCT:** if $|f_n| \leq h$ for some integrable h (a fixed dominating function), then $\int f_n\,d\mu \to \int f\,d\mu$ and $\int |f_n - f|\,d\mu \to 0$.

用途：凡是要“把极限挪进积分号”——交换 lim 与 $\mathbb{E}(\cdot)$、对参数求导穿过积分号（Leibniz 法则）、把分段逼近的积分取极限——都靠 MCT/DCT 持证放行。考场上找一个固定的可积控制函数 h，DCT 就成立。第四章里 bootstrap 一致性的“条件收敛升级为无条件收敛”用的也是 DCT（控制函数取常数 1）。

1.5 自测题 Self-Test

Example (Self-Test 1：可测性).

设 $(\Omega, \mathcal{F}) = ([0,1], \mathcal{B})$。函数 $Y(\omega) = \omega$ 若 $\omega \in [0, \frac{1}{2}]$，$Y(\omega) = 2 - \omega$ 若 $\omega \in (\frac{1}{2}, 1]$。$Y$ 是随机变量吗?

Solution.

是。$\{Y \leq t\}$ 对每个 t 都是两个区间之并（Y 在两段上都连续单调），是 Borel 集；或直接说 Y 是分段连续函数，分段端点构成 Borel 划分，故可测。

Example (Self-Test 2：标准机器).

设 X 有密度 f_X（即 $P_X(B) = \int_B f_X\,d\lambda$）。用换元公式与 Theorem 1.10 证明 $\mathbb{E}(g(X)) = \int_{\mathbb{R}} g(x) f_X(x)\,dx$。

Solution.

由 Theorem 1.9，$\mathbb{E}(g(X)) = \int g\,dP_X$。又 P_X 以 f_X 为关于 λ 的密度，由 Theorem 1.10(b)（取 $\mu = \lambda, \nu = P_X, f = f_X$），$\int g\,dP_X = \int g f_X\,d\lambda = \int g(x) f_X(x)\,dx$。这正是本科“$\mathbb{E}(g(X)) = \int gf$”那条公式的测度论根据。

Example (Self-Test 3：交换极限与积分).

设 $f_n(\omega) = n\,\mathbb{1}_{(0,1/n)}(\omega)$ 在 $([0,1], \lambda)$ 上。$\int f_n\,d\lambda$ 的极限是多少？f_n 的逐点极限的积分又是多少？这与 MCT/DCT 矛盾吗?

Solution.

$\int f_n \, d\lambda = 1$ 对所有 n，故极限为 1。但 $f_n(\omega) \to 0$ 逐点（对每个固定 $\omega > 0$，当 n 大时 $f_n(\omega) = 0$），其积分为 0。不矛盾：f_n 既不单调上升（MCT 不适用），也没有公共可积控制函数（要控制住需 $h(\omega) \geq \sup_n f_n(\omega)$，而 $\int \sup_n f_n = \infty$，DCT 不适用）。这正是“没有控制函数时极限与积分不能交换”的标准反例。

Chapter 2　分布、矩、累积量与不等式（Distributions, Moments, Cumulants & Inequalities）

导读 · Roadmap（先读这里，不含公式）

有了第1章的测度论地基，本章把“随机变量到底长什么样”这件事补齐：先点名一批*命名分布*（Bernoulli/Binomial/Poisson/Geometric、Uniform/Normal/Exponential/Gamma/χ^2），把它们的 pmf/pdf/CDF、期望、方差一次性钉死；再讲随机变量经函数 $g(X)$ 变换后密度怎么算（单调变换的 Jacobian、$Z = X^2$ 这类非单调情形）；接着是*矩生成函数* MGF 与*特征函数*——它们“唯一确定分布”的性质是资格考的常客；然后引入*累积量* cumulants 与累积量母函数 $K(t) = \log M(t)$，前四阶 cumulant 恰好对应 mean / variance / 偏度分子；最后一整套*概率不等式*（Markov、Chebyshev、Jensen、Cauchy–Schwarz、Hölder、Minkowski、Lyapunov、c_r）以及两条*尾界*（Hoeffding、Bernstein），它们是后面 LLN/CLT、检验功效、bootstrap 一致性的弹药库。
资格考在这一章**年年**有题，最爱考五件事：（1）从 hazard / CDF / MGF 这类*特征*反认出是哪个命名分布；（2）算 $g(X)$ 的密度（$Z = X^2$、阶统计量经 PIT 变 Beta）；（3）从 cumulant generating function 求 cumulants；（4）用 Markov + Jensen 在*只许用所给条件*（不许独立、不许同分布）下证一个尾概率界；（5）用 Hoeffding 构造一个有限样本 size-α 检验。本章逐一拆解，并配上 2020 / 2023 / 2024 / 2025 年资格考原题。

2.1　命名分布速查（Named distributions）

[优先级 ●●○ 中 Med]　*资格考足迹*：2020, 2023, 2024, 2025

资格考极少让你“背公式默写”，但几乎每道分布题都默认你*秒认*这几张脸：给一个 pmf/pdf 你要知道它是谁、期望方差是多少、MGF 长什么样。这一节就是这张速查表。约定：离散用 pmf $p(x) = \mathrm{P}(X = x)$，连续用 pdf f，CDF 一律 $F(x) = \mathrm{P}(X \le x)$。

2.1.1 离散分布

Definition 2.1: 四个离散命名分布

- *Bernoulli* $X \sim \mathrm{Bern}(p)$, $p \in [0,1]$: $p(1) = p$, $p(0) = 1-p$. $\mathbb{E}(X) = p$, $\mathrm{Var}(X) = p(1-p)$, $M(t) = 1 - p + pe^t$.
- *Binomial* $X \sim \mathrm{Bin}(n,p)$: $p(k) = \binom{n}{k} p^k (1-p)^{n-k}$, $k = 0, \dots, n$. $\mathbb{E}(X) = np$, $\mathrm{Var}(X) = np(1-p)$, $M(t) = (1-p+pe^t)^n$. ($X = \sum_{i=1}^n X_i$, $X_i \overset{i.i.d.}{\sim} \mathrm{Bern}(p)$.)
- *Poisson* $X \sim \mathrm{Pois}(\lambda)$, $\lambda > 0$: $p(k) = e^{-\lambda}\lambda^k / k!$, $k = 0, 1, 2, \dots$. $\mathbb{E}(X) = \mathrm{Var}(X) = \lambda$, $M(t) = \exp\{\lambda(e^t - 1)\}$.
- *Geometric* $X \sim \mathrm{Geom}(p)$（"首次成功前的失败次数"版）：$p(k) = (1-p)^k p$, $k = 0, 1, 2, \dots$. $\mathbb{E}(X) = (1-p)/p$, $\mathrm{Var}(X) = (1-p)/p^2$, $M(t) = p/(1-(1-p)e^t)$ for $e^t < 1/(1-p)$.

两个口诀。其一：Poisson 是"稀有事件计数"——$\mathrm{Bin}(n,p)$ 在 $n \to \infty$, $np \to \lambda$ 时收敛到 $\mathrm{Pois}(\lambda)$，所以它*均值等于方差*（这是诊断"是不是 Poisson"的体检指标）。其二：Geometric 的"无记忆性" $\mathrm{P}(X \geq s+t \mid X \geq s) = \mathrm{P}(X \geq t)$ 是离散世界里唯一具有此性质的分布，正对应连续世界的指数分布——见 §2.1.2 与 2020 Q2 的 hazard 识别。

2.1.2 连续分布

Definition 2.2: 五个连续命名分布

- *Uniform* $X \sim U[a,b]$: $f(x) = \frac{1}{b-a}\mathbb{1}_{[a,b]}(x)$, $F(x) = \frac{x-a}{b-a}$ on $[a,b]$. $\mathbb{E}(X) = \frac{a+b}{2}$, $\mathrm{Var}(X) = \frac{(b-a)^2}{12}$.
- *Normal* $X \sim N(\mu, \sigma^2)$: $f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\big(-\frac{(x-\mu)^2}{2\sigma^2}\big)$. $\mathbb{E}(X) = \mu$, $\mathrm{Var}(X) = \sigma^2$, $M(t) = \exp\big(\mu t + \frac{1}{2}\sigma^2 t^2\big)$.
- *Exponential* $X \sim \mathrm{Exp}(\lambda)$ (rate $\lambda > 0$): $f(x) = \lambda e^{-\lambda x}\mathbb{1}_{(0,\infty)}(x)$, $F(x) = 1 - e^{-\lambda x}$, $\mathrm{P}(X > x) = e^{-\lambda x}$. $\mathbb{E}(X) = 1/\lambda$, $\mathrm{Var}(X) = 1/\lambda^2$, $M(t) = \frac{\lambda}{\lambda - t}$ for $t < \lambda$.
- *Gamma* $X \sim \mathrm{Gamma}(r, \lambda)$ (shape $r > 0$, rate $\lambda > 0$): $f(x) = \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x}\mathbb{1}_{(0,\infty)}(x)$, where $\Gamma(r) = \int_0^\infty x^{r-1}e^{-x}\,\mathrm{d}x$. $\mathbb{E}(X) = r/\lambda$, $\mathrm{Var}(X) = r/\lambda^2$, $M(t) = \big(\frac{\lambda}{\lambda - t}\big)^r$ for $t < \lambda$.
- *Chi-square* $X \sim \chi^2_k$: the $\mathrm{Gamma}(r = k/2,\ \lambda = 1/2)$ special case, $f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}\mathbb{1}_{(0,\infty)}(x)$. $\mathbb{E}(X) = k$, $\mathrm{Var}(X) = 2k$, $M(t) = (1-2t)^{-k/2}$ for $t < \frac{1}{2}$.

Fact 2.3: 两条贯穿全书的关系链

(i) $\mathrm{Exp}(\lambda) = \mathrm{Gamma}(1,\lambda)$；独立 Gamma 同 rate 相加 shape 相加：$\mathrm{Gamma}(r_1,\lambda)+\mathrm{Gamma}(r_2,\lambda)=\mathrm{Gamma}(r_1+r_2,\lambda)$（由 MGF 相乘立得）。(ii) 若 $Z\sim N(0,1)$ 则 $Z^2\sim\chi^2_1$；独立标准正态平方和 $\sum_{j=1}^{k} Z_j^2\sim\chi^2_k$。后者是所有 Wald / LM 检验渐近卡方分布的根。

Remark (Gamma 的两种参数化别搞混).

Andrews 讲义用 *rate* λ（MGF $=(\lambda/(\lambda-t))^r$），不少教材用 *scale* $\theta=1/\lambda$（MGF $=(1-\theta t)^{-r}$）。本书统一跟 Andrews 用 rate。考场上看到 $f(x)\propto x^{r-1}e^{-\lambda x}$ 就读 rate，看到 $e^{-x/\theta}$ 就读 scale，对应均值分别是 r/λ 与 $r\theta$。

2.2 随机变量的变换：$g(X)$ 的密度

[优先级 ●●● 高 High] 资格考足迹：2020, 2023, 2024

给 X 的密度，问 $Y=g(X)$ 的密度——这是资格考的高频送分题，核心是 Jacobian。分两种情形：g 单调（一对一）用换元公式；g 非单调（如 $g(x)=x^2$）要把 Y 的每个值的所有原像加起来。

Theorem 2.4: 单调变换公式（change of variables / Jacobian）

[REPRODUCE —memorize this proof]

Let X have pdf f_X on an interval, and let g be *strictly monotone* and continuously differentiable on the support of X, with inverse $x=g^{-1}(y)$. Then $Y=g(X)$ has pdf

$$f_Y(y) \;=\; f_X\big(g^{-1}(y)\big)\left|\tfrac{\mathrm{d}}{\mathrm{d}y}\,g^{-1}(y)\right|, \qquad y\in g(\operatorname{supp} X).$$

Proof for Theorem

[REPRODUCE —memorize this proof]

Take g increasing (decreasing is symmetric). Then $F_Y(y)=\mathrm{P}(g(X)\le y)=\mathrm{P}\big(X\le g^{-1}(y)\big)=F_X(g^{-1}(y))$. Differentiate in y via the chain rule: $f_Y(y)=f_X(g^{-1}(y))\cdot\frac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y)$. For decreasing g the event flips to $\{X\ge g^{-1}(y)\}$, giving a minus sign that the absolute value absorbs. Hence the $|\cdot|$. ■

非单调变换（如 $Z=X^2$）的标准三步

当 g 不单调，不能直接套 Theorem 2.4。改走“先 CDF、后求导”或“分支求和”：

Step 1. 写出 Y 的支撑（g 的值域）。

Step 2. 对每个 y，找出所有满足 $g(x)=y$ 的根 $x_1(y),\dots,x_m(y)$；在每个根附近

g 局部单调。

Step 3. 各分支的 Jacobian 相加：$f_Y(y) = \sum_{j=1}^{m} f_X(x_j(y)) \left|\frac{\mathrm{d}x_j}{\mathrm{d}y}\right|$。

等价做法：直接算 $F_Y(y) = \mathrm{P}(g(X) \le y)$ 再求导——非单调时往往更稳，不易漏分支。

2.2.1 Worked example：2020 资格考 Q1(b)（$Z = X^2$ 的密度）

原题 (Econometrics prelim 2020, Q1)

Suppose that (X, Y) has a joint probability density function $f_{XY}(x, y) = 8xy\,\mathbb{1}\{0 \le x \le y \le 1\}$. *(b)* Let $Z = X^2$. Obtain the probability density function of Z.

Solution.

先求 X 的边际密度。固定 $x \in (0, 1)$，支撑要求 $y \in [x, 1]$，故

$$f_X(x) = \int_x^1 8xy\,\mathrm{d}y = 8x \cdot \frac{1 - x^2}{2} = 4x(1 - x^2), \qquad x \in (0, 1).$$

（顺手验证：$\int_0^1 4x(1 - x^2)\,\mathrm{d}x = 4(\frac{1}{2} - \frac{1}{4}) = 1$，是合法密度。）

再做单调变换。 $g(x) = x^2$ 在 X 的支撑 $(0, 1)$ 上严格单调递增（因为 $x > 0$），故可直接用 Theorem 2.4。逆变换 $x = g^{-1}(z) = \sqrt{z}$，$\frac{\mathrm{d}}{\mathrm{d}z}\sqrt{z} = \frac{1}{2\sqrt{z}}$。代入：

$$f_Z(z) = f_X(\sqrt{z}) \cdot \frac{1}{2\sqrt{z}} = 4\sqrt{z}\,(1 - z) \cdot \frac{1}{2\sqrt{z}} = 2(1 - z), \qquad z \in (0, 1).$$

即 $Z \sim \mathrm{Beta}(1, 2)$。验证 $\int_0^1 2(1 - z)\,\mathrm{d}z = 2(1 - \frac{1}{2}) = 1$。■

Remark (为什么这里不用分支求和).

> 虽然 $w \mapsto w^2$ 在整个 $\mathbb{R}$ 上非单调，但 X 的支撑被限制在 $(0, 1) \subset (0, \infty)$，在这一段 g 单调，只有一个根 $\sqrt{z}$，故 Theorem 2.4 直接适用。真正需要本节的非单调变换模板（“分支求和”三步，§2.2）的是支撑跨 0 的情形：若 $X \sim N(0, 1)$，则 $Z = X^2$ 有两个根 $\pm\sqrt{z}$，$f_Z(z) = 2 \cdot \phi(\sqrt{z}) \cdot \frac{1}{2\sqrt{z}} = \frac{1}{\sqrt{2\pi z}} e^{-z/2}$，正是 χ^2_1 的密度——这就是 Fact 2.3(ii) 的密度级证明。

2.2.2 Worked example：2020 资格考 Q2（hazard 识别指数分布 + PIT 得 Beta）

原题 (Econometrics prelim 2020, Q2)

Let $X_1, X_2, \ldots, X_n$ be i.i.d. with $\mathrm{P}(X_1 \le 0) = 0$ and $\lim_{h\downarrow 0} \mathrm{P}(X_1 \le x+h \mid X_1 > x)/h = 1$ for all $x > 0$. Let $X_{(j)}$ be the j-th smallest among $X_1, \ldots, X_n$. *(a)* Compute $\mathbb{E}(X_1)$ and $\mathbb{E}\big(X_{(1)}\big)$. *(b)* Let F be the CDF of X_1. Derive the CDF and pdf of $Y_{(j)} := F(X_{(j)})$ for $j = 1, 2, \ldots, n$.

Solution.

(a) 由 hazard 反认指数分布。定义 *hazard rate* $\lambda(x) := \lim_{h\downarrow 0} \frac{\mathrm{P}(x<X_1\le x+h \mid X_1>x)}{h} = \frac{f(x)}{1-F(x)}$（条件密度／生存概率）。题设 $\lambda(x) = 1$ 对一切 $x > 0$，且支撑在 $(0,\infty)$（因 $\mathrm{P}(X_1 \le 0) = 0$）。解 ODE：$\frac{f}{1-F} = -\frac{\mathrm{d}}{\mathrm{d}x}\log(1-F) = 1 \Rightarrow 1 - F(x) = e^{-x}$，即 $X_1 \sim \mathrm{Exp}(1)$。故

$$\mathbb{E}(X_1) = 1.$$

最小值 $X_{(1)}$ 的分布。 $\mathrm{P}\big(X_{(1)} > x\big) = \mathrm{P}(\text{all } X_i > x) = \big(e^{-x}\big)^n = e^{-nx}$（用 i.i.d.），即 $X_{(1)} \sim \mathrm{Exp}(n)$，于是

$$\mathbb{E}\big(X_{(1)}\big) = \frac{1}{n}.$$

(b) PIT：阶统计量经 CDF 变换得 Beta。关键事实——概率积分变换 *(Probability Integral Transform)*：若 X 连续、CDF F，则 $U := F(X) \sim U[0,1]$。而且 F 单调不减，把 $X_1, \ldots, X_n$ 的次序原封不动地搬到 $U_i := F(X_i)$ 上，故

$$Y_{(j)} = F(X_{(j)}) = U_{(j)}, \qquad U_1, \ldots, U_n \overset{i.i.d.}{\sim} U[0,1].$$

所以 $Y_{(j)}$ 就是 n 个 i.i.d. 均匀变量的第 j 个阶统计量。其分布是 $\mathrm{Beta}(j, n-j+1)$：CDF 用“至少 j 个落在 $[0,y]$”的二项和

$$F_{Y_{(j)}}(y) = \mathrm{P}\big(U_{(j)} \le y\big) = \sum_{k=j}^{n} \binom{n}{k} y^k (1-y)^{n-k}, \qquad y \in [0,1],$$

pdf（对其求导，或直接用阶统计量密度公式）

$$f_{Y_{(j)}}(y) = \frac{n!}{(j-1)!\,(n-j)!}\, y^{j-1}(1-y)^{n-j}, \qquad y \in (0,1).$$

■

Remark (阶统计量密度公式怎么记).

第 j 个阶统计量 $X_{(j)}$ 落在 y 附近的“故事”是：$j-1$ 个样本小于 y（概率各 $F(y)$）、1 个恰在 y（密度 $f(y)$）、$n-j$ 个大于 y（概率各 $1-F(y)$），再乘多项式系数 $\frac{n!}{(j-1)!\,1!\,(n-j)!}$：

$$f_{X_{(j)}}(y) = \frac{n!}{(j-1)!\,(n-j)!}\,F(y)^{j-1}\big(1-F(y)\big)^{n-j} f(y).$$

取 $F(y)=y$（均匀）就回到上面的 Beta。这个“多项式计数”记忆法考场上比硬背 Beta 参数靠谱。

2.3 矩、中心矩、方差

[优先级 ●●○ 中 Med] 资格考足迹：2020, 2024, 2025

Definition 2.5: 矩与中心矩 (Moments)

For a random variable X with $\mathbb{E}(|X|^r) < \infty$:

- the *r-th (raw) moment* is $\mu'_r := \mathbb{E}(X^r)$;
- the *r-th central moment* is $\mu_r := \mathbb{E}((X-\mu)^r)$ where $\mu = \mathbb{E}(X) = \mu'_1$.

In particular $\mu_2 = \mathrm{Var}(X) = \mathbb{E}\big(X^2\big) - \mu^2$. The *skewness* is $\gamma_1 = \mu_3/\sigma^3$ and the *(excess) kurtosis* is $\gamma_2 = \mu_4/\sigma^4 - 3$, with $\sigma = \sqrt{\mu_2}$.

Fact 2.6: 矩存在性的“向下传染”（Lyapunov 的推论）

If $\mathbb{E}(|X|^s) < \infty$ then $\mathbb{E}(|X|^r) < \infty$ for all $0 < r \le s$. （高阶矩存在 ⇒ 一切低阶矩存在。）证见 Lyapunov 不等式 2.13。

这条“向下传染”很常用：题目若说假设 $\mathbb{E}\big(X^4\big) < \infty$，你就免费拥有 $\mathbb{E}\big(X^3\big)$、$\mathbb{E}\big(X^2\big)$、$\mathbb{E}(|X|)$ 全部有限——CLT 里要二阶矩、delta 方法里要四阶矩，都是靠这条把“给的高阶矩”降级使用。反过来不成立：Cauchy 分布 $\mathbb{E}(|X|) = \infty$，连一阶矩都没有。

2.4 MGF、特征函数与“唯一确定分布”

[优先级 ●●● 高 High] 资格考足迹：2020, 2024

Definition 2.7: 矩生成函数 MGF

The *moment generating function* of X is $M_X(t) := \mathbb{E}\big(e^{tX}\big)$, defined for t in a neighborhood of 0 when it is finite there. When finite near 0 it generates

moments:

$$\mathbb{E}(X^r) = \frac{\mathrm{d}^r}{\mathrm{d}t^r} M_X(t)\Big|_{t=0}, \qquad r = 1, 2, \dots$$

Theorem 2.8: MGF / 特征函数唯一确定分布

[STRUCTURE —know the steps, fudge details]

(i) If M_X and M_Y are finite in a common neighborhood of 0 and $M_X(t) = M_Y(t)$ there, then $X \overset{d}{=} Y$ (same distribution). (ii) The *characteristic function* $\varphi_X(t) := \mathbb{E}\big(e^{itX}\big)$ ($i = \sqrt{-1}$) is *always* finite ($|e^{itX}| = 1$) and *always* determines the distribution: $\varphi_X = \varphi_Y \iff X \overset{d}{=} Y$. Moments are read off via $\frac{\mathrm{d}^r}{\mathrm{d}t^r}\varphi_X(t)|_{t=0} = i^r\mathbb{E}(X^r)$ whenever $\mathbb{E}(|X|^r) < \infty$.

为什么要分两个工具？MGF 直观、好算导数取矩，但可能不存在（重尾分布如 Cauchy、t 分布的 MGF 在 $t \neq 0$ 处为 $+\infty$）。特征函数把 t 换成 it，让指数变成单位圆上的旋转 $|e^{itX}| = 1$，因此对任何随机变量都有限、都好用——CLT 的标准证明就走特征函数。考场口诀：**要取矩、分布够轻**用 MGF；**要证 CLT、或分布重尾**用特征函数。

"认 MGF"识别分布（2020 Q3 / 2024 Q4 套路）

当题目给联合 MGF $\mathbb{E}\big(e^{tX+sY}\big)$ 让你"识别 (X, Y)"：

1. 看能否因式分解 $M_{X,Y}(t, s) = M_X(t)\, M_Y(s)$——若能，$X \perp Y$（独立）。
2. 把各因子与速查表（Def 2.2）对号：$e^{t^2/2}$ 就是 $N(0, 1)$ 的 MGF。
3. 任何线性组合 $aX + bY$ 的 MGF $= \mathbb{E}\big(e^{(at)X+(bt)Y}\big) = M_{X,Y}(at, bt)$；若结果形如 $e^{ct^2/2}$ 则该组合 $\sim N(0, c)$。联合正态的边际与条件全是正态。

Example (2024 资格考 Q4(a) 的 MGF 识别一步).

2024 Q4 给 $\mathbb{E}\big(e^{tX+sY}\big) = \exp\{(t^2 + s^2)/2\}$，问 $Y \mid 2X + Y$ 的条件密度。

Solution.

MGF 因式分解 $= \exp(t^2/2)\exp(s^2/2) = M_X(t)M_Y(s)$，每个因子是 $N(0, 1)$ 的 MGF，故 $X, Y \overset{i.i.d.}{\sim} N(0, 1)$。于是 $(Y, 2X+Y)$ 联合正态，$\mathrm{Var}(2X + Y) = 4+1 = 5$，$\mathrm{Cov}(Y, 2X + Y) = \mathrm{Var}(Y) = 1$。联合正态下条件分布仍正态：$Y \mid (2X+Y = w) \sim N\big(\frac{1}{5}w,\ 1 - \frac{1}{5}\big) = N\big(\frac{w}{5}, \frac{4}{5}\big)$。这里只示范第一步"认脸"；条件正态公式见第3章。

2.5 累积量与累积量母函数 $K(t) = \log M(t)$

[优先级 ●●● 高 High] 资格考足迹：2025

Definition 2.9: 累积量母函数与累积量 (Cumulants)

The *cumulant generating function* (cgf) of X is $K_X(t) := \log M_X(t)$ (defined where M_X is finite and positive near 0). The *k-th cumulant* κ_k is the k-th Taylor coefficient at 0:

$$K_X(t) = \sum_{k=1}^{\infty} \kappa_k \frac{t^k}{k!}, \qquad \kappa_k = \frac{\mathrm{d}^k}{\mathrm{d}t^k} K_X(t)\Big|_{t=0}.$$

Theorem 2.10: 前四阶 cumulant 与矩的对应

[REPRODUCE —memorize this proof]

$$\kappa_1 = \mu = \mathbb{E}(X), \qquad \kappa_2 = \sigma^2 = \mathrm{Var}(X), \qquad \kappa_3 = \mu_3 = \mathbb{E}\big((X-\mu)^3\big),$$

$$\kappa_4 = \mu_4 - 3\sigma^4.$$

Hence skewness $\gamma_1 = \kappa_3/\kappa_2^{3/2}$ and excess kurtosis $\gamma_2 = \kappa_4/\kappa_2^2$.

Proof for Theorem

[REPRODUCE —memorize this proof]

Write $M(t) = 1 + \mu_1' t + \frac{\mu_2'}{2}t^2 + \frac{\mu_3'}{6}t^3 + \cdots$ and expand $K = \log M = \log(1+u) = u - \frac{u^2}{2} + \frac{u^3}{3} - \cdots$ with $u = M(t) - 1$. Collecting powers of t:

$$\kappa_1 = \mu_1', \qquad \kappa_2 = \mu_2' - (\mu_1')^2, \qquad \kappa_3 = \mu_3' - 3\mu_1'\mu_2' + 2(\mu_1')^3,$$
$$\kappa_4 = \mu_4' - 4\mu_1'\mu_3' - 3(\mu_2')^2 + 12(\mu_1')^2\mu_2' - 6(\mu_1')^4.$$

Re-centering ($\mu_2' - (\mu_1')^2 = \sigma^2$, etc.) gives the stated $\kappa_2 = \sigma^2$, $\kappa_3 = \mu_3$, $\kappa_4 = \mu_4 - 3\sigma^4$. ■

Remark (cumulant 为什么好用).

两条杀手锏。**(1) 可加性**：独立随机变量之和，cumulant 直接相加 $\kappa_k(X+Y) = \kappa_k(X) + \kappa_k(Y)$（因为 $K_{X+Y} = K_X + K_Y$，独立时 MGF 相乘、取 log 即相加）。所以 $\bar{X} = \frac{1}{n}\sum X_i$ 的 $\kappa_k(\bar{X}) = n^{1-k}\kappa_k(X_1)$——$k \geq 3$ 的高阶 cumulant 以 n^{1-k} 速度消失，这正是 CLT 的"正态化"：$\kappa_3, \kappa_4 \to 0$。**(2) 正态的指纹**：$N(\mu, \sigma^2)$ 的 $K(t) = \mu t + \frac{1}{2}\sigma^2 t^2$ 是二次多项式，故 $\kappa_1 = \mu, \kappa_2 = \sigma^2$，而 $\kappa_k = 0$ 对一切 $k \geq 3$。一个分布是正态 $\iff$ 它三阶及以上 cumulant 全为零。

2.5.1 Worked example：2025 资格考 Q5(a)（指数分布的 cumulants）

原题 (Comprehensive Exam 2025, Pinkse, Q5)

Let $\{u_i\}$ be i.i.d. standard exponential (i.e. u_i has density $\exp(-u)\mathbb{1}(u \geq 0)$).
(a) Determine the cumulants of u_i.

Solution.

$u_i \sim \text{Exp}(1)$，rate $\lambda = 1$。由速查表（Def 2.2）其 MGF 为

$$M(t) = \frac{\lambda}{\lambda - t} = \frac{1}{1-t}, \qquad t < 1.$$

取对数得 cgf

$$K(t) = \log M(t) = -\log(1-t), \qquad t < 1.$$

对 $\log(1-t)$ 用幂级数 $-\log(1-t) = \sum_{k=1}^{\infty} \frac{t^k}{k}$，与定义 $K(t) = \sum_{k\geq 1} \kappa_k \frac{t^k}{k!}$ 对比系数：

$$\frac{\kappa_k}{k!} = \frac{1}{k} \implies \boxed{\kappa_k = \frac{k!}{k} = (k-1)!}, \qquad k = 1, 2, 3, \ldots$$

逐个核对：$\kappa_1 = 0! = 1$（均值 $= 1/\lambda = 1$，对）；$\kappa_2 = 1! = 1$（方差 $= 1/\lambda^2 = 1$，对）；$\kappa_3 = 2! = 2$，$\kappa_4 = 3! = 6$。■

Remark (考点提炼).

> 这题就是 Def 2.9 的直接应用：*写 MGF → 取 log → 套已知幂级数 → 比系数*。陷阱有二：(1) 一定要写 $K(t) = \sum \kappa_k t^k/\mathbf{k!}$（带阶乘!），漏了 $k!$ 会把 κ_k 算成 $1/k$。(2) 收敛域 $t < 1$ 要写，否则 MGF 在 $t \geq 1$ 发散。顺带：由 $\kappa_3 = 2 > 0$ 知指数分布右偏（skewness $= \kappa_3/\kappa_2^{3/2} = 2$），与图形吻合。

2.6 概率不等式总览

[优先级 ●●● 高 High] 资格考足迹：2023, 2024, 2025

这是全章弹药最密集的一节。资格考用它们来：控尾概率（Markov、Chebyshev、Hoeffding、Bernstein）、比矩的大小（Jensen、CBS、Hölder、Lyapunov、c_r）、证三角不等式（Minkowski）。下面按"先矩界、后尾界"分两组。

2.6.1 矩不等式（Jensen / CBS / Hölder / Minkowski / Lyapunov / c_r）

Theorem 2.11: Jensen 不等式

[REPRODUCE —memorize this proof]

Let $g:\mathbb{R}\to\mathbb{R}$ and X with $\mathbb{E}(|X|)<\infty$. *If g is convex*, then $g(\mathbb{E}(X))\le \mathbb{E}(g(X))$. *If g is concave*, the inequality reverses: $g(\mathbb{E}(X))\ge\mathbb{E}(g(X))$. If g is strictly convex (concave) and X is non-degenerate, the inequality is strict.

Proof for Theorem

[REPRODUCE —memorize this proof]

Convex case. By convexity there is a supporting line $\ell(x)=a+bx$ at the point $\mu=\mathbb{E}(X)$ with $\ell(\mu)=g(\mu)$ and $\ell(x)\le g(x)$ for all x. Take expectations of $\ell(X)\le g(X)$: by linearity $\mathbb{E}(\ell(X))=a+b\mu=\ell(\mu)=g(\mu)$, so $g(\mathbb{E}(X))=\mathbb{E}(\ell(X))\le\mathbb{E}(g(X))$. Concave case: apply to $-g$. ■

记方向的口诀：**凸函数“先期望后施函数”更小**（$g(\mathbb{E}(X))\le\mathbb{E}(g(X))$），凹的反过来。两个最常用的特例：(1) $g(x)=x^2$ 凸 $\Rightarrow (\mathbb{E}(X))^2\le\mathbb{E}(X^2)$，即方差 ≥ 0；(2) $g=\log$ 凹 $\Rightarrow \log\mathbb{E}(X)\ge\mathbb{E}(\log X)$——这条正是下面 2025 Q4 与第5章 MLE 一致性（Kullback–Leibler 非负）的引擎。还有 $g(x)=|x|$ 凸 $\Rightarrow |\mathbb{E}(X)|\le\mathbb{E}(|X|)$。

Theorem 2.12: Cauchy–Schwarz / Hölder / Minkowski

[STRUCTURE —know the steps, fudge details]

Let $p,q\ge 1$ with $\frac{1}{p}+\frac{1}{q}=1$ (conjugate exponents). Writing $\|X\|_p:=(\mathbb{E}(|X|^p))^{1/p}$:

- *Hölder:* $\mathbb{E}(|XY|)\le\|X\|_p\,\|Y\|_q$.
- *Cauchy–Schwarz (CBS)* ($p=q=2$): $\big|\mathbb{E}(XY)\big|\le\mathbb{E}(|XY|)\le\big(\mathbb{E}(X^2)\big)^{1/2}\big(\mathbb{E}(Y^2)\big)^{1/2}$.
- *Minkowski (triangle ineq. for L^p)* ($p\ge1$): $\|X+Y\|_p\le\|X\|_p+\|Y\|_p$.

三者层级：Hölder 是“母不等式”，$p=q=2$ 退化成 CBS，CBS 又喂给 Minkowski 的证明。直觉上 Hölder 是“积的期望 ≤ 各自范数之积”，CBS 是它在 L^2 的特例，正是统计里相关系数 $|\mathrm{Corr}|\le 1$ 的来源（取 $X-\mathbb{E}(X)$，$Y-\mathbb{E}(Y)$ 即得 $|\mathrm{Cov}(X,Y)|\le\sigma_X\sigma_Y$）。Minkowski 保证 $\|\cdot\|_p$ 是个合格的范数（满足三角不等式）。

Theorem 2.13: Lyapunov 不等式（L^r 范数随 r 单调）

[REPRODUCE —memorize this proof]

For $0<r\le s$: $\big(\mathbb{E}(|X|^r)\big)^{1/r}\le\big(\mathbb{E}(|X|^s)\big)^{1/s}$, i.e. $\|X\|_r\le\|X\|_s$. Hence

$\mathbb{E}(|X|^s) < \infty \Rightarrow \mathbb{E}(|X|^r) < \infty$ for all $0 < r \le s$.

Proof for Theorem

[REPRODUCE —memorize this proof]

Apply Hölder to $|X|^r \cdot 1$ with exponent $p = s/r \ge 1$ and its conjugate q: $\mathbb{E}(|X|^r) = \mathbb{E}(|X|^r \cdot 1) \le \big(\mathbb{E}((|X|^r)^p)\big)^{1/p} \cdot 1 = \big(\mathbb{E}(|X|^s)\big)^{r/s}$. Raise both sides to power $1/r$. ■

Theorem 2.14: c_r 不等式

[STRUCTURE —know the steps, fudge details]

For $r > 0$ and any random variables X, Y,

$$\mathbb{E}(|X+Y|^r) \le c_r\big(\mathbb{E}(|X|^r) + \mathbb{E}(|Y|^r)\big), \qquad c_r = \begin{cases} 1, & 0 < r \le 1, \\ 2^{r-1}, & r \ge 1. \end{cases}$$

c_r 不等式是“把和的矩拆成各项矩之和”的粗糙但万能的工具。证明：$r \ge 1$ 时 $x \mapsto |x|^r$ 凸，由 Jensen $\left|\frac{X+Y}{2}\right|^r \le \frac{1}{2}(|X|^r + |Y|^r)$，乘 2^r 即得 $c_r = 2^{r-1}$；$0 < r \le 1$ 时用 $|a+b|^r \le |a|^r + |b|^r$（次可加性）。它在第4章证 LLN（截断后控制余项的矩）时反复出现。

Theorem 2.15: von Bahr–Esseen 不等式（独立均值零和的 p 阶矩，$1 \le p \le 2$）

[STRUCTURE —know the steps, fudge details]

Let $x_1, \dots, x_n$ be *independent* with $\mathbb{E}(x_i) = 0$ and $\mathbb{E}(|x_i|^p) < \infty$ for some $1 \le p \le 2$. Write $S_n = \sum_{i=1}^n x_i$. Then

$$\mathbb{E}\big(|S_n|^p\big) \;\le\; C_p \sum_{i=1}^n \mathbb{E}(|x_i|^p), \qquad C_p = 2 \;\big(\text{可取更紧的} 2 - \tfrac{1}{n}\big).$$

特别地，i.i.d. 时 $\mathbb{E}(|S_n|^p) \le 2n\,\mathbb{E}(|x_1|^p) = O(n)$。

这是本节唯一**专为独立和打造**的矩不等式，也是资格考**收敛率**题的核心引擎。它把 c_r 不等式（Theorem 2.14 迭代 n 项给出的 $\mathbb{E}(|S_n|^p) \le n^{p-1}\sum \mathbb{E}(|x_i|^p)$，粗糙）在“独立 + 均值零 + $p \le 2$”下收紧到**线性** $O(n)$——关键正是这个 n 而非 n^{p-1}。**用途（$1 < p \le 2$、只有 p 阶矩、方差可能无穷）**：由它 $\mathbb{E}(|S_n|^p) = O(n)$，于是 $\mathbb{E}\big(|n^{1-1/p}\bar{x}|^p\big) = \mathbb{E}(|S_n|^p)/n = O(1)$，再对 $n^{1-1/p}\bar{x}$ 用 Markov（2.16）即得

$$\bar{x} - \mu = O_p\big(n^{-(1-1/p)}\big),$$

正是 2025 年资格考 Q1(b) 的解法（对应收敛率定理 4.18）。它是 Marcinkiewicz–

Zygmund 不等式在 $1 \le p \le 2$ 的上界特例（M-Z 是双边 $\mathbb{E}(|S_n|^p) \asymp \mathbb{E}\big((\sum x_i^2)^{p/2}\big)$，对一般 $p \ge 1$）；当 $p \ge 2$（方差有限）时不必动用它，直接 CLT 给更快的 $O_p(n^{-1/2})$。证明属专著级（von Bahr & Esseen 1965，用 $|1+z|^p$ 的凸性配独立性归纳），考场只需会用，故标 **[STRUCTURE —know the steps, fudge details]**

。

2.6.2 尾界（Markov / Chebyshev / Hoeffding / Bernstein）

Theorem 2.16: Markov 不等式

[REPRODUCE —memorize this proof]

For any random variable X and any $k > 0$,

$$\mathrm{P}(|X| \ge k) \le \frac{\mathbb{E}(|X|)}{k}.$$

More generally, for any nondecreasing nonnegative ϕ with $\phi(k) > 0$: $\mathrm{P}(|X| \ge k) \le \mathbb{E}(\phi(|X|))\,/\phi(k)$.

Proof for Theorem

[REPRODUCE —memorize this proof]

Split $|X| = |X|\mathbb{1}\{|X| \ge k\} + |X|\mathbb{1}\{|X| < k\}$. Drop the second (nonnegative) term and bound the first below by $k\mathbb{1}\{|X| \ge k\}$:

$$\mathbb{E}(|X|) \;\ge\; \mathbb{E}(|X|\mathbb{1}\{|X| \ge k\}) \;\ge\; k\,\mathbb{E}(\mathbb{1}\{|X| \ge k\}) = k\,\mathrm{P}(|X| \ge k)\,.$$

Divide by k. ■

Theorem 2.17: Chebyshev 不等式

[REPRODUCE —memorize this proof]

If $\mathbb{E}(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2 < \infty$, then for every $b > 0$,

$$\mathrm{P}\big(|X-\mu| \ge b\,\sigma\big) \le \frac{1}{b^2}, \qquad \text{equivalently} \qquad \mathrm{P}(|X-\mu| \ge \varepsilon) \le \frac{\sigma^2}{\varepsilon^2}.$$

Proof for Theorem

[REPRODUCE —memorize this proof]

Apply Markov to the nonnegative variable $(X-\mu)^2$ at level ε^2: $\mathrm{P}(|X-\mu| \ge \varepsilon) = \mathrm{P}\big((X-\mu)^2 \ge \varepsilon^2\big) \le \mathbb{E}\big((X-\mu)^2\big)\,/\varepsilon^2 = \sigma^2/\varepsilon^2$. Setting $\varepsilon = b\sigma$ recovers the first form: $\mathrm{P}(|X-\mu| \ge b\sigma) \le \sigma^2/(b\sigma)^2 = 1/b^2$. ■

Markov 只要一阶矩就能用（甚至非负随机变量直接用），但很弱；Chebyshev 用了二阶矩，强一些，是弱大数律（WLLN）标准证明的主力（$\mathrm{P}\big(|\bar{X}-\mu| \ge \varepsilon\big) \le$

$\mathrm{Var}(\bar{X})/\varepsilon^2 = \sigma^2/(n\varepsilon^2) \to 0$）。这两条给的是多项式尾界（$\sim 1/k$ 或 $1/k^2$）。要指数尾界，得上 Hoeffding / Bernstein。

Theorem 2.18: Hoeffding 不等式（有界变量的指数尾界）

[STRUCTURE —know the steps, fudge details]

Let $x_1, \dots, x_n$ be *independent* (not necessarily identically distributed) with $\mathrm{P}(a_i \le x_i \le b_i) = 1$. Then for all $z > 0$,

$$\mathrm{P}\left(\left| \sum_{i=1}^{n} \big(x_i - \mathbb{E}(x_i)\big) \right| > z \right) \le 2\exp\left(\frac{-2z^2}{\sum_{i=1}^{n}(b_i - a_i)^2} \right).$$

Theorem 2.19: Bernstein 不等式（用方差收紧的指数尾界）

[STRUCTURE —know the steps, fudge details]

Let $x_1, \dots, x_n$ be independent, mean zero, with $|x_i| \le M$ a.s. and $\sum_i \mathrm{Var}(x_i) =: v$. Then for all $z > 0$,

$$\mathrm{P}\left(\left| \sum_{i=1}^{n} x_i \right| > z \right) \le 2\exp\left(\frac{-z^2/2}{v + Mz/3} \right).$$

Hoeffding 只用支撑宽度 $b_i - a_i$，方差信息浪费掉了；当真实方差远小于 $(b-a)^2/4$ 时它太松。Bernstein 把分母换成方差 v（小 z 时主导），尾巴更细——代价是要知道方差界并假设有界。考场用法：要一个有限样本（非渐近）的尾概率界、变量有界，就报 Hoeffding；想更紧、且方差小，报 Bernstein。**Marcinkiewicz–Zygmund 不等式**（把 $\mathbb{E}(|\sum(x_i - \mathbb{E}(x_i))|^p)$ 用 $\mathbb{E}\big((\sum(x_i - \mathbb{E}(x_i))^2)^{p/2}\big)$ 控制）属于同一弹药库，但用于 LLN 的收敛率，第4章再细讲；其 $1 \le p \le 2$ 的上界特例 **von Bahr–Esseen 不等式**已在 Theorem 2.15 单列（收敛率题的主力）。

2.7 不等式的资格考实战

[优先级 ●●● 高 High] 资格考足迹：2023, 2024, 2025

2.7.1 Worked example：2025 资格考 Q4（Markov + Jensen，只许用所给条件）

原题 (Comprehensive Exam 2025, Pinkse, Q4)

Suppose that for any i, x_i is such that $\mathbb{E}(x_i) = 3$ and $\mathrm{P}(x_i \ge 1) = 1$. Show that $\mathrm{P}(\sum_{i=1}^{n} \log x_i > 3n) \le 1/3$. You may NOT assume anything beyond the conditions stated (e.g. no independence, no identical distribution).

Solution.

思路：题面禁了独立，所以不能把 $\mathrm{P}\big(\prod x_i > e^{3n}\big)$ 拆成乘积。但 Markov 只需一个非负随机变量和它的期望，而期望对求和是线性的（与依赖结构无关）。难点是 $\mathbb{E}(\sum \log x_i)$ 不好直接算——用 Jensen 把“对数的平均”换成“平均的对数”，正好把事件转成关于样本均值 $\bar{x}$ 的事件，而 $\mathbb{E}(\bar{x}) = 3$ 是已知的。

第一步：用 Jensen 把事件抬到 $\bar{x}$ 上。在事件 $A := \{\sum_{i=1}^n \log x_i > 3n\}$ 上，两边除以 n 得 $\frac{1}{n}\sum_{i=1}^n \log x_i > 3$。由 log 的凹性（Jensen，凹方向，Theorem 2.11），对确定性的算术平均也成立（取在 n 个点上均匀分布的“随机变量”）：

$$\frac{1}{n}\sum_{i=1}^n \log x_i \;\le\; \log\Big(\frac{1}{n}\sum_{i=1}^n x_i\Big) = \log \bar{x}.$$

于是在 A 上有 $3 < \log \bar{x}$，即 $\bar{x} > e^3$。这说明事件包含关系

$$A \;\subseteq\; \{\bar{x} > e^3\} \quad\Longrightarrow\quad \mathrm{P}(A) \le \mathrm{P}\big(\bar{x} > e^3\big).$$

第二步：对 $\bar{x}$ 用 Markov。因 $\mathrm{P}(x_i \ge 1) = 1$，故 $x_i \ge 1 > 0$ a.s.，从而 $\bar{x} > 0$ 是非负随机变量，可用 Markov（Theorem 2.16）。又由期望线性性（无需独立）$\mathbb{E}(\bar{x}) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(x_i) = \frac{1}{n}\cdot n \cdot 3 = 3$。故

$$\mathrm{P}\big(\bar{x} > e^3\big) \;\le\; \frac{\mathbb{E}(\bar{x})}{e^3} = \frac{3}{e^3}.$$

第三步：收尾。合并两步：

$$\mathrm{P}\left(\sum_{i=1}^n \log x_i > 3n\right) \;\le\; \frac{3}{e^3} \;\approx\; 0.149 \;\le\; \frac{1}{3}.$$

（数值上 $e^3 \approx 20.09$，$3/e^3 \approx 0.149 < 0.333$。）■

Remark（为什么不能走“$\mathbb{E}(\log x_i) \le \log 3$ 再 Markov 对 $\sum \log x_i$”）.

> 那条路给的是 $\mathrm{P}(A) \le \frac{\mathbb{E}(\sum \log x_i)}{3n} \le \frac{n \log 3}{3n} = \frac{\log 3}{3} \approx 0.366 > \frac{1}{3}$——不够！差就差在 Markov 的阈值用了 $3n$，而 $\mathbb{E}(\sum \log x_i) \le n \log 3$ 太大。正确的招是先用 Jensen 把整个事件搬到 $\bar{x}$ 上、再对均值用 Markov，这样阈值变成 e^3（巨大），分子 $\mathbb{E}(\bar{x}) = 3$（小），比值才压到 $1/3$ 以下。考场上这是“先 Jensen 改写事件、再 Markov 控尾”的经典两段式，且全程只用了 $\mathbb{E}(x_i) = 3$ 与 $x_i \ge 1$ 两个给定条件，没碰独立性。

2.7.2 Worked example：2024 资格考 Q3(a)(b)（中位数最小化 $\mathbb{E}(|\cdot|)$ + 均值–中位数距离的 Jensen 界）

原题 (Comprehensive Exam 2024, Sung Jae Jun, Q3)

Suppose that X has a probability density function f on $\mathbb{R}$ such that $f(x) > 0$ for all $x \in \mathbb{R}$, and $\int_{-\infty}^{\infty} x^2 f(x)\,\mathrm{d}x < \infty$. For $y \in \mathbb{R}$ define $g(y) := \int_{-\infty}^{\infty} |x-y| f(x)\,\mathrm{d}x$. *(a)* Derive a characterization of the value y^* that minimizes g using the density f. *(b)* Let $x^* := \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x$ (the mean). What is the largest possible distance between x^* and y^*?

Solution.

(a) 一阶条件 ⇒ 中位数。 $g(y) = \mathbb{E}(|X-y|)$。把绝对值按符号拆开：

$$g(y) = \int_{-\infty}^{y} (y-x) f(x)\,\mathrm{d}x + \int_{y}^{\infty} (x-y) f(x)\,\mathrm{d}x.$$

对 y 求导（积分上下限处被积函数为 0，故 Leibniz 边界项消失，只剩对 y 的偏导；交换求导与积分由 $|x-y|$ 的导数有界 ± 1 与 DCT 保证）：

$$g'(y) = \int_{-\infty}^{y} f(x)\,\mathrm{d}x - \int_{y}^{\infty} f(x)\,\mathrm{d}x = F(y) - \big(1 - F(y)\big) = 2F(y) - 1.$$

令 $g'(y^*) = 0$ 得 $F(y^*) = \frac{1}{2}$，即 y^* 是中位数。又 $g''(y) = 2f(y) > 0$（$f > 0$），g 严格凸，故 y^* 是唯一全局最小。结论：

$$\boxed{F(y^*) = \tfrac{1}{2}} \quad (y^* \text{ 即} X \text{ 的中位数}).$$

(b) 均值–中位数距离 $\le \sigma$。 记均值 $x^* = \mu = \mathbb{E}(X)$、中位数 $y^* = m$、$\sigma = \sqrt{\mathrm{Var}(X)}$（由 $\mathbb{E}\big(X^2\big) < \infty$ 有限）。关键是 m 最小化 g，故

$$\mathbb{E}(|X-m|) = g(m) \le g(\mu) = \mathbb{E}(|X-\mu|)\,.$$

对左边用 $|\mu - m| = |\mathbb{E}(X) - m| = |\mathbb{E}(X-m)| \le \mathbb{E}(|X-m|)$（Jensen，$|\cdot|$ 凸，Theorem 2.11）。对右边用 Jensen / CBS：$\mathbb{E}(|X-\mu|) = \mathbb{E}(|X-\mu| \cdot 1) \le \big(\mathbb{E}\big((X-\mu)^2\big)\big)^{1/2} = \sigma$（这是 $L^1 \le L^2$，即 Lyapunov，Theorem 2.13）。串起来：

$$|\mu - m| \;\le\; \mathbb{E}(|X-m|) \;\le\; \mathbb{E}(|X-\mu|) \;\le\; \sigma.$$

故最大可能距离是 $\sigma = \sqrt{\mathrm{Var}(X)}$（一个标准差）：

$$\boxed{\big|\,\text{mean} - \text{median}\,\big| \;\le\; \sqrt{\mathrm{Var}(X)}}\,.$$

■

Remark (这两段各踩一个考点).

(a) 是"$\mathbb{E}(|X-y|)$ 的最小值点是中位数"的 FOC 推导——核心是 $\frac{\mathrm{d}}{\mathrm{d}y}|x-y| = -\operatorname{sgn}(x-y)$，期望后变 $2F(y)-1$；务必说明可交换求导与积分（DCT，控制函数 1）。(b) 是著名的均值–中位数不等式 $|\mu - m| \le \sigma$，证法是"中位数最小化 $\mathbb{E}(|X-\cdot|)$"这一最优性夹在中间，两头各用一次 Jensen（$|\mathbb{E}(\cdot)| \le \mathbb{E}(|\cdot|)$ 与 $\mathbb{E}(|\cdot|) \le L^2$ 范数）。这是 Jensen 凸方向最漂亮的应用之一。

2.7.3 Worked example：2023 资格考 Q6（Hoeffding 构造 size-α 检验）

原题 (Comprehensive Exam 2023, Joris Pinkse, Q6)

Recall the Hoeffding inequality: if $\{x_i\}$ are independent (not necessarily identically distributed) and for numbers $\{a_i\}, \{b_i\}$, $\mathrm{P}(a_i \le x_i \le b_i) = 1$, then for all $z > 0$, $\mathrm{P}(|\sum_{i=1}^n (x_i - \mathbb{E}(x_i))| > z) \le 2\exp(-2z^2/\sum_{i=1}^n (b_i - a_i)^2)$. Suppose the support of the i.i.d. $\{x_i\}$ is $[-1/\sqrt{2}, 1/\sqrt{2}]$. Construct a test for $H_0 : \mu_0 = 0$ against $H_1 : \mu_0 \neq 0$, where $\mu_0 = \mathbb{E}(x_1)$, that has rejection probability no greater than α under H_0. Use a decision rule of the form "reject H_0 if $|\bar{x}| > C_\alpha$".

Solution.

把 Hoeffding 套到 $\bar{x}$ 上。支撑 $[-1/\sqrt{2}, 1/\sqrt{2}]$ 给出 $a_i = -1/\sqrt{2}$，$b_i = 1/\sqrt{2}$，故每个宽度 $b_i - a_i = \sqrt{2}$，于是 $\sum_{i=1}^n (b_i - a_i)^2 = n \cdot (\sqrt{2})^2 = 2n$。在 H_0 下 $\mathbb{E}(x_i) = \mu_0 = 0$，所以 $\sum_{i=1}^n (x_i - \mathbb{E}(x_i)) = \sum_{i=1}^n x_i = n\bar{x}$。检验统计量是 $|\bar{x}|$，事件 $\{|\bar{x}| > C\} = \{|\sum x_i| > nC\}$。取 Hoeffding 的 $z = nC$：

$$\mathrm{P}_{H_0}(|\bar{x}| > C) = \mathrm{P}\left(\left| \sum_{i=1}^n x_i \right| > nC \right) \;\le\; 2\exp\left(\frac{-2(nC)^2}{2n} \right) = 2\exp\big(-nC^2\big).$$

解临界值 C_α。令右端 $= \alpha$：

$$2e^{-nC_\alpha^2} = \alpha \implies C_\alpha = \sqrt{\frac{1}{n}\log\frac{2}{\alpha}}.$$

结论：检验规则

$$\boxed{\text{reject } H_0 \text{ iff } |\bar{x}| > C_\alpha = \sqrt{\tfrac{1}{n}\log(2/\alpha)}}.$$

由构造，在 H_0 下 $\mathrm{P}_{H_0}(\text{reject}) = \mathrm{P}_{H_0}(|\bar{x}| > C_\alpha) \le 2e^{-nC_\alpha^2} = \alpha$。这是一个有限样本（非渐近）、level-$\alpha$ 的检验——对任何样本量 n 都成立，不依赖 CLT 或正态近似。■

Remark (考点提炼).

> 这题的精髓：Hoeffding 把“尾概率 $\le$ 函数”反过来用——令尾界等于 α、解出临界值，就得到一个保证 size 的检验。关键代数是 $\sum(b_i-a_i)^2 = 2n$ 与 $z = nC$，让指数恰好化简成 $-nC^2$。对比基于 CLT 的 z-检验（$C_\alpha = z_{\alpha/2}\,\hat\sigma/\sqrt{n}$，只渐近有效），Hoeffding 检验更保守（临界值更大、功效更低）但有限样本严格有效。$\sqrt{\log(2/\alpha)/n}$ 这个 $1/\sqrt{n}$ 收缩率与 CLT 同阶，是它好用的原因。

2.8 自测题 Self-Test

Example (Self-Test 1：MGF 取矩与认分布).

设 $M_X(t) = (1-2t)^{-3}$，$t < \frac{1}{2}$。X 是什么分布？求 $\mathbb{E}(X)$ 与 $\mathrm{Var}(X)$。

Solution.

对照 Gamma 的 MGF $(\lambda/(\lambda-t))^r = (1-t/\lambda)^{-r}$：这里 $(1-2t)^{-3} = (1-t/(1/2))^{-3}$，故 $r=3$，$\lambda = 1/2$，即 $X \sim \mathrm{Gamma}(3, \frac{1}{2})$。又 $r = k/2, \lambda = 1/2$ 对应 $k=6$，所以 $X \sim \chi^2_6$。于是 $\mathbb{E}(X) = r/\lambda = 3/(1/2) = 6 = k$，$\mathrm{Var}(X) = r/\lambda^2 = 3/(1/4) = 12 = 2k$。（也可直接 $M'(0), M''(0)$ 验证。）

Example (Self-Test 2：cumulant 求偏度).

设 $X \sim \mathrm{Pois}(\lambda)$，其 MGF $M(t) = \exp\{\lambda(e^t-1)\}$。求 cgf $K(t)$、前三阶 cumulant，并给出偏度 γ_1。

Solution.

$K(t) = \log M(t) = \lambda(e^t-1)$。逐阶求导在 $t=0$：$K'(t) = \lambda e^t \Rightarrow \kappa_1 = \lambda$；$K'' = \lambda e^t \Rightarrow \kappa_2 = \lambda$；$K''' = \lambda e^t \Rightarrow \kappa_3 = \lambda$。事实上 Poisson 的所有 cumulant 都等于 λ。偏度 $\gamma_1 = \kappa_3/\kappa_2^{3/2} = \lambda/\lambda^{3/2} = 1/\sqrt{\lambda}$。直觉验证：$\lambda$ 越大越接近正态（偏度 $\to 0$），与 $\mathrm{Pois}(\lambda) \approx N(\lambda, \lambda)$ 一致。

Example (Self-Test 3：变换得 Uniform（PIT）).

设连续随机变量 X 有严格增 CDF F。证明 $U := F(X) \sim U[0,1]$。再问：若 $V \sim U[0,1]$，$X := F^{-1}(V)$ 的分布是什么？（这是模拟抽样的“逆变换法”。）

Solution.

对 $u \in (0,1)$，$\mathrm{P}(U \le u) = \mathrm{P}(F(X) \le u) = \mathrm{P}\big(X \le F^{-1}(u)\big) = F(F^{-1}(u)) = u$，正是 $U[0,1]$ 的 CDF，故 $U \sim U[0,1]$。反向：$\mathrm{P}(X \le x) = \mathrm{P}\big(F^{-1}(V) \le x\big) = \mathrm{P}(V \le F(x)) = F(x)$，故 X 的 CDF 恰为 F。这两条就是 2020 Q2(b) 用到的概率积分变换及其逆。

Example (Self-Test 4：Chebyshev 推 WLLN 一步).

设 $X_1, \dots, X_n$ i.i.d.，$\mathbb{E}(X_i) = \mu$，$\mathrm{Var}(X_i) = \sigma^2 < \infty$。用 Chebyshev 证明 $\bar X \xrightarrow{p} \mu$。

Solution.

$\mathrm{Var}\big(\bar{X}\big) = \sigma^2/n$。对任意 $\varepsilon > 0$，由 Chebyshev（Theorem 2.17）$\mathrm{P}\big(|\bar{X} - \mu| \geq \varepsilon\big) \leq \frac{\mathrm{Var}\big(\bar{X}\big)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0$（$n \to \infty$）。这正是依概率收敛 $\bar{X} \xrightarrow{p} \mu$ 的定义，即弱大数律。注意只用到二阶矩有限。

Example (Self-Test 5：Jensen 方向判断).

判断下列各式中 $\leq$ 还是 $\geq$（X 非退化、相应矩有限）：(i) $\mathbb{E}\big(e^X\big) \;\square\; e^{\mathbb{E}(X)}$；(ii) $\mathbb{E}(1/X) \;\square\; 1/\mathbb{E}(X)$（$X > 0$）；(iii) $\mathbb{E}\Big(\sqrt{X}\Big) \;\square\; \sqrt{\mathbb{E}(X)}$（$X > 0$）。

Solution.

(i) e^x 凸 $\Rightarrow \mathbb{E}\big(e^X\big) \geq e^{\mathbb{E}(X)}$。(ii) $1/x$ 在 $x > 0$ 上凸 $\Rightarrow \mathbb{E}(1/X) \geq 1/\mathbb{E}(X)$。(iii) $\sqrt{x}$ 凹 $\Rightarrow \mathbb{E}\Big(\sqrt{X}\Big) \leq \sqrt{\mathbb{E}(X)}$。口诀：凸则 $\mathbb{E}(g(X)) \geq g(\mathbb{E}(X))$，凹则反向；非退化时严格。

Chapter 3　随机向量、条件期望与多元正态（Random Vectors, Conditional Expectation & the Multivariate Normal）

导读 · Roadmap（先读这里，不含公式）

上一章把“随机变量”立成了可测函数。这一章把镜头从一维拉到多维：随机向量的联合、边际、条件分布；以及计量经济学真正的主角——条件期望 $\mathbb{E}(Y\mid X)$。要把握一件事：$\mathbb{E}(Y\mid X)$ 不是某个公式，而是“在所有用 X 造的预测里，均方误差最小的那一个”——它是 Y 在“X 的函数”这个空间上的L^2 投影。从这个投影视角，迭代期望定律（LIE / tower）、全方差分解 $\mathrm{Var}(Y)=\mathbb{E}(\mathrm{Var}(Y\mid X))+\mathrm{Var}(\mathbb{E}(Y\mid X))$ 全都顺理成章。

后半章是多元正态，它几乎统治了资格考的“分布题”。正态有几条神性质：线性变换还是正态；分块后边际、条件都是正态，且条件均值是 x_2 的线性函数、条件方差与 x_2 无关——这就是 $\mu_1+\Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2)$ 与 $\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ 那两个必须背到肌肉记忆的公式。再加上一个截断小工具——*inverse Mills ratio*（$\mathbb{E}(Z\mid Z>c)=\phi(c)/(1-\Phi(c))$），就能秒杀一大类“给个单边事件求条件均值”的题。

资格考在本章的固定考法有四种：（1）给联合密度/联合 MGF，识别出（独立标准）正态，再求某个条件期望；（2）分块正态条件分布的推导（2017 原样考过）；（3）带截断事件的条件期望（2024、2025 连考，全靠 Mills 比）；（4）“独立 vs 不相关 vs 条件均值独立”三者关系的反例构造（2020、2022、2023 反复出）。本章逐一拆解并配上原题。

3.1　随机向量：联合、边际与条件

[优先级 ●●○ 中 Med]　资格考足迹：2019, 2020, 2024

一个 $\mathbb{R}^d$ 值的**随机向量** $X=(X_1,\dots,X_d)'$ 就是一个可测映射 $X:\Omega\to\mathbb{R}^d$（等价地，每个分量 X_j 都是随机变量）。它的联合分布由联合 CDF $F_X(x)=\mathrm{P}(X_1\le x_1,\dots,X_d\le x_d)$ 决定；当存在联合密度 f_X 时，$\mathrm{P}(X\in B)=\int_B f_X(x)\,dx$。

Definition 3.1: 边际密度与条件密度 Marginal & conditional density

Let (X, Y) be a continuous random vector on $\mathbb{R}^{d_X} \times \mathbb{R}^{d_Y}$ with joint density $f_{X,Y}$.

(i) The *marginal* density of X is $f_X(x) = \int f_{X,Y}(x, y)\, dy$ (integrate out y).

(ii) For x with $f_X(x) > 0$, the *conditional* density of Y given $X = x$ is

$$f_{Y \mid X}(y \mid x) := \frac{f_{X,Y}(x, y)}{f_X(x)}.$$

条件密度的来历（Andrews 的讲法）：本想照搬"事件"的条件概率 $\mathrm{P}(A \mid B) = \mathrm{P}(A \cap B) / \mathrm{P}(B)$，但连续型里 $\{X = x\}$ 是零概率事件，分母为 0 没法直接除。出路是定义：让 $f_{Y \mid X}(y \mid x) \propto f_{X,Y}(x, y)$，再除以 $f_X(x)$ 使它对 y 积分为 1。如此一来，当 $X \perp Y$（独立）时 $f_{X,Y} = f_X f_Y$，恰好回收 $f_{Y \mid X} = f_Y$——"给定 X 不改变 Y 的分布"，与直觉一致。

Fact 3.2: 条件密度是一支正经密度，且可算条件概率/条件期望

对固定的 x（$f_X(x) > 0$），$y \mapsto f_{Y \mid X}(y \mid x)$ 非负且 $\int f_{Y \mid X}(y \mid x)\, dy = 1$。从而

$$\mathrm{P}(Y \in G \mid X = x) = \int_G f_{Y \mid X}(y \mid x)\, dy, \qquad \mathbb{E}(Y \mid X = x) = \int y\, f_{Y \mid X}(y \mid x)\, dy.$$

Remark（"先求边际，再做条件"是连续型条件题的万能起手式）.

> 凡是题目给你一个支撑古怪的联合密度（如三角形 $0 \le x \le y \le 1$），求 $\mathbb{E}(Y \mid X = x)$，三步走永远一样：(1) 对 y 在正确的积分限上积分得 $f_X(x)$（积分限随支撑变，是最易错处）；(2) 相除得 $f_{Y \mid X}$；(3) 再积分得条件均值。2020 Q1(a)、2024 Q2 都是这个套路，本章 3.2 与 3.3 各演示一次。

3.2 条件期望 = 最佳 L^2 预测（投影）

[优先级 ●●● 高 High] 资格考足迹：2020, 2021, 2024

这是全章的灵魂。$\mathbb{E}(Y \mid X)$ 应被理解为一个随机变量：把 $\mathbb{E}(Y \mid X = x)$ 看成 x 的函数 $g(x)$，再代回 $x = X$，得到的随机变量 $g(X)$ 就记作 $\mathbb{E}(Y \mid X)$。它的关键刻画是：在所有"X 的（可测）函数"里，$\mathbb{E}(Y \mid X)$ 是均方误差最小的那个预测。

Theorem 3.3: 条件期望是最佳 L^2 预测子 (Best predictor / L^2 projection)

[REPRODUCE —memorize this proof]

Let $\mathbb{E}(Y^2) < \infty$. Among all measurable functions $h : \mathbb{R}^{dx} \to \mathbb{R}$ with $\mathbb{E}\big(h(X)^2\big) < \infty$, the choice $g(X) = \mathbb{E}(Y \mid X)$ minimizes the mean squared error:

$$\mathbb{E}\Big([Y - g(X)]^2\Big) \le \mathbb{E}\Big([Y - h(X)]^2\Big) \qquad \text{for every such } h.$$

Equivalently, the prediction error $Y - \mathbb{E}(Y \mid X)$ is orthogonal to every function of X:

$$\mathbb{E}\big(\big(Y - \mathbb{E}(Y \mid X)\big)\, h(X)\big) = 0 \quad \text{for all } h.$$

Proof for Theorem

[REPRODUCE —memorize this proof]

Fix any candidate h and write $g(X) = \mathbb{E}(Y \mid X)$. Add and subtract $g(X)$:

$$\begin{aligned}\mathbb{E}\Big([Y - h(X)]^2\Big) &= \mathbb{E}\left(\left[\underbrace{(Y - g(X))}_{=:u} + \underbrace{(g(X) - h(X))}_{=:r(X)}\right]^2\right) \\ &= \mathbb{E}\big(u^2\big) + 2\,\mathbb{E}(u\, r(X)) + \mathbb{E}\big(r(X)^2\big)\,.\end{aligned}$$

The cross term vanishes. By the law of iterated expectations (Theorem 3.4 below),

$$\mathbb{E}(u\, r(X)) = \mathbb{E}(\mathbb{E}(u\, r(X) \mid X)) = \mathbb{E}(r(X)\, \mathbb{E}(u \mid X))\,,$$

since $r(X)$ is a function of X and pulls out of the inner conditional expectation. But

$$\mathbb{E}(u \mid X) = \mathbb{E}(Y - g(X) \mid X) = \mathbb{E}(Y \mid X) - g(X) = 0 \quad \text{a.s.}$$

Hence the cross term is 0, and $\mathbb{E}\Big([Y - h(X)]^2\Big) = \mathbb{E}\big(u^2\big) + \mathbb{E}\big(r(X)^2\big) \ge \mathbb{E}\big(u^2\big) = \mathbb{E}\Big([Y - g(X)]^2\Big)$, with equality iff $r(X) = 0$ a.s., i.e. $h(X) = g(X)$ a.s. The orthogonality statement is exactly the vanishing cross term with $g(X) = \mathbb{E}(Y \mid X)$. ■

把 L^2 空间想成欧氏空间，“X 的所有函数”是其中一个子空间。$\mathbb{E}(Y \mid X)$ 就是把 Y 垂直投影到这个子空间上的脚——投影到子空间的脚，与子空间里任何向量正交（这正是上面的“正交性”），而且投影脚是子空间里离 Y 最近的点（这正是“最小均方”）。OLS（投影到线性函数空间）只是把这个子空间换成“X 的线性函数”而已，所以 OLS 是 $\mathbb{E}(Y \mid X)$ 的“线性化版本”。

Remark (毕达哥拉斯分解一眼看穿全方差公式).

上面证明顺带给出 $\mathbb{E}\big((Y-h(X))^2\big)=\mathbb{E}\big((Y-\mathbb{E}(Y\,|\,X))^2\big)+\mathbb{E}\big((\mathbb{E}(Y\,|\,X)-h(X))^2\big)$。取 $h\equiv\mathbb{E}(Y)$（常数预测），左边是 $\mathrm{Var}(Y)$，右边第一项是 $\mathbb{E}(\mathrm{Var}(Y\,|\,X))$，第二项是 $\mathrm{Var}(\mathbb{E}(Y\,|\,X))$——这就是 3.3 的全方差分解，一行推完。

3.2.1 Worked example：2020 资格考 Q1(a)（最佳 L^2 预测）

原题 (Econometrics prelim 2020, Q1(a))

Suppose (X,Y) has a joint pdf $f_{XY}(x,y)=8xy\,\mathbb{1}(0\le x\le y\le 1)$. Obtain the function $g:(0,1)\to\mathbb{R}$ such that for any measurable function $h:(0,1)\to\mathbb{R}$,

$$\mathbb{E}\Big([Y-g(X)]^2\Big)\le\mathbb{E}\Big([Y-h(X)]^2\Big).$$

Solution.

由 Theorem 3.3，所求 $g(x)=\mathbb{E}(Y\,|\,X=x)$。先求边际。支撑是 $0\le x\le y\le 1$，故给定 x，y 从 x 跑到 1：

$$f_X(x)=\int_x^1 8xy\,dy=8x\cdot\frac{1-x^2}{2}=4x(1-x^2),\qquad 0<x<1.$$

条件密度

$$f_{Y\,|\,X}(y\,|\,x)=\frac{8xy}{4x(1-x^2)}=\frac{2y}{1-x^2},\qquad x\le y\le 1.$$

故

$$g(x)=\mathbb{E}(Y\,|\,X=x)=\int_x^1 y\cdot\frac{2y}{1-x^2}\,dy=\frac{2}{1-x^2}\cdot\frac{1-x^3}{3}=\frac{2}{3}\cdot\frac{1-x^3}{1-x^2}.$$

（可化简为 $g(x)=\frac{2}{3}\cdot\frac{1+x+x^2}{1+x}$，因为 $1-x^3=(1-x)(1+x+x^2)$、$1-x^2=(1-x)(1+x)$。）

Remark (考点提炼).

答的是“函数 g”，但本质是问 $\mathbb{E}(Y\,|\,X=x)$——一看到“对任意 h 均方误差最小”就该条件反射地写下 $g=\mathbb{E}(Y\,|\,X)$。最易扣分处是积分限：支撑 $0\le x\le y\le 1$ 意味着给定 x 后 $y\ge x$，所以 f_X 与条件期望都从 x 积到 1，不是从 0 到 1。

Example (正交性的直接用法（仿 2024 Q2(d))).

（仿题：真题 2024 Q2(d) 用的是 $f(x,y)=c\,x\,\mathbb{1}(0<x<y<1)$ 配 $\mathbb{1}(X>0.5)$；这里换成上面的 $8xy$ 密度配 $\mathbb{1}(X\le c)$ 练同一招。）延用上题密度。求 Corr 不必算——证明残差 $Y-\mathbb{E}(Y\,|\,X)$ 与任意 $\mathbb{1}(X\le c)$ 不相关。

Solution.

由 Theorem 3.3 的正交性，取 $h(X)=\mathbb{1}(X\le c)$，立得 $\mathbb{E}((Y-\mathbb{E}(Y\mid X))\mathbb{1}(X\le c))=0$；又 $\mathbb{E}(Y-\mathbb{E}(Y\mid X))=0$（取 $h\equiv 1$），故协方差 $=\mathbb{E}((Y-\mathbb{E}(Y\mid X))\mathbb{1}(X\le c))-\mathbb{E}(Y-\mathbb{E}(Y\mid X))\,\mathbb{E}(\mathbb{1}(X\le c))=0$。**要点**：条件均值残差与*任何* X 的函数都不相关——这是“投影脚正交于整个子空间”的直接推论，不需要任何积分。

3.3 迭代期望、条件方差与全方差分解

[优先级 ●●● 高 High] 资格考足迹：2018, 2024

Theorem 3.4: 迭代期望定律 (Law of Iterated Expectations, LIE / tower property)

[REPRODUCE —memorize this proof]

If $\mathbb{E}(|Y|)<\infty$, then

$$\mathbb{E}(Y)=\mathbb{E}(\mathbb{E}(Y\mid X))\,.$$

More generally (tower / smoothing), if $\sigma(X_1)\subseteq\sigma(X_1,X_2)$ then

$$\mathbb{E}(\mathbb{E}(Y\mid X_1,X_2)\mid X_1)=\mathbb{E}(Y\mid X_1)\quad\text{a.s.},$$

and any function of X pulls out: $\mathbb{E}(r(X)\,Y\mid X)=r(X)\,\mathbb{E}(Y\mid X)$ a.s.

证明. [**Proof (headline 式 $\mathbb{E}(Y)=\mathbb{E}(\mathbb{E}(Y\mid X))$，连续型).**] **[REPRODUCE — memorize this proof]**

连续型证明（Andrews）。这里只证 headline 式 $\mathbb{E}(Y)=\mathbb{E}(\mathbb{E}(Y\mid X))$；tower / pull-out 两条是条件期望定义的标准推论，见下方按语。按定义并用 Fubini 换积分次序：

$$\begin{aligned}\mathbb{E}(\mathbb{E}(Y\mid X))&=\int \mathbb{E}(Y\mid X=x)\;f_X(x)\,dx=\iint y\,f_{Y\mid X}(y\mid x)\,dy\,f_X(x)\,dx\\&=\iint y\,\frac{f_{X,Y}(x,y)}{f_X(x)}\,f_X(x)\,dy\,dx=\iint y\,f_{X,Y}(x,y)\,dx\,dy\\&=\int y\Big(\int f_{X,Y}(x,y)\,dx\Big)dy=\int y\,f_Y(y)\,dy=\mathbb{E}(Y)\,.\end{aligned}$$

换序由 Fubini 定理（$\mathbb{E}(|Y|)<\infty$ 保证可积），倒数第二步用边际密度公式。离散与混合情形完全类似。■ □

Remark (tower 与 pull-out 两条为何成立（标准引用结论）).

定理里的 tower $\mathbb{E}(\mathbb{E}(Y\mid X_1,X_2)\mid X_1)=\mathbb{E}(Y\mid X_1)$ 与 pull-out $\mathbb{E}(r(X)Y\mid X)=r(X)\mathbb{E}(Y\mid X)$ 是条件期望（作为在子 σ-域上的投影）定义的标准推论，本讲义直接引用、不逐行证：pull-out 因 $r(X)$ 关于 $\sigma(X)$ 可测、在对 X 取条件时当常数提出；tower 则是“$\sigma(X_1)\subseteq\sigma(X_1,X_2)$ 时，粗 σ-域上的投影吃掉细 σ-域上的投影”。

上面 Theorem 3.3 的 L^2 证明所用的正是 pull-out 这一条。

tower property 的口诀：粗的信息吃掉细的。先对 (X_1, X_2) 取条件（细），再对 X_1 取条件（粗），结果等于直接对 X_1 取条件——多余的 X_2 信息被外层的粗条件"抹平"了。一个特别有用的推论：若某个量已经是 X 的函数，对 X 取条件时它就是常数、原样提到期望外面。

Definition 3.5: 条件方差 Conditional variance

$\mathrm{Var}(Y \mid X) := \mathbb{E}\big((Y - \mathbb{E}(Y \mid X))^2 \mid X\big) = \mathbb{E}\big(Y^2 \mid X\big) - \big(\mathbb{E}(Y \mid X)\big)^2$，本身是 X 的函数（一个随机变量）。

Theorem 3.6: 全（总）方差分解 (Law of total variance / EVVE)

[STRUCTURE —know the steps, fudge details]

If $\mathbb{E}\big(Y^2\big) < \infty$, then

$$\mathrm{Var}(Y) = \underbrace{\mathbb{E}(\mathrm{Var}(Y \mid X))}_{\text{"组内"平均波动}} + \underbrace{\mathrm{Var}(\mathbb{E}(Y \mid X))}_{\text{"组间"条件均值波动}}.$$

Proof for Theorem

[STRUCTURE —know the steps, fudge details]

记 $m(X) = \mathbb{E}(Y \mid X)$。由 LIE，$\mathbb{E}(Y) = \mathbb{E}(m(X))$。拆开：

$$\mathrm{Var}(Y) = \mathbb{E}\big((Y - \mathbb{E}(Y))^2\big) = \mathbb{E}\Big(\big((Y - m(X)) + (m(X) - \mathbb{E}(Y))\big)^2\Big).$$

展开三项。交叉项 $2\mathbb{E}((Y - m(X))(m(X) - \mathbb{E}(Y)))$：先对 X 取条件，$m(X) - \mathbb{E}(Y)$ 提出，$\mathbb{E}(Y - m(X) \mid X) = 0$，故交叉项为 0。第一项 $\mathbb{E}\big((Y - m(X))^2\big) = \mathbb{E}\big(\mathbb{E}\big((Y - m(X))^2 \mid X\big)\big) = \mathbb{E}(\mathrm{Var}(Y \mid X))$（内层正是条件方差）。第二项 $\mathbb{E}\big((m(X) - \mathbb{E}(m(X)))^2\big) = \mathrm{Var}(m(X)) = \mathrm{Var}(\mathbb{E}(Y \mid X))$。相加即得。■

Remark (直觉与一句考场提醒).

把人群按 X 分组：$\mathbb{E}(\mathrm{Var}(Y \mid X))$ 是"各组内部方差的平均"，$\mathrm{Var}(\mathbb{E}(Y \mid X))$ 是"各组均值之间的方差"。总方差 = 组内 + 组间。提醒：外层那个方差作用在随机变量 $\mathbb{E}(Y \mid X)$ 上，不是数；很多人误把 $\mathrm{Var}(\mathbb{E}(Y \mid X))$ 当成 0，那是把它和 $\mathrm{Var}(\mathbb{E}(Y)) = 0$ 搞混了。

3.3.1 Worked example：2018 资格考 Q1(a)（测量误差模型）

原题 (Comprehensive Exam 2018, Guggenberger, Q1(a))

Student i's true performance is $X_i \sim N(\mu, \sigma^2)$. The recorded grade is $Y_i = X_i + u_i + v_i$, where $u_i \sim N(0,1)$ (first grader's measurement error), $v_i \sim N(0,1)$ (second grader's error), and X_i, u_i, v_i are mutually independent. Calculate $X_i^* = \mathbb{E}(X_i \,|\, Y_i)$ and $\mathrm{Var}(X_i \,|\, Y_i)$.

Solution.

(X_i, Y_i) 是 X_i, u_i, v_i 的线性组合，故联合正态（线性变换不变性，见 Theorem 3.10）。算出三件套：

$$\mathbb{E}(X_i) = \mu, \quad \mathbb{E}(Y_i) = \mu; \qquad \mathrm{Var}(X_i) = \sigma^2, \quad \mathrm{Var}(Y_i) = \sigma^2 + 1 + 1 = \sigma^2 + 2;$$

$$\mathrm{Cov}(X_i, Y_i) = \mathrm{Cov}(X_i, X_i + u_i + v_i) = \mathrm{Var}(X_i) = \sigma^2 \quad (\text{因} X_i \perp u_i, v_i).$$

套用标量版正态条件公式（Theorem 3.11 的一维特例 $\mathbb{E}(X \,|\, Y) = \mathbb{E}(X) + \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)}(Y - \mathbb{E}(Y))$）：

$$\boxed{X_i^* = \mathbb{E}(X_i \,|\, Y_i) = \mu + \frac{\sigma^2}{\sigma^2 + 2}\,(Y_i - \mu)}$$

$$\mathrm{Var}(X_i \,|\, Y_i) = \mathrm{Var}(X_i) - \frac{\mathrm{Cov}(X_i, Y_i)^2}{\mathrm{Var}(Y_i)} = \sigma^2 - \frac{\sigma^4}{\sigma^2 + 2} = \frac{2\sigma^2}{\sigma^2 + 2}.$$

Remark (考点提炼).

两记号要点：(1) 条件均值是 Y_i 的线性函数，系数 $\frac{\sigma^2}{\sigma^2+2} \in (0,1)$ 是"信噪比"——真信号方差占总方差的比例，把含噪记录 Y_i 朝先验均值 μ 收缩（shrinkage）。(2) 条件方差不依赖于 Y_i 的取值，这是正态独有的便利。这里 u_i, v_i 各贡献 1 单位噪声，所以分母是 $\sigma^2 + 2$ 而非 $\sigma^2 + 1$——漏掉一个 grader 是常见失分点。

3.3.2 Worked example：2024 资格考 Q2(c)（嵌套条件 / tower）

原题 (ECON 501 Econometrics 2024, Sung Jae Jun, Q2(c))

For the joint density $f(x,y) = c \cdot x \cdot \mathbb{1}(0 < x < y < 1)$, characterize the probability distribution of

$$Z := \mathbb{E}\big(\,\mathbb{E}\big(\,\mathbb{E}(Y \,|\, \mathbb{1}(X \le 0.5)) \,\big|\, \mathbb{1}(X \le 0.5), X^2\big) \,\big|\, X\,\big).$$

Solution.

逐层用 tower（Theorem 3.4）从里往外剥。

最内层 $W := \mathbb{E}(Y \mid \mathbb{1}(X \le 0.5))$ 只依赖二值随机变量 $D := \mathbb{1}(X \le 0.5)$，因此 W 本身只取两个值：在 $\{X \le 0.5\}$ 上取 $a := \mathbb{E}(Y \mid X \le 0.5)$，在 $\{X > 0.5\}$ 上取 $b := \mathbb{E}(Y \mid X > 0.5)$。

中间层：对 (D, X^2) 取条件。但 W 已经是 D 的函数，而 D 是 (D, X^2) 可测的，所以 $\mathbb{E}\big(W \mid D, X^2\big) = W$（已知量原样不变）。

最外层：对 X 取条件。$W = \mathbb{1}(X \le 0.5)\, a + \mathbb{1}(X > 0.5)\, b$ 是 X 的函数，故 $\mathbb{E}(W \mid X) = W$。

于是 $Z = W = \mathbb{E}(Y \mid \mathbb{1}(X \le 0.5))$，是一个两点（离散）随机变量：

$$Z = \begin{cases} a = \mathbb{E}(Y \mid X \le 0.5), & \text{w.p. } p := \mathrm{P}(X \le 0.5), \\ b = \mathbb{E}(Y \mid X > 0.5), & \text{w.p. } 1 - p. \end{cases}$$

（题目只要“characterize the distribution”，到此即可：两个原子，概率分别为 $\mathrm{P}(X \le 0.5)$ 与 $\mathrm{P}(X > 0.5)$。若要数值，先定常数 c：$\int_0^1 \int_0^y cx\, dx\, dy = \int_0^1 c\frac{y^2}{2} dy = \frac{c}{6} = 1 \Rightarrow c = 6$；则 $f_X(x) = \int_x^1 6x\, dy = 6x(1-x)$，$p = \int_0^{1/2} 6x(1-x) dx = \frac{1}{2}$。）

Remark (考点提炼).

> 别套 tower 的“粗吃细”：这里条件 σ-域其实是越条件越细（$\sigma(D) \subset \sigma(D, X^2) \subset \sigma(X)$），不存在“被粗外层抹掉”。真正的机制是可测性 / *pull-out*：最内层结果 W 已经是 D 的函数，因而对更细的 (D, X^2)、再对 X 都可测——把一个已可测的量再拿去取条件，原样吐回。关键判断：一旦某个量已是当前条件变量的函数，条件期望就把它当常数提出。答案是离散分布——别被三层期望唬住去做积分。

3.4 协方差、相关与 Cauchy–Schwarz

[优先级 ●●○ 中 Med] 资格考足迹：2020, 2021, 2023

Definition 3.7: 协方差矩阵与相关 Covariance matrix & correlation

For random vectors $X \in \mathbb{R}^n$, $Y \in \mathbb{R}^m$,

$$\mathrm{Var}(X) = \mathbb{E}\big((X - \mathbb{E}(X))(X - \mathbb{E}(X))'\big) \ \text{(psd)},$$
$$\mathrm{Cov}(X, Y) = \mathbb{E}\big((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))'\big) \in \mathbb{R}^{n \times m}.$$

Linear rules: $\operatorname{Var}(AX+b) = A\operatorname{Var}(X)A'$ and $\operatorname{Cov}(AX+c, DY+g) = A\operatorname{Cov}(X,Y)D'$. For scalars,

$$\rho_{XY} = \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)}\sqrt{\operatorname{Var}(Y)}} \in [-1,1],$$

invariant to affine rescaling $X \mapsto aX+b$, $Y \mapsto cY+d$ with $a, c > 0$.

Theorem 3.8: Cauchy–Bunyakovsky–Schwarz (CBS) 不等式

[REPRODUCE —memorize this proof]

For any random variables X, Y with finite second moments,

$$|\mathbb{E}(XY)| \le \big(\mathbb{E}\big(X^2\big)\big)^{1/2}\big(\mathbb{E}\big(Y^2\big)\big)^{1/2},$$

with equality iff X, Y are linearly dependent, i.e. $Y = aX$ a.s. or $X = aY$ a.s. for some constant a. Consequently $|\rho_{XY}| \le 1$.

Proof for Theorem

[REPRODUCE —memorize this proof]

Andrews 的“线性投影”证法。设 $a = \mathbb{E}(XY)/\mathbb{E}\big(X^2\big)$（若 $\mathbb{E}\big(X^2\big)=0$ 则 $X=0$ a.s.，平凡），写

$$Y = (Y - aX) + aX.$$

两边平方取期望：$\mathbb{E}\big(Y^2\big) = \mathbb{E}\big((Y-aX)^2\big) + 2a\,\mathbb{E}(X(Y-aX)) + a^2\mathbb{E}\big(X^2\big)$。交叉项

$$2a\,\mathbb{E}(X(Y-aX)) = 2a\big(\mathbb{E}(XY) - a\mathbb{E}\big(X^2\big)\big) = 2a\Big(\mathbb{E}(XY) - \tfrac{\mathbb{E}(XY)}{\mathbb{E}(X^2)}\mathbb{E}\big(X^2\big)\Big) = 0.$$

故 $\mathbb{E}\big(Y^2\big) - a^2\mathbb{E}\big(X^2\big) = \mathbb{E}\big((Y-aX)^2\big) \ge 0$，即 $\mathbb{E}\big(Y^2\big) \ge \dfrac{(\mathbb{E}(XY))^2}{\mathbb{E}(X^2)}$，移项即得。等号当且仅当 $\mathbb{E}\big((Y-aX)^2\big)=0$，即 $Y = aX$ a.s.。把 X, Y 换成 $X-\mathbb{E}(X)$，$Y-\mathbb{E}(Y)$ 即得 $|\rho_{XY}| \le 1$。■

CBS 的证法本身就是“把 Y 沿 X 方向做投影、留下垂直残差 $Y-aX$，残差平方非负”——和 Theorem 3.3 是同一招的线性版。后面分块正态条件分布的推导（Theorem 3.11）用的 $X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2$ 也正是这个“减掉投影、留正交残差”的多维翻版——Andrews 明说这是 CBS 分解的多维版本。

3.5 多元正态分布

[优先级 ●●● 高 High] 资格考足迹：2017, 2020, 2021, 2024, 2025

资格考最常考的分布。先立密度与两条神性质，再推分块条件分布。

Definition 3.9: 多元正态 Multivariate normal $N(\mu,\Sigma)$

Let $\mu\in\mathbb{R}^n$ and Σ be symmetric positive definite $n\times n$. $X\sim N(\mu,\Sigma)$ if it has density

$$f(x)=\frac{1}{(2\pi)^{n/2}}(\det\Sigma)^{-1/2}\exp\Big(-\tfrac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\Big),\quad x\in\mathbb{R}^n.$$

The distribution is determined entirely by its mean $\mu=\mathbb{E}(X)$ and variance $\Sigma=\mathrm{Var}(X)$. Equivalently, $X\sim N(\mu,\Sigma)$ iff its MGF is $M_X(t)=\mathbb{E}\Big(e^{t'X}\Big)=\exp\big(t'\mu+\tfrac{1}{2}t'\Sigma t\big)$ for all $t\in\mathbb{R}^n$.

Theorem 3.10: 标准化表示 + 线性变换不变性

[STRUCTURE —know the steps, fudge details]

(i) If $X\sim N(\mu,\Sigma)$ with Σ pd, then $Z=\Sigma^{-1/2}(X-\mu)\sim N(0,I_n)$, equivalently $X=\Sigma^{1/2}Z+\mu$ with $Z\sim N(0,I_n)$.
(ii) (*Linearity*) For any $k\times n$ constant A and k-vector b (with $A\Sigma A'$ pd), $AX+b\sim N(A\mu+b,\ A\Sigma A')$. In particular every linear combination / sub-vector of a multivariate normal is again normal.

Proof for Theorem

[STRUCTURE —know the steps, fudge details]

(i) 谱分解 $\Sigma=B\Lambda B'$（B 正交、Λ 对角且对角元为正特征值），定义 $\Sigma^{1/2}=B\Lambda^{1/2}B'$、$\Sigma^{-1/2}=B\Lambda^{-1/2}B'$，则 $\Sigma^{-1/2}\Sigma^{-1/2}=\Sigma^{-1}$、$\Sigma^{1/2}\Sigma^{1/2}=\Sigma$。令 $Z=\Sigma^{-1/2}(X-\mu)$，由换元公式（雅可比 $(\det\Sigma)^{-1/2}$）得 $f_Z(z)=(2\pi)^{-n/2}\exp(-\frac{1}{2}z'z)=\prod_i\frac{1}{\sqrt{2\pi}}e^{-z_i^2/2}$，即 $Z\sim N(0,I_n)$，分量独立标准正态。(ii) 由 (i) 写 $X=\Sigma^{1/2}Z+\mu$，则 $AX+b=(A\Sigma^{1/2})Z+(A\mu+b)$ 仍是 Z 的仿射变换；用 MGF 最快：$\mathbb{E}\Big(e^{t'(AX+b)}\Big)=e^{t'b}\mathbb{E}\Big(e^{(A't)'X}\Big)=\exp\big(t'(A\mu+b)+\frac{1}{2}t'A\Sigma A't\big)$，正是 $N(A\mu+b,A\Sigma A')$ 的 MGF。■

Remark (由 MGF 反认正态——资格考的高频开场).

若题目给 $\mathbb{E}\big(e^{tX+sY}\big)=\exp\{(t^2+s^2)/2\}$，比对 $N(\mu,\Sigma)$ 的 MGF $\exp(t'\mu+\frac{1}{2}t'\Sigma t)$：均值 0、$\Sigma=I_2$，即 $X,Y\overset{iid}{\sim}N(0,1)$（MGF 可乘 $\Rightarrow$ 独立）。这一步是 2020 Q3、2021 Q2、2024 Q4 的统一入口。**切记**：MGF 唯一决定分布，认出形状就能直接写下联合正态，无需再算密度。

Theorem 3.11: 分块多元正态的边际与条件分布

[REPRODUCE —memorize this proof]

Let $X = (X_1', X_2')' \sim N(\mu, \Sigma)$ with $X_i \in \mathbb{R}^{n_i}$, and partition

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Marginals: $X_1 \sim N(\mu_1, \Sigma_{11})$ and $X_2 \sim N(\mu_2, \Sigma_{22})$; moreover $X_1 \perp X_2$ iff $\Sigma_{12} = 0$.

Conditional: the conditional law of X_1 given $X_2 = x_2$ is normal,

$$X_1 \,|\, X_2 = x_2 \;\sim\; N\big(\mu_{1\cdot 2},\ \Sigma_{1\cdot 2}\big), \qquad \boxed{\begin{aligned} \mu_{1\cdot 2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \\ \Sigma_{1\cdot 2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \end{aligned}}$$

The conditional mean is *linear* in x_2; the conditional variance does *not* depend on x_2 and satisfies $\Sigma_{1\cdot 2} \preceq \Sigma_{11}$ (psd order).

Proof for Theorem

[REPRODUCE —memorize this proof]

边际：$X_1 = [\,I_{n_1} \;\; 0\,]X$ 是 X 的线性变换，由 Theorem 3.10(ii) 即 $N(\mu_1, \Sigma_{11})$。独立判据：联合正态时 $X_1 \perp X_2 \iff \Sigma_{12} = 0$——“仅当”因 Σ_{12} 就是协方差、独立必零协方差；“当”因 $\Sigma_{12} = 0$ 时联合密度因子化为两支边际密度之积（$\det\Sigma = \det\Sigma_{11}\det\Sigma_{22}$，二次型也分块拆开），即独立。

条件分布（Andrews 的“减掉投影”法）。作分解

$$X_1 = \underbrace{(X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2)}_{=:R} + \Sigma_{12}\Sigma_{22}^{-1}X_2.$$

(R, X_2) 是 X 的线性变换，故联合正态。算 R 与 X_2 的协方差：

$$\mathrm{Cov}(R, X_2) = \mathrm{Cov}(X_1, X_2) - \Sigma_{12}\Sigma_{22}^{-1}\,\mathrm{Cov}(X_2, X_2) = \Sigma_{12} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22} = 0.$$

联合正态 + 零协方差 $\Rightarrow R \perp X_2$。因此“给定 $X_2 = x_2$，R 的条件分布 $= R$ 的无条件分布”，即正态，均值

$$\mathbb{E}(R) = \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2,$$

方差

$$\begin{aligned}\operatorname{Var}(R) &= \operatorname{Var}(X_1) + \Sigma_{12}\Sigma_{22}^{-1}\operatorname{Var}(X_2)\,\Sigma_{22}^{-1}\Sigma_{21} \\ &\quad - \operatorname{Cov}(X_1, X_2)\,\Sigma_{22}^{-1}\Sigma_{21} - \Sigma_{12}\Sigma_{22}^{-1}\operatorname{Cov}(X_2, X_1) \\ &= \Sigma_{11} + \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \\ &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} =: \Sigma_{1\cdot 2}.\end{aligned}$$

而第二个加项 $\Sigma_{12}\Sigma_{22}^{-1}X_2$ 在给定 $X_2 = x_2$ 时是常数 $\Sigma_{12}\Sigma_{22}^{-1}x_2$，只平移条件分布。两段相加：

$$\begin{aligned}X_1 \,|\, X_2 = x_2 &\sim N\big(\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2 + \Sigma_{12}\Sigma_{22}^{-1}x_2,\ \Sigma_{1\cdot 2}\big) \\ &= N\big(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2),\ \Sigma_{1\cdot 2}\big).\end{aligned}$$

最后 $\Sigma_{11} - \Sigma_{1\cdot 2} = \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$：记 $M := \Sigma_{12}\Sigma_{22}^{-1}$，则因 $\Sigma_{21} = \Sigma_{12}'$、$\Sigma_{22}^{-1}\Sigma_{22}\Sigma_{22}^{-1} = \Sigma_{22}^{-1}$，有 $\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = M\Sigma_{22}M' \succeq 0$（正定阵 Σ_{22} 的合同变换），故条件方差 $\preceq$ 无条件方差——条件化只会减少不确定性。■

这两个公式必须背到肌肉记忆。记忆法：把 Σ_{22}^{-1} 想成“先把 x_2 的尺度标准化”，Σ_{12} 是 X_1 与 X_2 的协方差，于是 $\Sigma_{12}\Sigma_{22}^{-1}$ 就是“X_1 对 X_2 的回归系数”——它正是 X_1 对 X_2 的总体 *OLS* 斜率！所以在正态世界里，$\mathbb{E}(X_1 \,|\, X_2)$ 恰好等于线性投影 X_1 on X_2；条件期望 = 最佳线性预测。一维特例 $\mathbb{E}(X \,|\, Y) = \mathbb{E}(X) + \frac{\operatorname{Cov}(X,Y)}{\operatorname{Var}(Y)}(Y - \mathbb{E}(Y))$、$\operatorname{Var}(X \,|\, Y) = \operatorname{Var}(X) - \frac{\operatorname{Cov}(X,Y)^2}{\operatorname{Var}(Y)}$，就是 2018 测量误差题用的那套。

3.5.1 Worked example：2017 资格考 Q5（分块正态条件分布的推导）

原题 (Comprehensive Exam 2017, Q5)

Assume $X = (X_1', X_2')' \sim N(\mu, \Sigma)$, where $\mu = (\mu_1', \mu_2')'$ and Σ is partitioned with blocks $\Sigma_{11}, \Sigma_{12}, \Sigma_{21}, \Sigma_{22}$, and $X_i \in \mathbb{R}^{p_i}$ for $i = 1, 2$. Derive the conditional distribution of X_1 given $X_2 = x_2$. (You may use that linear transformations of X are again normal, and that two jointly normal random vectors are independent iff they have zero covariance.)

Solution.

这正是 Theorem 3.11 的条件分布部分，给定的两条“可用事实”精确对应证明所需的两块工具。完整推导见上面的证明；考场上把它默写出来即可，结论：

$$X_1 \,|\, X_2 = x_2 \sim N\big(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2),\ \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\big).$$

答题模板（分块正态条件分布，15 分）

第一步（构造正交残差）：写 $X_1 = R + \Sigma_{12}\Sigma_{22}^{-1}X_2$，其中 $R = X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2$。
第二步（联合正态）：$(R', X_2')'$ 是 X 的线性变换 $\begin{pmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{pmatrix} X$，故联合正态（用“事实一”）。
第三步（零协方差 $\Rightarrow$ 独立）：$\mathrm{Cov}(R, X_2) = \Sigma_{12} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22} = 0$，由“事实二”得 $R \perp X_2$。
第四步（条件 = 无条件 + 平移）：给定 $X_2 = x_2$，R 分布不变 $= N(\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2,\ \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$，加上常数 $\Sigma_{12}\Sigma_{22}^{-1}x_2$ 平移均值，得结论。

3.6 截断条件期望与 inverse Mills ratio

[优先级 ●●● 高 High] 资格考足迹：2024, 2025

最后一块工具，专门对付“给定一个单边事件 $\{Z > c\}$ 或 $\{Z < c\}$，求条件期望”的题。

Lemma 3.12: 标准正态截断均值 (Truncated mean / inverse Mills ratio)

[REPRODUCE —memorize this proof]

Let $Z \sim N(0,1)$ with density ϕ and cdf Φ. Then for any threshold c,

$$\mathbb{E}(Z \mid Z > c) = \frac{\phi(c)}{1 - \Phi(c)}, \qquad \mathbb{E}(Z \mid Z < c) = -\frac{\phi(c)}{\Phi(c)}.$$

The quantity $\lambda(c) := \phi(c)/(1 - \Phi(c))$ is the *inverse Mills ratio*. For $\sigma > 0$, a centered normal $Y = \sigma Z \sim N(0, \sigma^2)$ gives

$$\mathbb{E}(Y \mid Y > c) = \sigma\,\frac{\phi(c/\sigma)}{1 - \Phi(c/\sigma)}, \qquad \mathbb{E}(Y \mid Y < 0) = -\sigma\sqrt{\tfrac{2}{\pi}}.$$

Proof for Theorem

[REPRODUCE —memorize this proof]

核心是 $\phi'(z) = -z\phi(z)$，即 $z\phi(z) = -\phi'(z)$。于是

$$\mathbb{E}(Z \mid Z > c) = \frac{\int_c^\infty z\phi(z)\,dz}{\mathrm{P}(Z > c)} = \frac{\int_c^\infty -\phi'(z)\,dz}{1 - \Phi(c)} = \frac{[-\phi(z)]_c^\infty}{1 - \Phi(c)} = \frac{\phi(c)}{1 - \Phi(c)},$$

因 $\phi(\infty) = 0$。下截断同理：$\mathbb{E}(Z \mid Z < c) = \dfrac{\int_{-\infty}^c z\phi(z)dz}{\Phi(c)} = \dfrac{[-\phi(z)]_{-\infty}^c}{\Phi(c)} = -\dfrac{\phi(c)}{\Phi(c)}$。
取 $c = 0$：$\phi(0) = 1/\sqrt{2\pi}$、$\Phi(0) = 1/2$，得 $\mathbb{E}(Z \mid Z < 0) = -2\phi(0) = -\sqrt{2/\pi}$。尺度版 $Y = \sigma Z$ 时事件 $\{Y < 0\} = \{Z < 0\}$，故 $\mathbb{E}(Y \mid Y < 0) = \sigma\mathbb{E}(Z \mid Z < 0) = -\sigma\sqrt{2/\pi}$。■

记忆钩子：$\mathbb{E}(Z \mid Z > c) = \phi(c)/(1-\Phi(c))$ 永远为正且大于 c（你截掉了左半，剩下的当然更靠右）。半正态特例 $\mathbb{E}(Z \mid Z > 0) = \phi(0)/\Phi(0) = \sqrt{2/\pi} \approx 0.798$。Heckman 选择模型里的"逆米尔斯比修正项"就是这个 $\lambda(\cdot)$，所以这个工具在第二门课（510）还会重逢。

3.6.1 Worked example：2025 资格考 Q2（截断条件期望）

原题 (Comprehensive Exam 2025, Pinkse, Q2)

Suppose that $\mathbf{x}, \mathbf{y}$ are jointly normal with mean zero, unit variance, and correlation ρ for which $\rho^2 = 3/4$. Compute $\mathbb{E}(\mathbf{y} \mid \mathbf{y} \le \mathbf{x},\ \mathbf{x} = 0)$.

Solution.

先用条件"$\mathbf{x} = 0$"把二维问题降成一维。由分块正态条件公式（Theorem 3.11，一维），

$$\mathbf{y} \mid \mathbf{x} = 0 \ \sim\ N\Big(0 + \rho \cdot 1 \cdot (0-0),\ 1-\rho^2\Big) = N(0,\ 1-\rho^2).$$

在 $\mathbf{x} = 0$ 上，事件 $\{\mathbf{y} \le \mathbf{x}\}$ 变成 $\{\mathbf{y} \le 0\}$。于是要算

$$\mathbb{E}(\mathbf{y} \mid \mathbf{y} \le 0,\ \mathbf{x} = 0) = \mathbb{E}(Y \mid Y < 0), \quad Y \sim N(0, 1-\rho^2).$$

代 $\sigma = \sqrt{1-\rho^2}$ 进 Lemma 3.12 的下截断 $\mathbb{E}(Y \mid Y < 0) = -\sigma\sqrt{2/\pi}$。这里 $\rho^2 = 3/4 \Rightarrow 1-\rho^2 = 1/4 \Rightarrow \sigma = 1/2$：

$$\boxed{\mathbb{E}(\mathbf{y} \mid \mathbf{y} \le \mathbf{x},\ \mathbf{x} = 0) = -\tfrac{1}{2}\sqrt{\tfrac{2}{\pi}} = -\frac{1}{\sqrt{2\pi}} \approx -0.399.}$$

Remark (考点提炼).

> 两步法：(1) 先条件化到 $\mathbf{x} = 0$ 得 $\mathbf{y} \sim N(0, 1-\rho^2)$——注意 ρ 的符号不影响答案，因为只用到 ρ^2，条件均值那项 $\rho \cdot (0-0) = 0$。(2) 再对降维后的正态用截断均值。陷阱：把 $\{\mathbf{y} \le \mathbf{x}\}$ 误当 $\{\mathbf{y} \le \mathbb{E}(\mathbf{x})\}$ 之外的复杂事件——其实代入 $\mathbf{x} = 0$ 后它就是 $\{\mathbf{y} \le 0\}$，干净利落。

3.6.2 Worked example：2020 Q3 与 2021 Q2（联合 MGF → 正态 → 截断条件期望）

原题 (Econometrics prelim 2020, Q3 = Spring 2021, Q2)

Suppose X and Y are random variables with $\mathbb{E}(\exp(tX + sY)) = \exp\{(t^2 + s^2)/2\}$ for all $t, s \in \mathbb{R}$. Compute $\mathbb{E}(2X + Y \mid X + 2Y > 0)$.

Solution.

第一步：认正态。比对 $N(\mu,\Sigma)$ 的 MGF，$\exp\{(t^2+s^2)/2\}=\exp\{0+\frac{1}{2}(t^2+s^2)\}$ 给出 $\mu=0$、$\Sigma=I_2$，即 $X,Y\overset{iid}{\sim}N(0,1)$（MGF 因子化 $\Rightarrow$ 独立）。

第二步：造两个线性组合。令 $A=2X+Y$、$B=X+2Y$。两者联合正态（线性变换），均值 0，且

$$\mathrm{Var}(A)=4\,\mathrm{Var}(X)+\mathrm{Var}(Y)=5,\quad \mathrm{Var}(B)=\mathrm{Var}(X)+4\,\mathrm{Var}(Y)=5,$$

$$\mathrm{Cov}(A,B)=\mathrm{Cov}(2X+Y,X+2Y)=2\cdot1+1\cdot2=4$$

（其中用到 $\mathrm{Var}(X)=\mathrm{Var}(Y)=1$、$\mathrm{Cov}(X,Y)=0$。）

第三步：投影 + 截断。 A 对 B 的条件期望（正态条件公式）：$\mathbb{E}(A\,|\,B)=\mathbb{E}(A)+\frac{\mathrm{Cov}(A,B)}{\mathrm{Var}(B)}(B-\mathbb{E}(B))=\frac{4}{5}B$。再用 tower：

$$\mathbb{E}(A\,|\,B>0)=\mathbb{E}(\mathbb{E}(A\,|\,B)\,|\,B>0)=\tfrac{4}{5}\mathbb{E}(B\,|\,B>0)\,.$$

$B\sim N(0,5)$，由 Lemma 3.12（尺度 $\sigma=\sqrt{5}$，下截断换上截断、$c=0$）：$\mathbb{E}(B\,|\,B>0)=\sqrt{5}\cdot\sqrt{2/\pi}=\sqrt{10/\pi}$。故

$$\boxed{\mathbb{E}(2X+Y\,|\,X+2Y>0)=\frac{4}{5}\sqrt{\frac{10}{\pi}}=4\sqrt{\frac{2}{5\pi}}\approx1.427.}$$

Remark (考点提炼).

> 这是“MGF 认正态 → 线性组合联合正态 → 把被求量投影到条件变量上 → 截断均值”的四步标准流程，几乎每隔一两年考一次（2020、2021、2024 同源）。*关键技巧*：不要去算 A 在 $\{B>0\}$ 上的复杂积分；而是先把 A 沿 B 投影成 $\frac{4}{5}B$ + 与 B 独立的残差，残差在条件 $\{B>0\}$ 下均值仍为 0，只剩 $\frac{4}{5}\mathbb{E}(B\,|\,B>0)$。

3.6.3 Worked example：2024 Q4(a)(b)（条件密度 + 单边截断 / Mills）

原题 (ECON 501 Econometrics 2024, Sung Jae Jun, Q4)

Suppose X,Y have $\mathbb{E}(\exp(tX+sY))=\exp\{(t^2+s^2)/2\}$ for all $t,s\in\mathbb{R}$. (a) Obtain the conditional density of Y given $2X+Y$. (b) Obtain $\mathbb{E}(Y\,|\,2X+Y<0)$.

Solution.

同上认出 $X,Y\overset{iid}{\sim}N(0,1)$。令 $W=2X+Y$。

(a) (Y,W) 联合正态，$\mathbb{E}(Y)=\mathbb{E}(W)=0$，

$$\mathrm{Var}(Y)=1,\qquad \mathrm{Var}(W)=4\,\mathrm{Var}(X)+\mathrm{Var}(Y)=5,$$
$$\mathrm{Cov}(Y,W)=\mathrm{Cov}(Y,2X+Y)=\mathrm{Var}(Y)=1.$$

由正态条件公式，$Y\,|\,W=w \sim N\big(\frac{\mathrm{Cov}(Y,W)}{\mathrm{Var}(W)}w,\ \mathrm{Var}(Y)-\frac{\mathrm{Cov}(Y,W)^2}{\mathrm{Var}(W)}\big)=N\big(\frac{1}{5}w,\ 1-\frac{1}{5}\big)=N\big(\frac{w}{5},\frac{4}{5}\big)$。故条件密度

$$f_{Y\,|\,W}(y\,|\,w)=\frac{1}{\sqrt{2\pi\cdot\frac{4}{5}}}\exp\Big(-\frac{(y-w/5)^2}{2\cdot\frac{4}{5}}\Big)=\sqrt{\frac{5}{8\pi}}\exp\Big(-\frac{5}{8}\big(y-\tfrac{w}{5}\big)^2\Big).$$

(b) 用 tower + 投影：$\mathbb{E}(Y\,|\,W)=\frac{1}{5}W$，故

$$\mathbb{E}(Y\,|\,W<0)=\tfrac{1}{5}\mathbb{E}(W\,|\,W<0),\quad W\sim N(0,5).$$

下截断（Lemma 3.12，$\sigma=\sqrt{5}$）：$\mathbb{E}(W\,|\,W<0)=-\sqrt{5}\sqrt{2/\pi}=-\sqrt{10/\pi}$。于是

$$\boxed{\mathbb{E}(Y\,|\,2X+Y<0)=-\frac{1}{5}\sqrt{\frac{10}{\pi}}=-\sqrt{\frac{2}{5\pi}}\approx-0.357.}$$

Remark (考点提炼).

(a) 是把分块正态条件公式当“求条件密度”用，记得写完整的密度形式（均值 $w/5$、方差 $4/5$）。(b) 与上一题是镜像，只是把 $\{W>0\}$ 换成 $\{W<0\}$，答案变号、且系数 $1/5$ 而非 $4/5$（因为现在被求量是 Y、投影系数 $\mathrm{Cov}(Y,W)\,/\,\mathrm{Var}(W)=1/5$）。

3.7 独立 vs 不相关 vs 条件均值独立

[优先级 ●●● 高 High] 资格考足迹：2020, 2022, 2023

三个强弱不同的“无关性”概念，资格考极爱考它们的蕴含关系与**反例构造**。先把链条理清。

Definition 3.13: 三种无关性 Three notions

对随机变量 X,Y（二阶矩有限）：

(i) **独立 (independent)**：$f_{X,Y}=f_Xf_Y$，等价于对一切 g,h 有 $\mathbb{E}(g(X)h(Y))=\mathbb{E}(g(X))\,\mathbb{E}(h(Y))$。

(ii) **条件均值独立 (conditional mean independence, CMI)**：$\mathbb{E}(Y\,|\,X)=\mathbb{E}(Y)$ a.s.（$\mathbb{E}(Y\,|\,X)$ 不依赖 X）。

(iii) **不相关 (uncorrelated)**：$\mathrm{Cov}(X,Y)=0$，即 $\mathbb{E}(XY)=\mathbb{E}(X)\,\mathbb{E}(Y)$。

Proposition 3.14: 蕴含链：独立 ⇒ CMI ⇒ 不相关（反向均不成立）

[STRUCTURE —know the steps, fudge details]

独立 ⇒ (Y 关于 X) CMI ⇒ (X,Y) 不相关。一般而言三个箭头都不可逆；唯有当 (X,Y) 联合正态时，三者等价（此时不相关 ⇒ 独立）。

Proof for Theorem

[STRUCTURE —know the steps, fudge details]

独立 ⇒ CMI：独立时 $f_{Y\,|\,X}=f_Y$，故 $\mathbb{E}(Y\,|\,X)=\int y f_Y(y)dy=\mathbb{E}(Y)$，与 X 无关。**CMI ⇒ 不相关**：若 $\mathbb{E}(Y\,|\,X)=\mathbb{E}(Y)$（常数），则由 LIE

$$\mathbb{E}(XY)=\mathbb{E}(X\,\mathbb{E}(Y\,|\,X))=\mathbb{E}(X\cdot\mathbb{E}(Y))=\mathbb{E}(X)\,\mathbb{E}(Y)\;\Rightarrow\;\mathrm{Cov}(X,Y)=0.$$

联合正态下不相关 ⇒ 独立：$\mathrm{Cov}(X,Y)=0$ 时 Σ 对角，密度因子化（Theorem 3.11 的独立判据），故独立。反例见下面 3.7.1 与 3.7.2：可构造“CMI 但不独立”、“不相关但非 CMI”。■

一张图记牢：独立 ⊊ CMI ⊊ 不相关（越往右越弱、范围越大）。当心方向性：CMI 是不对称的——“Y 关于 X 条件均值独立”不蕴含“X 关于 Y 条件均值独立”。独立和不相关则是对称的。只有跳进联合正态的世界，三圈才坍缩成同一圈。

3.7.1 Worked example：2020 Q7（构造三种反例）

原题 (Econometrics prelim 2020, Q7)

Create a joint distribution of two random variables x, y with finite variances for each of: (a) x, y are independent; (b) y is conditionally mean independent of x but x, y are *not* independent; (c) y and x are uncorrelated but y is *not* conditionally mean independent of x.

Solution.

(a) 独立：取 $x,y\overset{iid}{\sim}N(0,1)$（或独立 Bernoulli(1/2)）。联合密度 = 边际之积，定义即独立。

(b) CMI 但不独立（异方差型）：取 $x\sim N(0,1)$，$e\sim N(0,1)$ 与 x 独立，令 $y=x\,e$。则

$$\mathbb{E}(y\,|\,x)=x\,\mathbb{E}(e\,|\,x)=x\cdot\mathbb{E}(e)=0\quad\text{(常数，故 CMI)},$$

但 $\mathrm{Var}(y\,|\,x)=x^2\,\mathrm{Var}(e)=x^2$ 依赖 x，故 x,y 不独立（条件分布随 x 变化）。

(c) 不相关但非 CMI：取 $x\sim N(0,1)$，$y=x^2-1$。则

$$\mathrm{Cov}(x,y)=\mathbb{E}\big(x(x^2-1)\big)=\mathbb{E}\big(x^3\big)-\mathbb{E}(x)=0-0=0\quad\text{(对称分布三阶矩为0, 不相关)},$$

但 $\mathbb{E}(y\,|\,x)=x^2-1$ 显然依赖 x，故非 CMI。

Remark (考点提炼).

> 两个“主力反例”值得背死：(b) 乘积型 $y=x\cdot e$ 给出“均值独立但异方差 $\Rightarrow$ 非独立”；(c) 平方型 $y=x^2-\mathbb{E}(x^2)$ 给出“不相关但条件均值随 x 变”。它们恰好卡在蕴含链 Proposition 3.14 的两个“不可逆”箭头上。注意 (b)(c) 都要 x 对称（奇阶矩为 0）才干净。

Example (2022 Q4(f)（判断题：相互条件均值零 $\iff$ 不相关?）).

原题 (2022 Part 501 Q4(f)): True or false: Let X,Y be mean-zero. Then “$\mathbb{E}(X\,|\,Y)=0$ a.s. and $\mathbb{E}(Y\,|\,X)=0$ a.s.”iff “X,Y are uncorrelated.”

Solution.

False（“iff”有一个方向错）。$\Rightarrow$ 成立：若 $\mathbb{E}(Y\,|\,X)=0$，由 LIE $\mathbb{E}(XY)=\mathbb{E}(X\mathbb{E}(Y\,|\,X))=0=\mathbb{E}(X)\mathbb{E}(Y)$，不相关。$\Leftarrow$ 不成立：不相关推不出条件均值为零。反例：$X\sim N(0,1)$，$Y=X^2-1$（均值零）。$\mathrm{Cov}(X,Y)=\mathbb{E}(X^3)=0$（不相关），但 $\mathbb{E}(Y\,|\,X)=X^2-1\neq 0$。故双向等价不成立，命题为假。要点：这正是 Proposition 3.14 中“CMI $\Rightarrow$ 不相关但反之不然”的判断题外衣。

3.7.2 Worked example：2023 Pinkse Q2（$y=x^2$ 的四问）

原题 (Comprehensive Exam 2023, Pinkse, Q2)

Let $y=x^2$, where $x\sim N(\mu,\sigma^2)$ for $\mu\in\mathbb{R}$ and $\sigma>0$. For what values of (μ,σ^2) are: (a) x,y uncorrelated; (b) x conditionally mean independent of y; (c) y conditionally mean independent of x; (d) x,y independent?

Solution.

(a) 不相关：$\mathrm{Cov}(x,x^2)=\mathbb{E}(x^3)-\mathbb{E}(x)\mathbb{E}(x^2)$。对 $N(\mu,\sigma^2)$，$\mathbb{E}(x^2)=\mu^2+\sigma^2$、$\mathbb{E}(x^3)=\mu^3+3\mu\sigma^2$，故

$$\mathrm{Cov}(x,x^2)=(\mu^3+3\mu\sigma^2)-\mu(\mu^2+\sigma^2)=2\mu\sigma^2.$$

因 $\sigma>0$，不相关 $\iff$ $\mu=0$。

(b) x 关于 y CMI（即 $\mathbb{E}(x\,|\,y)$ 不依赖 y）：$y=x^2$ 给定即知 $x\in\{+\sqrt{y},-\sqrt{y}\}$。当 $\mu=0$ 时 x 分布对称，给定 $x^2=y$ 时 $x=\pm\sqrt{y}$ 等概率，$\mathbb{E}(x\,|\,y)=0$（常数，CMI）。当 $\mu\neq 0$ 时正负不等概率，$\mathbb{E}(x\,|\,y)$ 随 y 变，非 CMI。故 (b) $\iff$ $\mu=0$。

(c) y 关于 x CMI（即 $\mathbb{E}(y\,|\,x)$ 不依赖 x）：$\mathbb{E}(y\,|\,x)=\mathbb{E}(x^2\,|\,x)=x^2$，对任意 $\sigma>0$ 都是 x 的非平凡函数。故 y 永不关于 x 条件均值独立——对一切 (μ,σ^2) 均不成立。

(d) 独立：$y=x^2$ 是 x 的（确定性）函数，故 x,y 永不独立（$\sigma>0$ 下 y 不退

化为常数）。

Remark (考点提炼).

这题把 Proposition 3.14 的“不对称性”逼到极致：(b) 与 (c) 问的是同一对 x, y 的两个方向，答案截然不同（一个 $\mu = 0$ 时成立、一个永不成立）——CMI 不对称的最佳注脚。(a) 不相关只需 $\mu = 0$（比 (b) 弱不了，因为这里恰好同条件），但 (d) 独立永不成立——“函数依赖”是最强的依赖。

3.8 答题策略与速查

本章题型识别 + 一招制胜

1. **给联合密度求 $\mathbb{E}(Y \mid X = x)$**：先求边际 f_X（注意支撑给出的积分限），再相除得 $f_{Y \mid X}$，再积分。看到“对任意 h 均方最小”立刻写 $g = \mathbb{E}(Y \mid X)$。
2. **给联合 MGF**：比对 $\exp(t'\mu + \frac{1}{2}t'\Sigma t)$ 认出（独立标准）正态。$\exp\{(t^2 + s^2)/2\} \Rightarrow X, Y \overset{iid}{\sim} N(0,1)$。
3. **求 $\mathbb{E}\big(aX + bY \mid$ 线性事件 $> 0\big)$**：令 $A = aX + bY$、$B =$ 事件里的线性组合；把 A 投影成 $\frac{\mathrm{Cov}(A,B)}{\mathrm{Var}(B)}B +$ 独立残差；用 tower 得 $\mathbb{E}(A \mid B > 0) = \frac{\mathrm{Cov}(A,B)}{\mathrm{Var}(B)}\mathbb{E}(B \mid B > 0)$；最后套 Mills。
4. **截断均值**：$\mathbb{E}(Z \mid Z > c) = \dfrac{\phi(c)}{1 - \Phi(c)}$，$\mathbb{E}(Z \mid Z < c) = -\dfrac{\phi(c)}{\Phi(c)}$；尺度 σ 时把 c 换 c/σ 再乘 σ。半正态 $\mathbb{E}(|Z|) = \mathbb{E}(Z \mid Z > 0) = \sqrt{2/\pi}$。
5. **分块正态条件分布**：默写 $\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$、$\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$；要推导就走“减投影 $R = X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2$、证 $R \perp X_2$”四步（模板见 §3.5.1）。
6. **独立/不相关/CMI 反例**：异方差 $y = xe$（CMI 非独立）、平方 $y = x^2 - \mathbb{E}\big(x^2\big)$（不相关非 CMI）；判断题先验“箭头方向”再找反例。

3.9 自测题 Self-Test

Example (Self-Test 1：分块正态条件期望).

设 $(X, Y)' \sim N((\frac{1}{2}), (\begin{smallmatrix} 4 & 2 \\ 2 & 3 \end{smallmatrix}))$。求 $\mathbb{E}(X \mid Y = y)$ 与 $\mathrm{Var}(X \mid Y)$。

Solution.

用一维分块公式：$\mathbb{E}(X \mid Y = y) = \mu_X + \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)}(y - \mu_Y) = 1 + \frac{2}{3}(y - 2)$；$\mathrm{Var}(X \mid Y) = \mathrm{Var}(X) - \frac{\mathrm{Cov}(X,Y)^2}{\mathrm{Var}(Y)} = 4 - \frac{4}{3} = \frac{8}{3}$。条件方差是常数（不依赖 y），且 $< \mathrm{Var}(X) = 4$。

Example (Self-Test 2：全方差分解).

设 $N \sim \text{Poisson}(\lambda)$，给定 $N = n$ 时 $Y \mid N = n \sim \text{Binomial}(n, p)$。求 $\text{Var}(Y)$。

Solution.

用 EVVE（Theorem 3.6）。$\mathbb{E}(Y \mid N) = Np$，$\text{Var}(Y \mid N) = Np(1-p)$。故 $\mathbb{E}(\text{Var}(Y \mid N)) = \lambda p(1-p)$，$\text{Var}(\mathbb{E}(Y \mid N)) = p^2 \text{Var}(N) = p^2\lambda$。相加 $\text{Var}(Y) = \lambda p(1-p) + \lambda p^2 = \lambda p$。（其实 $Y \sim \text{Poisson}(\lambda p)$，方差正是 λp，自洽。）

Example (Self-Test 3：截断 + 投影（仿 2020 Q3）).

$X, Y \overset{iid}{\sim} N(0,1)$。求 $\mathbb{E}(X - Y \mid X + Y > 0)$。

Solution.

令 $A = X - Y$、$B = X + Y$。$\text{Var}(B) = 2$，$\text{Cov}(A, B) = \text{Cov}(X - Y, X + Y) = \text{Var}(X) - \text{Var}(Y) = 0$。故 $\mathbb{E}(A \mid B) = 0$，从而 $\mathbb{E}(A \mid B > 0) = 0$。要点：$A, B$ 不相关 + 联合正态 $\Rightarrow$ 独立，故条件不改变 A 的均值。不必算任何 Mills 比。

Example (Self-Test 4：CMI 反例的方向性).

判断真假：若 $\mathbb{E}(Y \mid X) = \mathbb{E}(Y)$（$Y$ 关于 X 条件均值独立），则必有 $\mathbb{E}(X \mid Y) = \mathbb{E}(X)$。

Solution.

假。CMI 不对称。反例（即 3.7.2 里 $y = x^2$ 那套）：$X \sim N(0,1)$，令 $Y = X^2$。**一个方向成立**：$\mathbb{E}(X \mid Y) = \mathbb{E}\big(X \mid X^2\big) = 0 = \mathbb{E}(X)$——给定 $X^2 = y$ 时 $X = \pm\sqrt{y}$ 等概率，对称相消，故 X 是关于 Y 的 CMI。**另一方向失败**：$\mathbb{E}(Y \mid X) = \mathbb{E}\big(X^2 \mid X\big) = X^2$ 是 X 的非平凡函数，故 Y 不是关于 X 的 CMI。一边成立、另一边不成立，正是 CMI 不对称的样子。这与 2023 Pinkse Q2 的 (b)(c) 同一教训。（注意：常被误用的 $Y = XU$ 这里不能当反例——由 $(x, u) \mapsto (-x, -u)$ 的对称性，X 给定 Y 仍对称居中，$\mathbb{E}(X \mid Y) = 0 = \mathbb{E}(X)$，两个方向都是 CMI。）

Example (Self-Test 5：inverse Mills 数值).

$Z \sim N(0,1)$。求 $\mathbb{E}(Z \mid Z > 1)$（保留 ϕ, Φ）与 $\mathbb{E}(Z \mid |Z| > 1)$。

Solution.

$\mathbb{E}(Z \mid Z > 1) = \dfrac{\phi(1)}{1 - \Phi(1)}$。对第二个，$\{|Z| > 1\}$ 对称，由对称性 $\mathbb{E}(Z \mid |Z| > 1) = 0$（正负两半相消）。陷阱：别误用单边 Mills 公式去算双边对称事件——对称事件的条件均值是 0。

Chapter 4　收敛与渐近理论（Modes of Convergence & Asymptotics）

导读 · Roadmap（先读这里，不含公式）

有限样本里我们几乎算不出任何有趣统计量的精确分布——$\bar{X}_n$ 的分布、样本方差的分布、非线性变换 $g(\bar{X}_n)$ 的分布，统统没有闭式。渐近理论的整个使命，就是当 $n \to \infty$ 时给这些量一个近似：用一个我们认得的极限（常数、或正态、或卡方）去逼近真分布，从而能算标准误、造置信区间、做检验。

这一章先把"逼近"这个词讲精确——四种收敛模式（依概率 $\xrightarrow{p}$、依分布 $\xrightarrow{d}$、几乎必然 $\xrightarrow{a.s.}$、L^r 均方），它们互相蕴含还是不蕴含（带反例）；再给随机阶 O_p, o_p 这套"数量级演算"，它是把一长串误差项一眼判成可忽略的速记法。然后是三大支柱：大数定律（WLLN，含可能无穷方差时的截断证明）、中心极限定理（CLT，单变量、多变量加 Cramér–Wold 装置）、以及把它们串起来用的 Slutsky 定理与连续映射定理（CMT）。最后是两件干活的利器：*delta 法*（含一阶导为零时退化出卡方的二阶 delta 法）和收敛率（只有 p 阶矩时 $\bar{X}_n$ 的率，Marcinkiewicz–Zygmund）。

资格考在这一章的命中率极高，且年年换皮：2025 考 $\sqrt{n}(\hat{\sigma}^2 - \sigma_0^2)$ 的极限（二阶矩上的 delta）与"仅 p 阶矩时样本均值的率"；2020/2021 一道 true/false 横扫所有收敛模式；2019 一道纯证 WLLN 的截断题、一道二阶 delta 出 χ^2_1；2019/2020/2021 反复考 $(\bar{X}, \hat{\sigma}^2)$ 的联合极限分布；2022 考样本均值之非线性函数的 delta 法 + $o_p(1)$ 线性化。本章把这些题逐一独立重做，每个公式都对过 Andrews 讲义。

4.1　四种收敛模式及其蕴含关系

[优先级 ●●● 高 High]　资格考足迹：2020, 2022, 2024, 2025

我们有一列随机变量（或随机向量）$\{Y_n\}$，想说它"趋于"某个极限 Y。但"趋于"有四种正式含义，强弱不同，考场上必须分清。下面 Y_n, Y 都定义在同一概率空间 $(\Omega, \mathcal{F}, \mathrm{P}(\cdot))$ 上。

Definition 4.1: 四种收敛模式 Four modes of convergence

(i) 依概率收敛 *(in probability)* $Y_n \xrightarrow{p} Y$: for every $\varepsilon > 0$, $\mathrm{P}(\|Y_n - Y\| > \varepsilon) \to 0$ as $n \to \infty$.

(ii) 依分布收敛 *(in distribution)* $Y_n \xrightarrow{d} Y$: $F_{Y_n}(y) \to F_Y(y)$ at every continuity point y of F_Y.

(iii) 几乎必然收敛 *(almost surely)* $Y_n \xrightarrow{a.s.} Y$: $\mathrm{P}(\{\omega : Y_n(\omega) \to Y(\omega)\}) = 1$.

(iv) L^r *(均方, $r = 2$)* 收敛 $Y_n \xrightarrow{L^r} Y$ $(r \geq 1)$: $\mathbb{E}(\|Y_n - Y\|^r) \to 0$, assuming $\mathbb{E}(\|Y_n\|^r) < \infty$ for all n.

记法直觉：依概率说的是“Y_n 偏离 Y 超过 ε 这件事件的概率消失”；几乎必然说的是“几乎每一条样本路径 ω 本身都收敛”（强得多，是逐点收敛 + 全概率）；L^r 说的是“偏差的 r 阶矩消失”，是平均意义上的收敛；依分布最弱，只要求分布函数逐点（连续点）收敛，Y_n 和 Y 甚至不必在同一空间、不必“靠近”任何具体路径。

Fact 4.2: 蕴含关系（强 $\Rightarrow$ 弱）

$$\left(Y_n \xrightarrow{a.s.} Y\right) \text{ 或 } \left(Y_n \xrightarrow{L^r} Y\right) \implies Y_n \xrightarrow{p} Y \implies Y_n \xrightarrow{d} Y.$$

所有反方向一般都不成立。两条例外：

(a) 极限是常数时，$\xrightarrow{d}$ 与 $\xrightarrow{p}$ 等价：$Y_n \xrightarrow{d} c \iff Y_n \xrightarrow{p} c$（Lemma 4.13 之 Lemma 2）。

(b) $\xrightarrow{a.s.}$ 与 L^r 之间无单向蕴含——各有反例（见下）。

Remark (一张“强弱图”帮你记).

把四种收敛按强弱排：$\xrightarrow{a.s.}$ 与 L^r 同处“最强”一档（但互不蕴含），往下到 $\xrightarrow{p}$，再往下到 $\xrightarrow{d}$。考 true/false（2020 Q6、2022 Q4、2024 Q6）时，先在脑中画这张图：题目问的箭头若是“强推弱”则真，若是“弱推强”则假，且十有八九要你给反例。

下面把每个“弱不推强”的反例钉死，这些正是资格考的标准弹药。

Example (反例库：哪些箭头不成立).

(1) $\xrightarrow{p}$ **不推** $\xrightarrow{a.s.}$**.** 经典“打字机”序列：在 $([0,1], \lambda)$ 上让 Y_n 为越来越窄但在 $[0,1]$ 上来回扫动的区间的指示函数。则 $\mathrm{P}(Y_n \neq 0) \to 0$（故 $\xrightarrow{p} 0$），但每个 ω 都被无穷多个区间命中，$Y_n(\omega)$ 不收敛，故非 $\xrightarrow{a.s.}$。

(2) $\xrightarrow{d}$ **不推** $\xrightarrow{p}$**.** 取 $Y \sim N(0,1)$，令 $Y_n := -Y$ 对所有 n。则 $Y_n \xrightarrow{d} Y$（同分布），但 $|Y_n - Y| = 2|Y|$ 不趋于 0，故非 $\xrightarrow{p}$。

(3) $\xrightarrow{p}$ **不推** L^1**（即不推期望收敛）.** $U \sim \mathrm{Unif}(0,1)$, $Y_n := n\,\mathbb{1}(U \le 1/n)$。则 $Y_n \xrightarrow{p} 0$（$\mathrm{P}(Y_n \neq 0) = 1/n \to 0$），甚至 $Y_n \xrightarrow{a.s.} 0$，但 $\mathbb{E}(Y_n) = n \cdot \frac{1}{n} = 1 \not\to 0$。这就是 2020 Q6(a)/2022 Q4 的核心反例（缺一致可积性，见 4.3）；2025 Q3 则是它的反陷阱变体——约束 $f_n \le 1$ 恰好封杀这个尖峰反例、命题反而为真，见 4.3。

(4) L^r **不推** $\xrightarrow{a.s.}$**.** 上面 (1) 的打字机序列（取高度为 1）满足 $\mathbb{E}\left(Y_n^2\right) = \mathrm{P}(Y_n = 1) \to 0$，故 $L^2 \to 0$，但非 $\xrightarrow{a.s.}$。

Theorem 4.3: Markov / Chebyshev 不等式（造反例与证 WLLN 的引擎）

[REPRODUCE —memorize this proof]

For a random variable X and any $\varepsilon > 0$, $r > 0$,

$$\mathrm{P}(|X| \ge \varepsilon) \;\le\; \frac{\mathbb{E}(|X|^r)}{\varepsilon^r} \qquad \text{(Markov)}.$$

With $r = 2$ applied to $X - \mathbb{E}(X)$: $\mathrm{P}(|X - \mathbb{E}(X)| \ge \varepsilon) \le \mathrm{Var}(X)/\varepsilon^2$ (Chebyshev).

Proof for Theorem

[REPRODUCE —memorize this proof]

Fix $\varepsilon > 0$. Since $|X|^r \ge \varepsilon^r \mathbb{1}(|X| \ge \varepsilon)$ pointwise (the indicator is 0 or 1, and where it is 1 we have $|X|^r \ge \varepsilon^r$), take expectations: $\mathbb{E}(|X|^r) \ge \varepsilon^r\,\mathbb{E}(\mathbb{1}(|X| \ge \varepsilon)) = \varepsilon^r\,\mathrm{P}(|X| \ge \varepsilon)$. Divide by ε^r. ■

Markov 是整章的发动机：它把“矩有界”翻译成“尾概率小”，于是 (a) L^r 收敛 $\Rightarrow$ 依概率收敛（取 $X = Y_n - Y$）；(b) 方差 $\to 0 \Rightarrow$ 依概率收敛（Chebyshev，正是下面 WLLN 的证法）；(c) 矩有界 $\Rightarrow O_p$（见 4.2）。一条不等式，三处通吃。

4.2 随机阶 O_p 与 o_p 的演算

[优先级 ●●● 高 High] 资格考足迹：2020, 2021, 2022

证渐近结果时，误差项常常是一长串相加相乘的随机量。我们需要一套“数量级速记”——像微积分里的 O, o，但带上“依概率”——好让“这一项可忽略”一眼可判。

Definition 4.4: 随机阶 O_p, o_p (stochastic order)

Let $\{X_n\}$ be random and $\{a_n\}$ positive constants.

- $X_n = O_p(a_n)$ (*bounded in probability*, 随机有界): for every $\eta > 0$ there exist $M < \infty$ and N such that $\mathrm{P}(|X_n|/a_n > M) < \eta$ for all $n \ge N$. Equivalently $\{X_n/a_n\}$ is *tight*.
- $X_n = o_p(a_n)$: $X_n/a_n \xrightarrow{p} 0$, i.e. $\mathrm{P}(|X_n|/a_n > \varepsilon) \to 0$ for every $\varepsilon > 0$.

特例：$X_n = O_p(1)$ 即“随机有界”；$X_n = o_p(1)$ 即 $X_n \xrightarrow{p} 0$。

Fact 4.5: 随机阶演算规则（背下来当反射）

设 $a_n, b_n > 0$。则

(i) $o_p(a_n) = O_p(a_n)$（o_p 更强）；常数 $X_n = c$ 是 $O_p(1)$。

(ii) 乘法：$O_p(a_n)\,O_p(b_n) = O_p(a_n b_n)$；$O_p(a_n)\,o_p(b_n) = o_p(a_n b_n)$；$o_p(a_n)\,o_p(b_n) = o_p(a_n b_n)$.

(iii) 加法：$O_p(a_n) + O_p(b_n) = O_p(\max\{a_n, b_n\})$；$o_p(a_n) + o_p(b_n) = o_p(\max\{a_n, b_n\})$.

(iv) 由 $\xrightarrow{d}$ 得 O_p：若 $X_n \xrightarrow{d} X$（X 是某随机变量），则 $X_n = O_p(1)$.

(v) 由矩界得 O_p（Markov）：若 $\mathbb{E}(|X_n|^r) = O(a_n^r)$ 对某 $r > 0$，则 $X_n = O_p(a_n)$.

Proof for Theorem

[STRUCTURE —know the steps, fudge details]

只证最常用的两条。**(iv)** $\xrightarrow{d} \Rightarrow O_p(1)$：给 $\eta > 0$，取 $\pm M$ 为 F_X 的连续点且 $\mathrm{P}(|X| > M) < \eta/2$（$F_X$ 是分布函数，这样的 M 存在）。由 $X_n \xrightarrow{d} X$，$\mathrm{P}(|X_n| > M) \to \mathrm{P}(|X| > M) < \eta/2$，故对大 n 有 $\mathrm{P}(|X_n| > M) < \eta$。即 $\{X_n\}$ tight。**(v) 矩界** $\Rightarrow O_p$：由 Markov（Thm 4.3），$\mathrm{P}(|X_n|/a_n > M) = \mathrm{P}(|X_n| > M a_n) \le \mathbb{E}(|X_n|^r)/(M^r a_n^r) = O(1)/M^r$，对大 M 一致地小。■

规则 (iv) 是 Slutsky/delta 法证明里那句“$\sqrt{n}(\widehat{\theta} - \theta_0) = O_p(1)$ 故被 $o_p(1)$ 一吸就没了”的根据：凡是依分布收敛到某分布的东西，就是 $O_p(1)$，再乘以一个 $o_p(1)$ 即 $o_p(1)$，可丢。规则 (v) 则是 2020 Q6(b)/2021 Q6(a) 的全部内容——见下。

Example (2021 Q6(a) / 2020 Q6(b)：由三阶矩界得 $O_p(n^{1/3})$).

原题

Prove or disprove (all r.v.s on a common probability space): If $\mathbb{E}\left(|X_n|^3\right) = O(n)$, then $X_n = O_p(n^{1/3})$.

Solution.

结论：真。用 Markov（$r = 3$），对任意 $M > 0$：

$$\mathrm{P}\left(|X_n| > M\, n^{1/3}\right) \le \frac{\mathbb{E}\left(|X_n|^3\right)}{M^3\, n} = \frac{O(n)}{M^3\, n} = \frac{O(1)}{M^3}.$$

给 $\eta > 0$，由 $\mathbb{E}\left(|X_n|^3\right)/n \le C$（某常数 C，因为 $= O(n)$），取 $M = (C/\eta)^{1/3}$ 即

得 $\mathrm{P}\big(|X_n| > Mn^{1/3}\big) \le \eta$ 对所有 n。故 $X_n/n^{1/3}$ tight，即 $X_n = O_p(n^{1/3})$。■

这正是 Fact 4.5(v) 取 $a_n = n^{1/3}, r = 3$ 的特例：$\mathbb{E}(|X_n|^r) = O(n) = O((n^{1/3})^3) = O(a_n^r)$。

4.3 一致可积性：为什么依概率收敛不推期望收敛

[优先级 ●●● 高 High] 资格考足迹：2020, 2022, 2025

Fact 4.2 说 $\xrightarrow{a.s.}/L^r/\xrightarrow{p}$ 都不自动给出 $\mathbb{E}(Y_n) \to \mathbb{E}(Y)$。资格考极爱在这里设陷阱（2020 Q6a、2021、2022 Q4、2024 Q6a、2025 Q3）。补上这道缺口的标准概念，正是**一致可积性**（UI）：它是「$\xrightarrow{p}$ 何时才升级为期望收敛」的关键充分条件。**但别把它用过头**——下面 2025 Q3 就是一道反陷阱：题目那条密度约束不是让你去证 UI，而是让前提 $x_n \xrightarrow{p} 0$ 本身无法成立。

Definition 4.6: 一致可积 Uniform integrability (UI)

A sequence $\{Y_n\}$ is *uniformly integrable* if

$$\lim_{K\to\infty} \sup_n \mathbb{E}(|Y_n|\,\mathbb{1}(|Y_n| > K)) = 0.$$

Theorem 4.7: $\xrightarrow{p}$ + UI $\Rightarrow L^1$（期望收敛）

[STRUCTURE —know the steps, fudge details]

If $Y_n \xrightarrow{p} Y$ and $\{Y_n\}$ is uniformly integrable, then $Y_n \xrightarrow{L^1} Y$, and in particular $\mathbb{E}(Y_n) \to \mathbb{E}(Y)$. A sufficient condition for UI is a *dominating* integrable bound $|Y_n| \le Z$ with $\mathbb{E}(Z) < \infty$ (then DCT applies), or a *uniform moment bound* $\sup_n \mathbb{E}\left(|Y_n|^{1+\delta}\right) < \infty$ for some $\delta > 0$.

口诀：依概率收敛只管住了"概率质量的位置"，管不住"跑到无穷远的那一小撮质量带走多少期望"。UI 就是禁止"质量逃逸到无穷远"。控制收敛定理 DCT 是 UI 的一个充分条件（有公共可积控制函数）；"一致 $1+\delta$ 阶矩有界"是另一个更易查的充分条件。

Example (2025 Q3（反陷阱）：密度 ≤ 1 让前提 $x_n \xrightarrow{p} 0$ 无法成立).

原题 (2025 Q3, Pinkse)

Let x_n be continuously distributed random variables with bounded support and density $f_n \le 1$. Suppose $x_n \xrightarrow{p} 0$. Prove or disprove: "$\mathbb{E}(x_n)$ necessarily converges to zero."

Solution.

答案：命题「空虚为真」——因为在 $f_n \le 1$ 之下，前提 $x_n \xrightarrow{p} 0$ 根本不可能发生。这题表面问「$\xrightarrow{p}$ 推不推期望收敛」，实为一道反陷阱：真正的关卡在前提本身。

先破陷阱。顺着上一节的直觉，会想搬出「又高又窄的尖峰」$n\,\mathbb{1}(U \le 1/n)$（见下条 2020 Q6a）答「可证伪」。但那个反例把质量 $1-\frac{1}{n}$ 堆在单点 0、密度趋于无穷，违反 $f_n \le 1$——在本题非法，用不了。

密度封顶直接杀死前提（关键一步）。密度 ≤ 1 让任何长度 2ε 的区间至多承载概率 2ε：对每个 n、每个 $\varepsilon > 0$，

$$\mathrm{P}(|x_n| \le \varepsilon) = \int_{-\varepsilon}^{\varepsilon} f_n(t)\,dt \;\le\; 2\varepsilon \qquad \Longrightarrow \qquad \mathrm{P}(|x_n| > \varepsilon) \;\ge\; 1 - 2\varepsilon.$$

取 $\varepsilon = \frac{1}{4}$：$\mathrm{P}\big(|x_n| > \frac{1}{4}\big) \ge \frac{1}{2}$ 对所有 n 成立，与 n 无关、永不趋于 0。而 $x_n \xrightarrow{p} 0$ 按定义要求 $\mathrm{P}(|x_n| > \varepsilon) \to 0$ 对每个 ε 成立——正相矛盾。故满足 $f_n \le 1$ 的序列不可能有 $x_n \xrightarrow{p} 0$。

结论：前提自相矛盾，蕴含式空虚为真。既然没有任何序列能同时满足 $f_n \le 1$ 与 $x_n \xrightarrow{p} 0$，蕴含式「$x_n \xrightarrow{p} 0 \Rightarrow \mathbb{E}(x_n) \to 0$」的适用对象是空集，空前提的蕴含式恒真。■

Remark (本题真正考点：先查前提可不可满足，别硬套 UI).

这也是一堂「工具别用过头」的课。**(a) 一般原理**：$\xrightarrow{p}$ 不推期望收敛，桥梁是一致可积（UI / Vitali），无约束时「远端尖峰」正是反例（下条 2020 Q6a）。**(b) 本题的扣杀**：题给的 $f_n \le 1$ 不是让你去证 UI，而是把尖峰彻底封死——密度摊平后质量无法向 0 集中，前提 $\xrightarrow{p} 0$ 本身就不成立。逻辑务必闭合：若坚持判「假」，就必须拿出一个同时满足 $f_n \le 1$、$x_n \xrightarrow{p} 0$、$\mathbb{E}(x_n) \not\to 0$ 的反例；上面已证这种序列不存在，所以「判假却给不出合规反例」自相矛盾，正确判定是空虚为真。**反向印证**：把 $f_n \le 1$ 去掉、只留「连续」，尖峰立刻借尸还魂成连续版——以概率 $1-\frac{1}{n}$ 落在 $[0, \frac{1}{n}]$、以概率 $\frac{1}{n}$ 落在 $[n, n+\frac{1}{n}]$（近团密度 $\approx n$，已破 1），此时 $x_n \xrightarrow{p} 0$ 但 $\mathbb{E}(x_n) \approx (1-\frac{1}{n})\frac{1}{2n} + \frac{1}{n}\cdot n \to 1 \ne 0$，答案立刻翻成「可证伪」。可见 $f_n \le 1$ 的作用真实而决定性，只是强到让整题变平凡。

Example (2020 Q6(a) / 2022 Q4：$\xrightarrow{a.s.} 0$（或 $\xrightarrow{p} b$）不推期望收敛——干净反例).

原题 (2020 Q6a)

Prove or disprove: If $X_n \to 0$ almost surely, then $\mathbb{E}(X_n) \to 0$.

Solution.

可证伪。取 $U \sim \mathrm{Unif}(0,1)$，$X_n := n\,\mathbb{1}(U \le 1/n)$。对固定 ω（即固定 $U(\omega) > 0$），当 $n > 1/U$ 时 $X_n = 0$，故 $X_n \xrightarrow{a.s.} 0$（从而也 $\xrightarrow{p} 0$）。但

$$\mathbb{E}(X_n) = n \cdot \mathrm{P}(U \le 1/n) = n \cdot \tfrac{1}{n} = 1 \quad \forall n, \qquad \text{故} \mathbb{E}(X_n) \to 1 \ne 0.$$

缺的就是一致可积性：$\mathbb{E}(|X_n|\mathbb{1}(|X_n|>K))=1$ 对所有 $n>K$，sup 不趋于 0。同一反例答 2022 Q4（“$X_n \xrightarrow{p} b \Rightarrow \mathbb{E}(X_n)\to b$?”——假，取 $b=0$）。■

4.4 大数定律 WLLN 与 SLLN

[优先级 ●●● 高 High] 资格考足迹：2018, 2019, 2022

大数定律说样本均值依（概率 / 几乎必然）收敛到总体均值——这是“样本矩估计总体矩”一切一致性论证的源头。

Theorem 4.8: 弱大数定律 WLLN（L^2 版，有限方差）

[REPRODUCE —memorize this proof]

Let $\{X_i\}$ be iid with $\mathbb{E}(X_1)=\mu$ and $\mathrm{Var}(X_1)=\sigma^2<\infty$. Then $\bar{X}_n \xrightarrow{p} \mu$.

Proof for Theorem

[REPRODUCE —memorize this proof]

$\mathbb{E}(\bar{X}_n)=\mu$ and, by independence, $\mathrm{Var}(\bar{X}_n)=\sigma^2/n$. By Chebyshev (Thm 4.3, $r=2$), for any $\varepsilon>0$,

$$\mathrm{P}\big(|\bar{X}_n-\mu|>\varepsilon\big) \;\le\; \frac{\mathrm{Var}(\bar{X}_n)}{\varepsilon^2} \;=\; \frac{\sigma^2}{n\varepsilon^2} \;\to\; 0. \qquad \blacksquare$$

Khinchin WLLN（只需一阶矩）

其实只要 $\mathbb{E}(|X_1|)<\infty$（无须有限方差）即有 $\bar{X}_n \xrightarrow{p} \mu$。证明走截断（下题）或特征函数。资格考真正考的是自己用截断把它证出来。

Theorem 4.9: 强大数定律 SLLN（陈述）

[STRUCTURE —know the steps, fudge details]

Let $\{X_i\}$ be iid with $\mathbb{E}(|X_1|)<\infty$ and $\mathbb{E}(X_1)=\mu$. Then $\bar{X}_n \xrightarrow{a.s.} \mu$ (Kolmogorov). 几乎必然收敛蕴含依概率收敛，故 SLLN 比 WLLN 强。

Example (2019 Q4：用截断法证 WLLN（方差可能无穷，对称）).

原题 (2019 Comp Q4, Guggenberger)

Let $X_1,\dots,X_n$ be iid mean-zero random variables with a distribution symmetric about zero and variance *not necessarily finite*. Prove the law of large numbers by a truncation argument, using $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i\mathbb{1}(|X_i|\le c)+$

$n^{-1}\sum_{i=1}^{n} X_i \mathbb{1}(|X_i| > c)$ for any constant $c > 0$.

Solution.

[STRUCTURE —know the steps, fudge details]

目标：$\bar{X}_n \xrightarrow{p} 0$。假设 $\mathbb{E}(X_1) = 0$ 且 $\mathbb{E}(|X_1|) < \infty$（Khinchin 的标准前提；若连一阶矩都无穷则均值无定义）。对称性不保证可积（如对称 Cauchy 有 $\mathbb{E}(|X_1|) = \infty$），它在下面仅用来保证*截断后仍均值 0*。

第 1 步：固定 c 做截断分解。写 $X_i = Y_i + Z_i$，其中

$$Y_i := X_i \mathbb{1}(|X_i| \le c) \ \text{（截断部分）}, \qquad Z_i := X_i \mathbb{1}(|X_i| > c) \ \text{（尾部）}.$$

于是 $\bar{X}_n = \bar{Y}_n + \bar{Z}_n$。我们分别控制两块，再让 c 适当大。

第 2 步：截断部分用 Chebyshev。Y_i 有界（$|Y_i| \le c$），故方差有限：$\mathrm{Var}(Y_i) \le \mathbb{E}(Y_i^2) \le c^2$。由对称性，$X_1$ 关于 0 对称 $\Rightarrow$ $X_1 \mathbb{1}(|X_1| \le c)$ 也关于 0 对称 $\Rightarrow$ $\mathbb{E}(Y_i) = 0$。所以 $\bar{Y}_n$ 是均值 0、方差 $\le c^2/n$ 的和，Chebyshev 给

$$\mathrm{P}\big(|\bar{Y}_n| > \varepsilon/2\big) \ \le \ \frac{\mathrm{Var}(Y_1)}{n(\varepsilon/2)^2} \ \le \ \frac{4c^2}{n\varepsilon^2} \ \xrightarrow[n\to\infty]{} \ 0 \quad (\text{固定} c).$$

第 3 步：尾部用 Markov + 可积性。$\mathbb{E}(|Z_i|) = \mathbb{E}(|X_1|\,\mathbb{1}(|X_1| > c)) =: \tau(c)$。由 $\mathbb{E}(|X_1|) < \infty$ 与控制收敛，$\tau(c) \to 0$ as $c \to \infty$。由三角不等式与 Markov，

$$\mathrm{P}\big(|\bar{Z}_n| > \varepsilon/2\big) \ \le \ \frac{\mathbb{E}\big(|\bar{Z}_n|\big)}{\varepsilon/2} \ \le \ \frac{n^{-1}\sum_i \mathbb{E}(|Z_i|)}{\varepsilon/2} \ = \ \frac{2\tau(c)}{\varepsilon}.$$

第 4 步：先 n 后 c，夹出 0。由 $\bar{X}_n = \bar{Y}_n + \bar{Z}_n$ 与并事件，

$$\mathrm{P}\big(|\bar{X}_n| > \varepsilon\big) \ \le \ \mathrm{P}\big(|\bar{Y}_n| > \varepsilon/2\big) + \mathrm{P}\big(|\bar{Z}_n| > \varepsilon/2\big) \ \le \ \frac{4c^2}{n\varepsilon^2} + \frac{2\tau(c)}{\varepsilon}.$$

取 $\limsup_{n\to\infty}$（第一项消失）：$\limsup_n \mathrm{P}\big(|\bar{X}_n| > \varepsilon\big) \le 2\tau(c)/\varepsilon$。此式对*每个* c 成立，再令 $c \to \infty$（$\tau(c) \to 0$）得 $\limsup_n \mathrm{P}\big(|\bar{X}_n| > \varepsilon\big) = 0$。故 $\bar{X}_n \xrightarrow{p} 0$。■

Remark（截断证明的“三段套路”）.

这套“截断 + Chebyshev（中间块）+ Markov 尾部 + 先 n 后 c”是处理*可能无穷方差*的标准武器，务必背熟节奏：(1) 按 $|X_i| \lessgtr c$ 劈成中间块与尾部；(2) 中间块有界 $\Rightarrow$ 方差有限 $\Rightarrow$ Chebyshev，让 $n \to \infty$ 杀掉；(3) 尾部用 $\mathbb{E}(|X_1|) < \infty$ 让 $\tau(c) = \mathbb{E}(|X_1|\,\mathbb{1}(|X_1| > c)) \to 0$；(4) 关键是*先对 n 取 $\limsup$、再让 $c \to \infty$*——顺序反了就漏。对称性在此只用来保证截断后*仍均值 0*（否则要再减去 $\mathbb{E}(Y_i)$ 并验证它 $\to 0$）。

4.5 中心极限定理 CLT

[优先级 ●●● 高 High] 资格考足迹：2019, 2020, 2021, 2022, 2025

WLLN 只说 $\bar{X}_n \to \mu$，是个常数极限，没法量"离 μ 多远"。CLT 把标准化后的 $\bar{X}_n$ 的分布钉成正态，这才让我们能造置信区间、做检验。

Theorem 4.10: 中心极限定理（单变量，iid）

[STRUCTURE —know the steps, fudge details]

Let $\{X_i\}$ be iid with $\mathbb{E}(X_1) = \mu$ and $0 < \mathrm{Var}(X_1) = \sigma^2 < \infty$. Then

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2), \qquad \text{equivalently} \quad \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} Z \sim N(0,1).$$

Proof for Theorem

[STRUCTURE —know the steps, fudge details]

（一般有限方差情形 $0 < \sigma^2 < \infty$ 的证明用特征函数 + Lévy 连续性定理——特征函数总存在，无须任何额外矩条件，是标准结论、考场直接引用即可，此处不展开。下面给的 MGF 论证覆盖 $M_{X_1}(t) < \infty$ 于 0 邻域的特例，步骤与特征函数版完全平行，只是把 φ 换成 M，故作为可默写的代表证法列出。）

（MGF 法，假设 $M_{X_1}(t) < \infty$ 于 0 的邻域。）令 $Y_i = (X_i - \mu)/\sigma$，则 $\mathbb{E}(Y_i) = 0, \mathrm{Var}(Y_i) = 1$，且 $\sqrt{n}(\bar{X}_n - \mu)/\sigma = \sqrt{n}\,\bar{Y}_n = n^{-1/2}\sum Y_i$。由独立同分布，

$$M_{n^{-1/2}\sum Y_i}(t) = \big(M_{Y_1}(t/\sqrt{n})\big)^n.$$

Y_1 的 MGF 在 0 处二阶 Taylor 展开：$M_{Y_1}(s) = 1 + M'_{Y_1}(0)s + \frac{1}{2}M''_{Y_1}(\tilde{s})s^2 = 1 + \frac{1}{2}M''_{Y_1}(\tilde{s})s^2$（因 $M'(0) = \mathbb{E}(Y_1) = 0$）。代入 $s = t/\sqrt{n}$，M''_{Y_1} 在 0 处连续且 $\tilde{s}/\sqrt{n} \to 0$，故 $M''_{Y_1}(t/\sqrt{n}) \to M''_{Y_1}(0) = \mathbb{E}\big(Y_1^2\big) = 1$。于是

$$\big(M_{Y_1}(t/\sqrt{n})\big)^n = \left(1 + \frac{\frac{1}{2}M''_{Y_1}(\tilde{s})\,t^2}{n}\right)^n \longrightarrow e^{t^2/2} = M_Z(t),$$

用 $\lim_n (1 + a/n)^n = e^a$。MGF 逐点收敛到 $N(0,1)$ 的 MGF ⇒ 分布函数收敛（连续极限故处处），即 $\sqrt{n}\bar{Y}_n \xrightarrow{d} N(0,1)$。■

两点考场提醒：(1) CLT 必须先标准化——**乘以** $\sqrt{n}$ 是为了让方差不塌（$\mathrm{Var}\big(\sqrt{n}(\bar{X}_n - \mu)\big) = \sigma^2$ 不随 n 变），不乘则 $\bar{X}_n - \mu \xrightarrow{p} 0$ 退化、乘以 n 则爆掉（$\mathrm{Var}\big(n(\bar{X}_n - \mu)\big) = n\sigma^2 \to \infty$）。$\sqrt{n}$ 是唯一让方差稳定的尺度。(2) 极限方差就是单个 X_1 的方差 σ^2，不是 σ^2/n。

Theorem 4.11: 多变量 CLT

[STRUCTURE —know the steps, fudge details]

Let $\{X_i\}$ be iid random k-vectors with $\mathbb{E}(X_1) = \mu$ and $\mathrm{Var}(X_1) = \Sigma =$

$\mathbb{E}((X_1-\mu)(X_1-\mu)')$ (finite). Then

$$\sqrt{n}(\bar{X}_n-\mu) \xrightarrow{d} N(0,\Sigma).$$

若 Σ 非奇异，可写 $\Sigma^{-1/2}\sqrt{n}(\bar{X}_n-\mu)\xrightarrow{d}N(0,I_k)$。
（证明借助下面的 *Cramér–Wold* 装置——见 *Theorem 4.12* 之后的证明。）

Theorem 4.12: Cramér–Wold 装置（把多变量 CLT 归约到单变量）

[INTUITION ONLY —read once, do not memorize]

A sequence of random k-vectors $\{Y_n\}$ satisfies $Y_n \xrightarrow{d} Y$ *if and only if* $c'Y_n \xrightarrow{d} c'Y$ for every $c\in\mathbb{R}^k$.

（*Cramér–Wold* 装置本身的证明用特征函数 + *Lévy* 连续性定理，是标准结论、考场直接引用即可，此处不展开。下面用它来证多变量 ***CLT***。）

证明. [**Proof (多变量 CLT（Theorem 4.11）的证明——用 Cramér–Wold 归约到单变量).**] **[STRUCTURE —know the steps, fudge details]**

设 X_i iid，$\mathbb{E}(X_1)=\mu, \mathrm{Var}(X_1)=\Sigma$。固定 $c\in\mathbb{R}^k, c\neq 0$。则 $\{c'X_i\}$ 是 iid 标量，均值 $c'\mu$、方差 $c'\Sigma c$。由单变量 CLT，

$$\sqrt{n}\big(c'\bar{X}_n-c'\mu\big)=\sqrt{n}\,n^{-1}\sum_i(c'X_i-c'\mu) \xrightarrow{d} N(0,\,c'\Sigma c).$$

注意 $N(0,c'\Sigma c)$ 恰是 $c'Y$ 的分布，$Y\sim N(0,\Sigma)$（当 $c'\Sigma c=0$，如 Σ 奇异且 c 落在其零空间时，单变量 CLT 退化为 $c'\sqrt{n}(\bar{X}_n-\mu)\xrightarrow{p}0=c'Y$ 的退化正态 δ_0，仍成立）。再补上平凡的 $c=0$（两边恒为 0）。故对每个 $c\in\mathbb{R}^k$，$c'\big[\sqrt{n}(\bar{X}_n-\mu)\big]\xrightarrow{d}c'Y$。由 Cramér–Wold（$\Leftarrow$ 方向），$\sqrt{n}(\bar{X}_n-\mu)\xrightarrow{d}Y\sim N(0,\Sigma)$。 □

Cramér–Wold 的威力：它把“k 维收敛”降维成“沿每个方向 c 的一维收敛”。证多变量 CLT 时，沿任一方向投影就是 iid 标量和，单变量 CLT 直接适用；再由 Cramér–Wold 收回多维。这是把多变量结果免费从单变量搬过来的标准桥。

4.6 Slutsky 定理与连续映射定理

[优先级 ●●● 高 High] 资格考足迹：2019, 2020, 2022

CLT 给的是 $\sqrt{n}(\bar{X}_n-\mu)/\sigma\xrightarrow{d}N(0,1)$，但实际中 σ 未知、要用 $\hat{\sigma}$ 替；又常要对收敛的东西再作连续变换。Slutsky 与 CMT 就是合法做这些替换/变换的执照。

Lemma 4.13: Slutsky 引理（$\xrightarrow{p}$ 内部 + $\xrightarrow{d}$ 组合）

[STRUCTURE —know the steps, fudge details]

Slutsky (in prob, CMT version): If $Y_n \xrightarrow{p} c$ (c a constant vector) and $h(\cdot)$ is continuous at c, then $h(Y_n) \xrightarrow{p} h(c)$.

Slutsky (mixing $\xrightarrow{d}$ with $\xrightarrow{p}$, Lemma 1): If $Y_n \xrightarrow{d} Y$ and $Z_n \xrightarrow{p} r$ (r constant), then

$$\text{(a) } Z_n + Y_n \xrightarrow{d} r + Y, \quad \text{(b) } Z_n Y_n \xrightarrow{d} rY, \quad \text{(c) } Y_n / Z_n \xrightarrow{d} Y/r \ (r \neq 0).$$

Lemma 2: $Y_n \xrightarrow{d} c \iff Y_n \xrightarrow{p} c$ for a constant c.

（本引理中只有第一条“*in-prob CMT* 版”下面给出可复现证明（***[REPRODUCE]***，见下）——这是考场务必默写的一条。混合部分 *(a)(b)(c)* 由 *CMT* +「向常数收敛自动升级为联合收敛」（*Lemma 3*，见 *4.6* 末 `rmkb`）得出；*Lemma 2* 的 $\xrightarrow{d} c \Rightarrow \xrightarrow{p} c$ 是标准结论，均直接引用。）

Theorem 4.14: 连续映射定理 CMT

[INTUITION ONLY —read once, do not memorize]

If $Y_n \xrightarrow{d} Y$ and $h : \mathbb{R}^m \to \mathbb{R}^p$ is continuous on a set B with $\mathrm{P}(Y \in B) = 1$ (i.e. continuous a.s.-P_Y), then $h(Y_n) \xrightarrow{d} h(Y)$.

（一般 $\xrightarrow{d}$ 版的证明用 *portmanteau* 定理 / *Skorokhod* 表示，是标准结论、考场直接引用即可，此处不展开。下面证最常用的 $\xrightarrow{p}$ 常数特例。）

证明. [**Proof (连续映射的 $\xrightarrow{p}$ 常数特例（$Y_n \xrightarrow{p} c$、h 连续于 $c \Rightarrow h(Y_n) \xrightarrow{p} h(c)$；即 Lemma 4.13 的 Slutsky-CMT 版）.**] **[REPRODUCE —memorize this proof]**

由连续性，$\forall \varepsilon > 0\, \exists \delta > 0$ 使 $\|y - c\| \le \delta \Rightarrow \|h(y) - h(c)\| \le \varepsilon$。逆否得事件包含 $\{\|h(Y_n) - h(c)\| > \varepsilon\} \subseteq \{\|Y_n - c\| > \delta\}$。故

$$\mathrm{P}(\|h(Y_n) - h(c)\| > \varepsilon) \le \mathrm{P}(\|Y_n - c\| > \delta) \to 0.$$

□

Remark (CMT 与 Slutsky 的统一视角).

CMT 是更一般的：若 $(Y_n, Z_n) \xrightarrow{d} (Y, r)$（联合），对连续 h 有 $h(Y_n, Z_n) \xrightarrow{d} h(Y, r)$。而当其中一支收敛到常数 r 时，“边际收敛”自动升级为“联合收敛”（Lemma 3：$Y_n \xrightarrow{d} Y, Z_n \xrightarrow{p} r \Rightarrow (Y_n, Z_n)' \xrightarrow{d} (Y, r)'$），于是取 $h(y, z) = y + z$ 或 yz 就得到 Slutsky 的 (a)(b)。**关键警告**：若 $Z_n \xrightarrow{d} Z$ 收敛到随机的 Z（非常数），则一般 $Y_n + Z_n \not\xrightarrow{d} Y + Z$——因为边际收敛说不清联合分布（反例：$Y_n \xrightarrow{d} N(0,1)$，

$Z_n = -Y_n \xrightarrow{d} N(0,1)$，但 $Y_n + Z_n \equiv 0 \not\xrightarrow{d} N(0,2)$）。

Example (CMT 应用：t 统计量与 Wald 统计量).

(a) t-比化为正态. 由 CLT $\sqrt{n}(\bar{X}_n - \mu)/\sigma \xrightarrow{d} N(0,1)$，由 WLLN+CMT（$h(x) = \sqrt{x}$）$S_{Xn} \xrightarrow{p} \sigma$，故 $\sigma/S_{Xn} \xrightarrow{p} 1$。Slutsky (b)：

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_{Xn}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \cdot \frac{\sigma}{S_{Xn}} \xrightarrow{d} N(0,1) \cdot 1 = N(0,1).$$

(b) Wald $\to \chi^2$. 多变量 CLT 给 $\Sigma^{-1/2}\sqrt{n}(\bar{X}_n - \mu_0) \xrightarrow{d} Z \sim N(0, I_k)$。取连续 $h(y) = y'y$，CMT：

$$W_n = n(\bar{X}_n - \mu_0)'\Sigma^{-1}(\bar{X}_n - \mu_0) = h\big(\Sigma^{-1/2}\sqrt{n}(\bar{X}_n - \mu_0)\big) \xrightarrow{d} Z'Z \sim \chi^2_k.$$

若 Σ 未知，用 $\hat{\Sigma} \xrightarrow{p} \Sigma$ 替换：因 $h(y, \Sigma) = y'\Sigma^{-1}y$ 在 Σ 正定处连续，CMT 给同样的 χ^2_k 极限——估计 Σ 不改变渐近分布。

4.7 Delta 法（一阶与二阶）

[优先级 ●●● 高 High] 资格考足迹：2019, 2020, 2021, 2022, 2025

我们常关心的不是 $\widehat{\theta}_n$ 本身，而是它的非线性变换 $g(\widehat{\theta}_n)$（比如方差是二阶矩的函数、弹性是比值）。Delta 法把 $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ 的渐近正态传播到 $\sqrt{n}(g(\widehat{\theta}_n) - g(\theta_0))$。

Theorem 4.15: Delta 法（一维）

[REPRODUCE —memorize this proof]

Suppose $\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} Z^* \sim N(0, \sigma^2)$ and g is continuously differentiable at θ_0 with $g'(\theta_0) \neq 0$. Then

$$\sqrt{n}\big(g(\widehat{\theta}_n) - g(\theta_0)\big) \xrightarrow{d} g'(\theta_0)\, Z^* \sim N\big(0, \{g'(\theta_0)\}^2\sigma^2\big).$$

Proof for Theorem

[REPRODUCE —memorize this proof]

均值定理：$g(\widehat{\theta}_n) = g(\theta_0) + g'(\theta_n^*)(\widehat{\theta}_n - \theta_0)$，$\theta_n^*$ 在 $\widehat{\theta}_n, \theta_0$ 之间。乘 $\sqrt{n}$：

$$\sqrt{n}\big(g(\widehat{\theta}_n) - g(\theta_0)\big) = g'(\theta_n^*) \cdot \sqrt{n}(\widehat{\theta}_n - \theta_0).$$

由 $\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} Z^*$ 知 $\widehat{\theta}_n - \theta_0 \xrightarrow{p} 0$（$\widehat{\theta}_n - \theta_0 = \sqrt{n}(\widehat{\theta}_n - \theta_0)/\sqrt{n} = O_p(1) \cdot o(1) = o_p(1)$），且 $|\theta_n^* - \theta_0| \le \left|\widehat{\theta}_n - \theta_0\right| \xrightarrow{p} 0$，故 $\theta_n^* \xrightarrow{p} \theta_0$。$g'$ 在 θ_0 连续 $\Rightarrow g'(\theta_n^*) \xrightarrow{p} g'(\theta_0)$（Slutsky in-prob）。再由 Slutsky (b)：$g'(\theta_n^*) \cdot \sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} g'(\theta_0)Z^* \sim N(0, \{g'(\theta_0)\}^2\sigma^2)$。■

Theorem 4.16: 多变量 delta 法

[STRUCTURE —know the steps, fudge details]

Suppose $\sqrt{n}(\widehat{\theta}_n-\theta_0)\xrightarrow{d}Z\sim N(0,\Sigma)$ $(\theta_0\in\mathbb{R}^m)$ and $g:\mathbb{R}^m\to\mathbb{R}^\ell$ is differentiable at θ_0 with Jacobian $G(\theta_0)=\partial g(\theta_0)/\partial\theta'$ $(\ell\times m)$. Then

$$\sqrt{n}\big(g(\widehat{\theta}_n)-g(\theta_0)\big)\ \xrightarrow{d}\ G(\theta_0)Z\ \sim\ N\big(0,\,G(\theta_0)\,\Sigma\,G(\theta_0)'\big).$$

Theorem 4.17: 二阶 delta 法（$g'(\theta_0)=0$）

[STRUCTURE —know the steps, fudge details]

If $\sqrt{n}(\widehat{\theta}_n-\theta_0)\xrightarrow{d}Z^*\sim N(0,\sigma^2)$, $g\in C^2$ near θ_0, $g'(\theta_0)=0$ but $g''(\theta_0)\neq 0$, then the $\sqrt{n}$-rate degenerates and the *correct* rate is n:

$$n\big(g(\widehat{\theta}_n)-g(\theta_0)\big)\ \xrightarrow{d}\ \tfrac{1}{2}\,g''(\theta_0)\,(Z^*)^2\ =\ \tfrac{1}{2}\,g''(\theta_0)\,\sigma^2\,\chi^2_1,$$

其中 $(Z^*)^2=\sigma^2\cdot(Z^*/\sigma)^2$，$(Z^*/\sigma)^2\sim\chi^2_1$。

Proof for Theorem

[STRUCTURE —know the steps, fudge details]

二阶 Taylor（$g'(\theta_0)=0$）：$g(\widehat{\theta}_n)=g(\theta_0)+\frac{1}{2}g''(\theta_n^*)(\widehat{\theta}_n-\theta_0)^2$，$\theta_n^*$ 在中间。乘 n：

$$n\big(g(\widehat{\theta}_n)-g(\theta_0)\big)=\tfrac{1}{2}g''(\theta_n^*)\big[\sqrt{n}(\widehat{\theta}_n-\theta_0)\big]^2.$$

$\theta_n^*\xrightarrow{p}\theta_0$ 故 $g''(\theta_n^*)\xrightarrow{p}g''(\theta_0)$；又 $\sqrt{n}(\widehat{\theta}_n-\theta_0)\xrightarrow{d}Z^*$，由 CMT（$h(x)=x^2$）$[\sqrt{n}(\widehat{\theta}_n-\theta_0)]^2\xrightarrow{d}(Z^*)^2$。Slutsky (b)：

$$n\big(g(\widehat{\theta}_n)-g(\theta_0)\big)\ \xrightarrow{d}\ \tfrac{1}{2}g''(\theta_0)(Z^*)^2=\tfrac{1}{2}g''(\theta_0)\sigma^2\chi^2_1.\qquad\blacksquare$$

为什么 $g'(\theta_0)=0$ 要换挡到 n？一阶 delta 的极限是 $g'(\theta_0)Z^*$；当 $g'(\theta_0)=0$ 这是退化的 0，说明 $\sqrt{n}$ 这个放大倍数不够、信息全藏在二阶项里。把放大倍数从 $\sqrt{n}$ 升到 n，二次项 $[\sqrt{n}(\widehat{\theta}_n-\theta_0)]^2$ 恰好 $O_p(1)$、收敛到 $(Z^*)^2$，于是极限是卡方（正态的平方）而非正态。一句话：一阶导没了，分布从正态升一阶变卡方，率从 $\sqrt{n}$ 快一倍变 n。

Example (2019 Q6：二阶 delta 法，导出 $\frac{1}{2}g''\sigma^2\chi^2_1$).

原题 (2019 Comp Q6, Guggenberger)

$\sqrt{n}(Y_n-\theta)\xrightarrow{d}Z\sim N(0,\sigma^2)$, $g\in C^2(U)$ near θ, scalar-valued, with $g'(\theta)=0$. Find the limiting distribution of $g(Y_n)$ (with correct normalization and

centering). State any extra assumption.

Solution.

额外假设：$g''(\theta)\neq 0$（否则二阶项也退化，需更高阶）。直接套 Thm 4.17（$\widehat{\theta}_n = Y_n, \theta_0=\theta$）：

$$\boxed{n\big(g(Y_n)-g(\theta)\big)\ \xrightarrow{d}\ \tfrac{1}{2}g''(\theta)\,\sigma^2\,\chi^2_1\,.}$$

正确的归一化与中心化是：中心化在 $g(\theta)$、放大倍数为 n（不是 $\sqrt{n}$）。等价写法 $n(g(Y_n)-g(\theta))\xrightarrow{d}\frac{1}{2}g''(\theta)Z^2$，$Z\sim N(0,\sigma^2)$。■

4.8 收敛率：只有 p 阶矩时样本均值的速度

[优先级 ●●○ 中 Med] 资格考足迹：2025

CLT 假设了二阶矩有限，给出 $\bar{X}_n-\mu=O_p(n^{-1/2})$。若只有 p 阶矩（$1<p<2$），方差可能无穷，CLT 不适用——这时 $\bar{X}_n$ 收敛得更慢。率由 Marcinkiewicz–Zygmund 给出。

Theorem 4.18: Marcinkiewicz–Zygmund 型收敛率

[STRUCTURE —know the steps, fudge details]

Let $\{X_i\}$ be iid with $\mathbb{E}(X_1)=\mu$ and $0<\mathbb{E}(|X_1-\mu|^p)<\infty$ for some $p\in(1,\infty)$. Then

$$\bar{X}_n-\mu\ =\ O_p\big(n^{-(1-1/p)}\big)\quad\text{for } 1<p\le 2,\qquad \bar{X}_n-\mu=O_p\big(n^{-1/2}\big)\quad\text{for } p\ge 2.$$

合起来：$\bar{X}_n-\mu=O_p\big(n^{-\min\{1-1/p,\,1/2\}}\big)$。等价地，$n^{1-1/p}(\bar{X}_n-\mu)=O_p(1)$ $(1<p\le 2)$。

读法：率指数 $\min\{1-1/p,\,1/2\}$ 随可用矩 p 单调。$p=2$ 时 $1-1/p=1/2$ 与 CLT 的 $n^{-1/2}$ 接榫；$p\to 1^+$ 时 $1-1/p\to 0$，几乎不收敛（只有一阶矩，恰好够 WLLN 但慢如龟）。超过 $p=2$ 再多的矩不再加速率（仍 $n^{-1/2}$，只让 CLT 的正态逼近更准）。

Example (2025 Q1(b)：仅 p 阶矩时 $\bar{x}$ 的收敛率).

原题 (2025 Q1b, Pinkse)

Establish the convergence rate of $\bar{x}$ if $\bar{x}$ is the sample mean of an iid sequence with $0<\mathbb{E}(\|x\|^p)<\infty$ for some $1<p<\infty$. The answer should be a function of p and work for all such p.

Solution.

[STRUCTURE —know the steps, fudge details]

答案：$\bar{x}-\mu=O_p\big(n^{-\min\{1-1/p,\,1/2\}}\big)$，即率为 $n^{-(1-1/p)}$ 当 $1<p\le 2$、率为 $n^{-1/2}$ 当 $p\ge 2$。

$p\ge 2$ 的情形（有二阶矩）：此时 $\mathrm{Var}(x_1)<\infty$，CLT 给 $\sqrt{n}(\bar{x}-\mu)\xrightarrow{d}N(0,\sigma^2)$，由 Fact 4.5(iv) 得 $\sqrt{n}(\bar{x}-\mu)=O_p(1)$，即 $\bar{x}-\mu=O_p(n^{-1/2})$。

$1<p\le 2$ 的情形（可能无穷方差）：设 $\mu=0$（中心化）。记 $S_n=\sum_{i=1}^n x_i$。von Bahr–Esseen 不等式（2.15）给出，对独立、均值 0、p 阶矩有限（$1\le p\le 2$）的和，

$$\mathbb{E}(|S_n|^p)\;\le\;C_p\sum_{i=1}^{n}\mathbb{E}(|x_i|^p)\;=\;C_p\,n\,\mathbb{E}(|x_1|^p)\,.$$

于是对 $\bar{x}=S_n/n$，

$$\mathbb{E}\Big(\big|n^{1-1/p}\,\bar{x}\big|^p\Big)=\mathbb{E}\Big(\Big|\tfrac{S_n}{n^{1/p}}\Big|^p\Big)=\frac{\mathbb{E}(|S_n|^p)}{n}\le C_p\,\mathbb{E}(|x_1|^p)=O(1).$$

由 Fact 4.5(v)（矩界 $\Rightarrow O_p$，取 $a_n=1, r=p$ 作用于 $n^{1-1/p}\bar{x}$），$n^{1-1/p}\bar{x}=O_p(1)$，即 $\bar{x}=O_p(n^{-(1-1/p)})$。■

Remark (考点与口径).

> 此题正是 Thm 4.18。关键洞察：率 $=n^{-(1-1/p)}$ 对 $1<p\le 2$、$=n^{-1/2}$ 对 $p\ge 2$，合写 $n^{-\min\{1-1/p,1/2\}}$。证 $1<p\le 2$ 段用 von Bahr–Esseen 把 $\mathbb{E}(|S_n|^p)$ 控成 $O(n)$，再 Markov $\to O_p$。*若考场未学 von Bahr–Esseen*：可退一步只给答案 $+\,p\ge 2$ 的 *CLT* 论证 $+\,1<p<2$ 时方差可能无穷故 *CLT* 失效、率退化为 $n^{-(1-1/p)}$ 的定性说明，通常已得大部分分。

4.9 样本方差及其与样本均值之联合极限分布

[优先级 ●●● 高 High] 资格考足迹：2019, 2020, 2021, 2025

资格考最稳的一类题：求 $\hat{\sigma}^2$ 的极限、或 $(\bar{X},\hat{\sigma}^2)$ 的联合极限。套路统一——把样本矩写成 iid 量之和，多变量 CLT + delta 法/Slutsky 收尾。需要四阶矩。

Theorem 4.19: 样本方差的渐近分布

[REPRODUCE —memorize this proof]

Let $\{X_i\}$ be iid with $\mathbb{E}(X_1^4)<\infty$, $\mu=\mathbb{E}(X_1)$, $\sigma^2=\mathrm{Var}(X_1)>0$. Let $\hat{\sigma}^2=n^{-1}\sum_i(X_i-\bar{X}_n)^2$. Then

$$\sqrt{n}\big(\hat{\sigma}^2-\sigma^2\big)\;\xrightarrow{d}\;N\big(0,\;\mu_4-\sigma^4\big),\qquad \mu_4:=\mathbb{E}\big((X_1-\mu)^4\big)=\mathrm{Var}\big((X_1-\mu)^2\big)+\sigma^4.$$

故极限方差 $=\mathrm{Var}\big((X_1-\mu)^2\big)=\mu_4-\sigma^4$。（用 $S^2_{Xn}=\frac{n}{n-1}\hat{\sigma}^2$ 替换不改变极

限。）

Proof for Theorem

[REPRODUCE —memorize this proof]

关键分解（避开未知 μ 直接处理）：

$$\hat{\sigma}^2 = \frac{1}{n}\sum_i (X_i-\mu)^2 - (\bar{X}_n-\mu)^2.$$

乘 $\sqrt{n}$ 并中心化：

$$\sqrt{n}(\hat{\sigma}^2-\sigma^2) = \underbrace{\frac{1}{\sqrt{n}}\sum_i \left[(X_i-\mu)^2-\sigma^2\right]}_{\text{(A)}} - \underbrace{(\bar{X}_n-\mu)\sqrt{n}(\bar{X}_n-\mu)}_{\text{(B)}}.$$

(A)：$\{(X_i-\mu)^2\}$ iid，均值 σ^2，方差 $\mathrm{Var}\big((X_1-\mu)^2\big)=\mu_4-\sigma^4$（需 $\mathbb{E}\big(X_1^4\big)<\infty$）。CLT：(A) $\xrightarrow{d} N(0,\mu_4-\sigma^4)$。**(B)**：$\bar{X}_n-\mu \xrightarrow{p} 0$（WLLN）而 $\sqrt{n}(\bar{X}_n-\mu)\xrightarrow{d} N(0,\sigma^2)=O_p(1)$，故 (B) $=o_p(1)\cdot O_p(1)=o_p(1)\xrightarrow{p}0$（Slutsky/Lemma 1b：常数 0 乘以收敛量 $\xrightarrow{d}0$）。由 Slutsky (a)：$\sqrt{n}(\hat{\sigma}^2-\sigma^2)=$ (A) − (B) $\xrightarrow{d} N(0,\mu_4-\sigma^4)+0$。■

“估计 μ 不影响渐近方差”是这题的灵魂：项 (B) 含的就是“用 $\bar{X}_n$ 替真 μ”的代价，但它是 $o_p(1)$——因为 $(\bar{X}_n-\mu)$ 已 $\xrightarrow{p}0$，再乘一个 $O_p(1)$ 仍可忽略。所以渐近上样本方差就好像 μ 已知、直接对 iid 量 $(X_i-\mu)^2$ 跑 CLT。

Example (2025 Q1(a) / 2020 Q5(a) / 2021 Q5(a)：$\sqrt{n}(\hat{\sigma}^2-\sigma_0^2)$ 的极限).

原题 (2025 Q1a, Pinkse)

Let $\hat{\sigma}^2, \sigma_0^2$ be the sample and population variance of an iid sample of size n with finite fourth moments. Derive the limit distribution of $\sqrt{n}(\hat{\sigma}^2-\sigma_0^2)$.

Solution.

直接是 Thm 4.19：

$$\sqrt{n}(\hat{\sigma}^2-\sigma_0^2) \xrightarrow{d} N\big(0,\ \mu_4-\sigma_0^4\big), \qquad \mu_4=\mathbb{E}\big((X_1-\mu)^4\big).$$

2020 Q5(a)/2021 Q5(a) 是同题加问“率 r”：$r=1/2$，且额外条件 $\mathbb{E}\big((X_1-\mu)^4\big)>\sigma_0^4$ 保证极限方差 $\mu_4-\sigma_0^4>0$（非退化）。■

这也可视为对二阶矩的 delta 法：$\hat{\sigma}^2=\hat{m}_2-\hat{m}_1^2$（$\hat{m}_j$ 为样本 j 阶矩），对 $g(m_1,m_2)=m_2-m_1^2$ 用多变量 delta，梯度在真值处 $G=(-2\mu,1)$，最终方差 $G\Sigma G'=\mu_4-\sigma^4$（与上等价）。

Example (2025 Q1(c)：$\limsup_n \mathrm{P}\big(n\hat\sigma^2 > z_n\big)$（两条 $\sqrt{n}$-涨落赛跑）).

原题 (2025 Q1c, Pinkse)

延续 Q1(a)（取 $\sigma_0^2 = 1$）。设 $z_n \sim \chi^2_n$ 且独立于 $\hat\sigma^2$。求 $\limsup_n \mathrm{P}\big(n\hat\sigma^2 > z_n\big)$。

Solution.

[STRUCTURE —know the steps, fudge details]

答案：1/2。关键是把两边都绕开 n 这个共同的一阶项、只比较各自的 $\sqrt{n}$-涨落。

第 1 步：两边各减 n，化成两个 $\sqrt{n}$-量。由 Thm 4.19（$\sigma_0^2 = 1$），$\sqrt{n}(\hat\sigma^2 - 1) \xrightarrow{d} N(0, \mu_4 - 1)$，故

$$A_n := \sqrt{n}\big(\hat\sigma^2 - 1\big) = \frac{n\hat\sigma^2 - n}{\sqrt{n}} \xrightarrow{d} A \sim N(0,\, \mu_4 - 1).$$

对 $z_n \sim \chi^2_n$（可写成 n 个 iid χ^2_1 之和，均值 n、方差 $2n$），由 CLT

$$B_n := \frac{z_n - n}{\sqrt{n}} \xrightarrow{d} B \sim N(0,\, 2).$$

第 2 步：作差，赛跑。因 $n\hat\sigma^2 - n = \sqrt{n}\,A_n$、$z_n - n = \sqrt{n}\,B_n$，共同的 n 相消：

$$\mathrm{P}\big(n\hat\sigma^2 > z_n\big) = \mathrm{P}\big(n\hat\sigma^2 - n > z_n - n\big) = \mathrm{P}\big(\sqrt{n}\,A_n > \sqrt{n}\,B_n\big) = \mathrm{P}(A_n - B_n > 0)\,.$$

z_n 独立于 $\hat\sigma^2$ 给出 $A_n \perp B_n$，故 $(A_n, B_n)' \xrightarrow{d} N(0, \mathrm{Diag}(\mu_4 - 1,\, 2))$，由 CMT（$h(a,b) = a - b$ 连续）

$$A_n - B_n \xrightarrow{d} A - B \sim N\big(0,\ (\mu_4 - 1) + 2\big).$$

第 3 步：对称极限给 1/2。极限 $A - B$ 是均值 0、方差 $(\mu_4 - 1) + 2 > 0$ 的正态，0 是其连续点，故

$$\mathrm{P}(A_n - B_n > 0) \longrightarrow \mathrm{P}(A - B > 0) = \tfrac{1}{2}.$$

极限存在故 lim sup 即该极限：$\boxed{\limsup_n \mathrm{P}\big(n\hat\sigma^2 > z_n\big) = \tfrac{1}{2}.}$ ■

Remark (本题精髓：减掉公共漂移、比 $\sqrt{n}$-涨落).

> $n\hat\sigma^2$ 与 z_n 都以相同的一阶漂移 n 发散（$\hat\sigma^2 \xrightarrow{p} 1$、$\mathbb{E}(z_n) = n$），直接比大小会被 $\pm\infty$ 淹没。诀窍是两边同减 n，余下的恰是各自的 $\sqrt{n}$-涨落 A_n, B_n，二者都 $\xrightarrow{d}$ 到中心 0 的正态。独立性把它们拼成对角协方差的联合正态，差 $A_n - B_n$ 仍是对称的均值-0 正态——胜负各半，故概率 $\to 1/2$，与 μ_4 的具体值无关（只要 $\mu_4 < \infty$ 保证非退化）。这类“两条同阶发散量赛跑 $\to 1/2$”是 Pinkse 偏爱的考法。

Example (2019 Q3(b) / 2020 Q5 / 2021 Q5：$(\bar{X}_n, \hat{\sigma}^2)$ 的联合极限分布).

原题 (2019 Comp Q3b, Guggenberger)

$X_1, \dots, X_n$ iid, mean μ, variance σ^2. Find the asymptotic distribution of $\sqrt{n}\big(\bar{X}_n - \mu,\ \hat{\sigma}_n^2 - \sigma^2\big)'$. State assumptions.

Solution.

[STRUCTURE —know the steps, fudge details]

假设：$\mathbb{E}\big(X_1^4\big) < \infty$（联合协方差含四阶矩）。设 $\mu_3 = \mathbb{E}\big((X_1-\mu)^3\big), \mu_4 = \mathbb{E}\big((X_1-\mu)^4\big)$。

构造 iid 向量. 渐近上 $\hat{\sigma}^2$ 等价于 $\frac{1}{n}\sum(X_i-\mu)^2$（估计 μ 的代价 $o_p(1)$，见 Thm 4.19 证）。故考虑 iid 二维向量

$$V_i = \begin{pmatrix} X_i - \mu \\ (X_i-\mu)^2 - \sigma^2 \end{pmatrix}, \qquad \mathbb{E}(V_i) = 0.$$

其协方差矩阵

$$\Sigma = \mathrm{Var}(V_1) = \begin{pmatrix} \mathbb{E}\big((X_1-\mu)^2\big) & \mathbb{E}\big((X_1-\mu)^3\big) \\ \mathbb{E}\big((X_1-\mu)^3\big) & \mathrm{Var}\big((X_1-\mu)^2\big) \end{pmatrix} = \begin{pmatrix} \sigma^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma^4 \end{pmatrix}.$$

（右上元 $\mathrm{Cov}\big(X_1-\mu, (X_1-\mu)^2\big) = \mathbb{E}\big((X_1-\mu)^3\big) = \mu_3$；右下 $\mathrm{Var}\big((X_1-\mu)^2\big) = \mu_4 - \sigma^4$。）

多变量 CLT + 收尾. $\sqrt{n}\, n^{-1}\sum_i V_i \xrightarrow{d} N(0,\Sigma)$。而

$$\sqrt{n}\begin{pmatrix} \bar{X}_n - \mu \\ \hat{\sigma}^2 - \sigma^2 \end{pmatrix} = \sqrt{n}\,\frac{1}{n}\sum_i V_i \ + \ \begin{pmatrix} 0 \\ -(\bar{X}_n - \mu)\sqrt{n}(\bar{X}_n - \mu) \end{pmatrix},$$

第二项是 $o_p(1)$（同前），由 Slutsky 不改变极限。故

$$\boxed{\sqrt{n}\begin{pmatrix} \bar{X}_n - \mu \\ \hat{\sigma}^2 - \sigma^2 \end{pmatrix} \xrightarrow{d} N\left(0,\ \begin{pmatrix} \sigma^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma^4 \end{pmatrix}\right).}$$

Remark (读三个边际/相关结论).

(1) 第一分量边际 $\to N(0,\sigma^2)$（CLT）；(2) 第二分量边际 $\to N(0, \mu_4 - \sigma^4)$（Thm 4.19）；(3) 二者*渐近独立* $\iff \mu_3 = 0$，即 X 的分布关于 μ 对称（如正态）时 $\bar{X}_n$ 与 $\hat{\sigma}^2$ 渐近独立——这正是正态样本下 $\bar{X}$ 与 S^2 独立的渐近版本。2020/2021 的 Q5 把这道拆成 (a)$\hat{\sigma}^2$、(b)$\bar{X}^2$ 两小问，方法同此。

4.10 样本均值之非线性函数：delta 法 + $o_p(1)$ 线性化

[优先级 ●●● 高 High] 资格考足迹：2018, 2019, 2022

实际估计量常是样本平均之比值/指数/对数。处理三步走：(1) 多变量 CLT 给平均向量的联合正态；(2) delta 法传到非线性变换；(3) 必要时证"真估计量 − 其线性化 $= o_p(1)/\sqrt{n}$"，把极限搬过去。

Example (2022 Q3(c)(d)：非线性函数的 delta 法 + $o_p(1)$ 等价).

原题 (2022 Q3, Pinkse)

$Y_i = \beta X_i + \varepsilon_i$, (X_i, ε_i) iid, $X_i \perp \varepsilon_i$, $\mathbb{E}(X_i) = \mu \neq 0$, $\mathbb{E}(\varepsilon_i) = 0$. Let $M_n = \big(n^{-1}\sum X_i\varepsilon_i,\ n^{-1}\sum \varepsilon_i\big)'$, $\tilde{T}_n = e^{2\beta}\exp\big(\frac{n^{-1}\sum X_i\varepsilon_i}{\mathbb{E}(X_i^2)} + \frac{n^{-1}\sum \varepsilon_i}{\mu}\big)$, $T_n = \exp(\hat{\beta}+\bar{\beta})$ with $\hat{\beta} = \frac{\sum X_iY_i}{\sum X_i^2}$, $\bar{\beta} = \frac{\sum Y_i}{\sum X_i}$. (c) Derive the asymptotic distribution of $\tilde{T}_n$. (d) Show $\sqrt{n}(\tilde{T}_n - T_n) = o_p(1)$ and give the asymptotic distribution of T_n.

Solution.

[STRUCTURE —know the steps, fudge details]

铺垫：先求 M_n 的联合极限。 $\{(X_i\varepsilon_i, \varepsilon_i)'\}$ iid，均值 $(\mathbb{E}(X\varepsilon), \mathbb{E}(\varepsilon))' = (0,0)'$（用 $X \perp \varepsilon, \mathbb{E}(\varepsilon) = 0$）。协方差：设 $\sigma_\varepsilon^2 = \mathrm{Var}(\varepsilon)$。

$$\begin{aligned}\mathrm{Var}(X\varepsilon) &= \mathbb{E}\big(X^2\varepsilon^2\big) = \mathbb{E}\big(X^2\big)\,\sigma_\varepsilon^2,\\ \mathrm{Cov}(X\varepsilon, \varepsilon) &= \mathbb{E}\big(X\varepsilon^2\big) = \mathbb{E}(X)\ \sigma_\varepsilon^2 = \mu\sigma_\varepsilon^2,\\ \mathrm{Var}(\varepsilon) &= \sigma_\varepsilon^2.\end{aligned}$$

多变量 CLT：$\sqrt{n}M_n \xrightarrow{d} N(0,\Omega)$，$\Omega = \sigma_\varepsilon^2\begin{pmatrix}\mathbb{E}\big(X^2\big) & \mu\\ \mu & 1\end{pmatrix}$.

(c) Delta 法. 写 $\tilde{T}_n = e^{2\beta}\,g(M_n)$，$g(a,b) = \exp\big(\frac{a}{\mathbb{E}(X^2)} + \frac{b}{\mu}\big)$。在 M_n 的中心 $(0,0)$ 处 $g(0,0) = 1$，梯度

$$G = \nabla g(0,0) = \Big(\tfrac{1}{\mathbb{E}(X^2)},\ \tfrac{1}{\mu}\Big)\cdot g(0,0) = \Big(\tfrac{1}{\mathbb{E}(X^2)},\ \tfrac{1}{\mu}\Big).$$

delta 法（中心 $(0,0)$，$\tilde{T}_n = e^{2\beta}g(M_n)$，$e^{2\beta}g(0,0) = e^{2\beta}$）：

$$\sqrt{n}\big(\tilde{T}_n - e^{2\beta}\big)\ \xrightarrow{d}\ e^{2\beta}\,G\,N(0,\Omega)\ =\ N\big(0,\ e^{4\beta}\,G\Omega G'\big),$$

其中

$$G\Omega G' = \sigma_\varepsilon^2\Big[\frac{\mathbb{E}(X^2)}{\mathbb{E}(X^2)^2} + \frac{2\mu}{\mathbb{E}(X^2)\mu} + \frac{1}{\mu^2}\Big] = \sigma_\varepsilon^2\Big[\frac{1}{\mathbb{E}(X^2)} + \frac{2}{\mathbb{E}(X^2)} + \frac{1}{\mu^2}\Big] = \sigma_\varepsilon^2\Big(\frac{3}{\mathbb{E}(X^2)} + \frac{1}{\mu^2}\Big).$$

（化简用 Ω 三元：$G\Omega G' = \sigma_\varepsilon^2[\,G_1^2\mathbb{E}\big(X^2\big) + 2G_1G_2\mu + G_2^2\,]$，代 $G_1 = 1/\mathbb{E}\big(X^2\big)$，$G_2 = 1/\mu$ 即得上式。）

(d) 证 $\sqrt{n}(\tilde{T}_n - T_n) = o_p(1)$. 思路：$\hat{\beta}, \bar{\beta}$ 都是 β 的一致估计，其偏差恰好被

$\tilde{T}_n$ 的指数项一阶捕捉。先线性化两估计量：

$$\hat{\beta}-\beta=\frac{n^{-1}\sum X_i\varepsilon_i}{n^{-1}\sum X_i^2}=\frac{n^{-1}\sum X_i\varepsilon_i}{\mathbb{E}(X^2)}+o_p(n^{-1/2}),$$

因分母 $n^{-1}\sum X_i^2\xrightarrow{p}\mathbb{E}\big(X^2\big)$（WLLN），分子 $=O_p(n^{-1/2})$，用 $\frac{1}{\hat{a}}=\frac{1}{a}+o_p(1)$ 得替换误差 $o_p(n^{-1/2})$。同理

$$\bar{\beta}-\beta=\frac{n^{-1}\sum\varepsilon_i}{n^{-1}\sum X_i}=\frac{n^{-1}\sum\varepsilon_i}{\mu}+o_p(n^{-1/2})$$

（用 $Y_i=\beta X_i+\varepsilon_i$ 故 $\sum Y_i=\beta\sum X_i+\sum\varepsilon_i$，$\bar{\beta}=\beta+\frac{\sum\varepsilon_i}{\sum X_i}$）。相加：

$$\hat{\beta}+\bar{\beta}=2\beta+\Big(\tfrac{n^{-1}\sum X_i\varepsilon_i}{\mathbb{E}(X^2)}+\tfrac{n^{-1}\sum\varepsilon_i}{\mu}\Big)+o_p(n^{-1/2}).$$

故 $T_n=\exp(\hat{\beta}+\bar{\beta})=e^{2\beta}\exp(\cdots+o_p(n^{-1/2}))=\tilde{T}_n\cdot e^{o_p(n^{-1/2})}=\tilde{T}_n(1+o_p(n^{-1/2}))$。因 $\tilde{T}_n=O_p(1)$（$\xrightarrow{d}$ 极限存在），

$$\sqrt{n}(T_n-\tilde{T}_n)=\sqrt{n}\,\tilde{T}_n\,o_p(n^{-1/2})=\tilde{T}_n\cdot o_p(1)=o_p(1).$$

由 Slutsky (a)，T_n 与 $\tilde{T}_n$ 享有同一极限：

$$\sqrt{n}\big(T_n-e^{2\beta}\big)\ \xrightarrow{d}\ N\big(0,\ e^{4\beta}\sigma_\varepsilon^2(\tfrac{3}{\mathbb{E}(X^2)}+\tfrac{1}{\mu^2})\big).\qquad\blacksquare$$

Remark (“$o_p(1)$ 线性化”范式).

(d) 的逻辑是渐近理论里最常用的“先线性化再搬运”：真估计量 T_n 难直接求分布，但它与一个**已知分布**的近似 $\tilde{T}_n$ 只差 $o_p(1)/\sqrt{n}$，于是 $\sqrt{n}(T_n-\tilde{T}_n)=o_p(1)$，Slutsky 让二者**渐近等价**、共享极限。**两个易错点**：(1) 替换分母（$\frac{1}{\hat{a}}\to\frac{1}{a}$）的误差必须确认是 $o_p(n^{-1/2})$ 级而非 $O_p(n^{-1/2})$，否则会改变极限；(2) exp 的线性化用 $e^{x+\delta}=e^x(1+o_p(1))$，再乘 $O_p(1)$ 的 $\tilde{T}_n$ 才得 $o_p(1)$。

4.11 自测题 Self-Test

Example (Self-Test 1：收敛模式 true/false（2020 Q6 / 2024 Q6 综合）).

判断真假并简述/给反例：**(a)** $X_n\xrightarrow{d}X\Rightarrow h(X_n)\xrightarrow{d}h(X)$ 对连续 h。**(b)** $X_n\xrightarrow{d}0\Rightarrow X_n\xrightarrow{p}0$。**(c)** $\mathbb{E}\big((X_n-Y)^2\big)\to0\Rightarrow\mathbb{E}\big(X_n^2\big)\to\mathbb{E}\big(Y^2\big)$。**(d)** $X_n\xrightarrow{p}0\Rightarrow\mathrm{P}(X_n\le0)\to1$。

Solution.

(a) 真——CMT（Thm 4.14）。**(b) 真**——依分布到常数等价于依概率到该常数（Lemma 2）。**(c) 真**（前提 $Y \in L^2$，即 $\mathbb{E}(Y^2) < \infty$，才能谈 $\|Y\|_2$；这是 L^2 收敛的默认语境）——L^2 收敛即范数 $\|X_n\|_2 \to \|Y\|_2$（逆三角不等式 $|\,\|X_n\|_2 - \|Y\|_2\,| \le \|X_n - Y\|_2 \to 0$），平方即得。**(d) 假**——依概率到 0 只管 $\mathrm{P}(|X_n| > \varepsilon) \to 0$；取 $X_n \sim N(0, 1/n)$，则 $X_n \xrightarrow{p} 0$ 但 $\mathrm{P}(X_n \le 0) = 1/2 \not\to 1$（关于 0 对称）。

Example (Self-Test 2：$\bar{X}_n^2$ 的退化情形（2020/2021 Q5(b))).

$\{X_i\}$ iid，$\mathbb{E}(X_1^4) < \infty$，$\mu = \mathbb{E}(X_1), \sigma^2 = \mathrm{Var}(X_1) > 0$。求 $n^r(\hat{m}_1^2 - \mu^2)$ 的非退化极限（$\hat{m}_1 = \bar{X}_n$），找出合适的 r。

Solution.

对 $g(m) = m^2$ 用 delta 法，$g'(\mu) = 2\mu$。**若** $\mu \neq 0$：$r = 1/2$，$\sqrt{n}(\bar{X}_n^2 - \mu^2) \xrightarrow{d} 2\mu \cdot N(0, \sigma^2) = N(0, 4\mu^2\sigma^2)$。**若** $\mu = 0$：$g'(0) = 0$，一阶 delta 退化，用二阶 delta（$g'' = 2$）：$r = 1$，$n\bar{X}_n^2 = (\sqrt{n}\bar{X}_n)^2 \xrightarrow{d} (\sigma Z)^2 = \sigma^2\chi_1^2$。“合适的 r”依 μ 是否为 0 分两档，满分要把两档都点出。

Example (Self-Test 3：变异系数的渐近分布（延伸 2019 Q3）).

$\{X_i\}$ iid，$\mu \neq 0, \sigma^2 > 0, \mathbb{E}(X_1^4) < \infty$。$\hat{C}_n = \hat{\sigma}_n / \bar{X}_n$ 估变异系数 $C = \sigma/\mu$。说明 $\hat{C}_n$ 一致，并（提示性地）说出 $\sqrt{n}(\hat{C}_n - C)$ 的极限取决于哪些矩。

Solution.

一致性：$\bar{X}_n \xrightarrow{p} \mu$（WLLN），$\hat{\sigma}_n^2 \xrightarrow{p} \sigma^2$（WLLN+CMT），$\hat{\sigma}_n = \sqrt{\hat{\sigma}_n^2} \xrightarrow{p} \sigma$（CMT，$\sqrt{\cdot}$ 连续于 $\sigma^2 > 0$），故 $\hat{C}_n = \hat{\sigma}_n/\bar{X}_n \xrightarrow{p} \sigma/\mu = C$（Slutsky，$\mu \neq 0$）。**渐近分布**：$\hat{C}_n = g(\bar{X}_n, \hat{\sigma}_n^2)$，$g(m, v) = \sqrt{v}/m$。由 Example 4.9 的联合极限 $\sqrt{n}(\bar{X}_n - \mu, \hat{\sigma}^2 - \sigma^2)' \xrightarrow{d} N(0, \Sigma)$（$\Sigma$ 含 σ^2, μ_3, μ_4），对 g 用多变量 delta，梯度 $G = \left(-\sqrt{v}/m^2,\ \frac{1}{2m\sqrt{v}}\right)\Big|_{(\mu,\sigma^2)} = \left(-\sigma/\mu^2,\ \frac{1}{2\mu\sigma}\right)$，得 $\sqrt{n}(\hat{C}_n - C) \xrightarrow{d} N(0, G\Sigma G')$。故极限方差依赖到四阶矩 μ_4 与三阶矩 μ_3（通过 Σ）。

Example (Self-Test 4：O_p/o_p 演算).

设 $\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \sigma^2)$，$\hat{a}_n \xrightarrow{p} a \neq 0$。判断并化简：(a) $\widehat{\theta}_n - \theta_0 = ?$（$O_p$ 几何率）；(b) $(\widehat{\theta}_n - \theta_0)^2 = ?$；(c) $\hat{a}_n \cdot \sqrt{n}(\widehat{\theta}_n - \theta_0) = ?$（极限分布）。

Solution.

(a) $\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \sigma^2) \Rightarrow \sqrt{n}(\widehat{\theta}_n - \theta_0) = O_p(1)$（Fact 4.5iv），故 $\widehat{\theta}_n - \theta_0 = O_p(n^{-1/2})$。(b) $(O_p(n^{-1/2}))^2 = O_p(n^{-1})$。(c) $\hat{a}_n \xrightarrow{p} a$ 故 $= a + o_p(1)$；Slutsky (b)：$\hat{a}_n \cdot \sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} a \cdot N(0, \sigma^2) = N(0, a^2\sigma^2)$。

Chapter 5　估计与极大似然（Estimation & Maximum Likelihood）

导读 · Roadmap（先读这里，不含公式）

前几章我们把概率搭好了地基（1），又把“样本量趋于无穷时统计量收敛到哪里、以多快速度、按什么分布”讲清了（4）。本章把这些工具对准一个真实目标：从数据里估出一个未知参数 θ_0，并问两件事——这个估计量准不准（有限样本的偏差与均方误差），以及它好不好（大样本下的相合性、渐近正态、有没有触到效率天花板）。

主线是**极大似然（MLE）**。它的哲学只有一句话：在所有候选参数里，挑出“最可能生成你手上这组数据”的那个。围绕这句话，资格考反复考五件事：（1）写出似然 / 对数似然、求一阶条件解出 $\widehat{\theta}$；（2）用 *Jensen* 不等式 + *KLIC* 论证 MLE 相合；（3）背出渐近正态 $\sqrt{n}(\widehat{\theta}-\theta_0)\xrightarrow{d}N(0,I(\theta_0)^{-1})$ 并会用信息矩阵等式；（4）Cramér–Rao 下界与“MLE 渐近有效”；（5）几个经典陷阱——参数落在支撑边界时 MLE 不再是“求导置零”（$U[-\theta,\theta]$、平坦似然），以及给定协变量的条件 *MLE*。每一件都配一道 PSU 资格考原题：2025 的指数条件 MLE、2017 的不变性证明、2018 的均匀分布边界、2019 的二项 MLE、2024 关于“已知方差帮不帮得上估均值”的判断题。

5.1　估计量、偏差、MSE

[优先级 ●●○ 中 Med]　资格考足迹：2018, 2024

先把语言铺好。我们有一族候选分布 $\{f(\cdot;\theta):\theta\in\Theta\}$（参数族），数据 $\{Y_i\}_{i=1}^n$ 由其中某一个真值 θ_0 生成。**估计量**就是一个把数据映成参数空间一点的函数 $\widehat{\theta}_n=\widehat{\theta}_n(Y_1,\dots,Y_n)$——它本身是随机变量。衡量它好坏，最朴素的两把尺子是偏差和均方误差。

Definition 5.1: 无偏、偏差、MSE（Unbiasedness, bias, MSE）

For an estimator $\widehat{\theta}_n$ of a scalar $\theta\in\Theta$:

- *Bias*: $\mathrm{Bias}(\widehat{\theta}_n,\theta):=\mathbb{E}_\theta\left(\widehat{\theta}_n\right)-\theta$. The estimator is *unbiased* if $\mathbb{E}_\theta\left(\widehat{\theta}_n\right)=\theta$ for all $\theta\in\Theta$.

- *Mean-squared error*: $\mathrm{MSE}(\widehat{\theta}_n, \theta) := \mathbb{E}_\theta\Big((\widehat{\theta}_n - \theta)^2\Big)$.

($\mathbb{E}_\theta(\cdot)$ means expectation taken under the true value θ.)

Theorem 5.2: 偏差–方差分解（Bias–variance decomposition）

[REPRODUCE —memorize this proof]

For any estimator with finite second moment,

$$\mathrm{MSE}(\widehat{\theta}_n, \theta) \;=\; \mathrm{Var}_\theta\Big(\widehat{\theta}_n\Big) \;+\; \big[\mathrm{Bias}(\widehat{\theta}_n, \theta)\big]^2.$$

Proof for Theorem

[REPRODUCE —memorize this proof]

Write $\widehat{\theta}_n - \theta = (\widehat{\theta}_n - \mathbb{E}_\theta\Big(\widehat{\theta}_n\Big)) + (\mathbb{E}_\theta\Big(\widehat{\theta}_n\Big) - \theta)$, square, and take expectations. The cross term vanishes:

$$\mathbb{E}_\theta\left(2\,(\widehat{\theta}_n - \mathbb{E}_\theta\Big(\widehat{\theta}_n\Big))\,\underbrace{(\mathbb{E}_\theta\Big(\widehat{\theta}_n\Big) - \theta)}_{\text{constant}}\right) = 2(\mathbb{E}_\theta\Big(\widehat{\theta}_n\Big) - \theta)\,\mathbb{E}_\theta\Big(\widehat{\theta}_n - \mathbb{E}_\theta\Big(\widehat{\theta}_n\Big)\Big) = 0,$$

leaving $\mathbb{E}_\theta\Big((\widehat{\theta}_n - \mathbb{E}_\theta\Big(\widehat{\theta}_n\Big))^2\Big) + (\mathbb{E}_\theta\Big(\widehat{\theta}_n\Big) - \theta)^2 = \mathrm{Var}_\theta\Big(\widehat{\theta}_n\Big) + \mathrm{Bias}^2$. ■

这条分解是全章的“记账公式”：一个估计量的总不准 = 方差（抖动）+ 偏差 ^{2}（系统性偏移）。它解释了为什么有时稍微有偏的估计量反而更好——只要它把方差压下去得够多，总 MSE 反而更小。MLE 的 $\hat{\sigma}^2$ 就是典型：它对 σ^2 有偏（除以 n 而非 $n-1$），但这通常无伤大雅。

Definition 5.3: 相对效率（Relative efficiency）

For two estimators $\widehat{\theta}_{n,1}, \widehat{\theta}_{n,2}$ of θ, the *relative efficiency* of $\widehat{\theta}_{n,1}$ to $\widehat{\theta}_{n,2}$ is $\mathrm{MSE}(\widehat{\theta}_{n,2}, \theta) / \mathrm{MSE}(\widehat{\theta}_{n,1}, \theta)$. If both are unbiased this reduces to the variance ratio $\mathrm{Var}\Big(\widehat{\theta}_{n,2}\Big) / \mathrm{Var}\Big(\widehat{\theta}_{n,1}\Big)$; a value > 1 means $\widehat{\theta}_{n,1}$ is more efficient.

Remark (为什么不能直接“最小化 MSE 找最优估计量”).

因为 $\mathrm{MSE}(\widehat{\theta}_n, \theta)$ 依赖于未知的真值 θ。坏估计量 $\tilde{\theta}_n \equiv 13$（无论数据如何都报 13）在 $\theta = 13$ 处 MSE$= 0$，但别处巨大。所以“对每个 θ 都 MSE 最小”的估计量一般不存在——除非先用某种约束（如无偏）把这种作弊估计量排除掉。这正是下一节 UMVU 的动机。

Definition 5.4: 一致最小方差无偏（UMVU）

An estimator $\widehat{\theta}_n$ is *Uniformly Minimum Variance Unbiased* (UMVU) if (i) $\mathbb{E}_\theta\left(\widehat{\theta}_n\right) = \theta$ for all $\theta \in \Theta$, and (ii) $\mathrm{Var}_\theta\left(\widehat{\theta}_n\right) \le \mathrm{Var}_\theta\left(\tilde{\theta}_n\right)$ for every other unbiased $\tilde{\theta}_n$ and all $\theta \in \Theta$.

UMVU 是“先把候选缩到无偏估计量，再在里头找方差最小的”。课上给的三个例子里，$\bar{X}_n$ 是 $N(\mu,\sigma^2)$ 中 μ 的 UMVU，$S^2_{X,n} = \frac{1}{n-1}\sum(X_i-\bar{X})^2$ 是 σ^2 的 UMVU，$\hat{p} = \bar{X}_n$ 是 Bernoulli 中 p 的 UMVU。但 UMVU 有三个毛病——无偏太苛刻（可能存在略有偏但方差小得多的更好估计量）、要求对*所有* θ 一致最优导致它常常*不存在*、即便存在也常*难求*。资格考更偏爱的最优性语言是下面的 Cramér–Rao 与“MLE 渐近有效”。

5.1.1 用 MSE 证相合（MSE $\to 0$ + Markov）

证“估计量相合 $(\widehat{\theta}_n \xrightarrow{p} \theta)$”有两条标准路子：(A) WLLN + 连续映射 / Slutsky；(B) 证 MSE $\to 0$，再用 Markov 不等式。后者尤其干净，是答题的快通道。

Proposition 5.5: 由 MSE 推相合（Consistency via MSE）

[REPRODUCE —memorize this proof]

If $\mathrm{MSE}(\widehat{\theta}_n,\theta) \to 0$ as $n \to \infty$ for every $\theta \in \Theta$, then $\widehat{\theta}_n \xrightarrow{p} \theta$.

Proof for Theorem

[REPRODUCE —memorize this proof]

Fix $\varepsilon > 0$. By Markov's inequality applied to the nonnegative random variable $(\widehat{\theta}_n - \theta)^2$,

$$\mathrm{P}\left(|\widehat{\theta}_n - \theta| > \varepsilon\right) = \mathrm{P}\left((\widehat{\theta}_n-\theta)^2 > \varepsilon^2\right) \le \frac{\mathbb{E}_\theta\left((\widehat{\theta}_n-\theta)^2\right)}{\varepsilon^2} = \frac{\mathrm{MSE}(\widehat{\theta}_n,\theta)}{\varepsilon^2} \to 0.$$

Hence $\widehat{\theta}_n \xrightarrow{p} \theta$. ■

记住这条等价于“L^2 收敛 $\Rightarrow$ 依概率收敛”。它把一个*概率*陈述（难算）换成两个*矩*的陈述（好算）：只要偏差 $\to 0$ 且方差 $\to 0$，由分解 $\mathrm{MSE} = \mathrm{Var} + \mathrm{Bias}^2 \to 0$，相合立得。例如 $\bar{X}_n$：偏差恒为 0，方差 $\sigma^2/n \to 0$，故 $\bar{X}_n \xrightarrow{p} \mu$。

5.2 极大似然：似然、对数似然、一阶条件

[优先级 ●●● 高 High] 资格考足迹：2017, 2018, 2019, 2025

Definition 5.6: 似然与对数似然（Likelihood and log-likelihood）

Let $\{Y_i\}_{i=1}^n$ be iid with density $f(y;\theta_0)$ (w.r.t. a common dominating measure μ), $\theta_0 \in \Theta$. The *likelihood function* is the joint density viewed as a function of the parameter,

$$L_n(\theta) \;=\; \prod_{i=1}^{n} f(Y_i;\theta), \qquad \ell_n(\theta) := \log L_n(\theta) = \sum_{i=1}^{n} \log f(Y_i;\theta).$$

The *maximum likelihood estimator (MLE)* is any maximizer

$$\widehat{\theta}_n \;\in\; \arg\max_{\theta\in\Theta} L_n(\theta) \;=\; \arg\max_{\theta\in\Theta} \ell_n(\theta).$$

两点直觉。其一，$L_n(\theta)$ 不是“θ 的概率”——θ 是固定未知数，不是随机变量；$L_n(\theta)$ 是“*若*真值是 θ，则恰好观测到手上这组 Y 的可能性”。MLE 就是反过来问：哪个 θ 让“看到我所看到的”最不意外。其二，取对数把连乘变连加（积分 / 求导都好办），且 log 严格递增、不改变 argmax，所以最大化 ℓ_n 与 L_n 完全等价。

答 MLE 题的标准三步（Likelihood → FOC → solve）

Step 1. Write $L_n(\theta) = \prod_i f(Y_i;\theta)$ and take logs: $\ell_n(\theta) = \sum_i \log f(Y_i;\theta)$.
Step 2 (interior case). Take the *score* $S_n(\theta) := \partial\ell_n/\partial\theta$, set $S_n(\widehat{\theta}_n) = 0$, solve. Check it's a max (e.g. $\partial^2\ell_n/\partial\theta^2 < 0$).
Step 3 (boundary case). If ℓ_n has no interior stationary point—e.g. ℓ_n is monotone in θ on the feasible set, as in support-boundary problems—do *not* differentiate. Maximize directly by inspecting the constraint set (见 5.7 节).

Remark (一阶条件不是万能的——这是资格考最爱埋的雷).

> “求导置零”只在*真值落在参数空间内部、似然光滑*时有效。一旦参数进入分布的*支撑边界*（如 $U[0,\theta]$ 里 θ 卡着数据的最大值），似然在最优点处*不可导甚至不连续*，FOC 给出错误答案。5.7.1 节的 2018 Q4(a) 就是要你识别这一点。

5.2.1 例：Bernoulli 与正态

Example (热身：$\hat{p} = \bar{X}_n$ for iid Bernoulli).
$Y_i \overset{i.i.d.}{\sim} \text{Bern}(p)$, $p \in [0,1]$. 求 MLE。

Solution.
$f(y;p) = p^y(1-p)^{1-y}$，故 $\ell_n(p) = (\sum_{i=1}^n Y_i)\log p + (n - \sum_{i=1}^n Y_i)\log(1-p)$。score

$$S_n(p) = \frac{\sum_{i=1}^n Y_i}{p} - \frac{n - \sum_{i=1}^n Y_i}{1-p} \overset{!}{=} 0 \implies \hat{p} = \frac{1}{n}\sum_{i=1}^n Y_i = \bar{X}_n.$$

二阶导 <0，确为极大。这就是 2019 Q5(a) 的（多观测版）骨架，下一节直接用到。

Example（正态：$\hat{\mu}=\bar{X}_n$，$\hat{\sigma}^2$ 有偏）.
$Y_i \overset{i.i.d.}{\sim} N(\mu,\sigma^2)$，$\theta=(\mu,\sigma^2)$。求 MLE。

Solution.
$\ell_n=-\frac{n}{2}\log(2\pi)-\frac{n}{2}\log\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^n(Y_i-\mu)^2$。对 μ 求导置零得 $\hat{\mu}=\bar{X}_n$；代回对 σ^2 求导置零得

$$\hat{\sigma}^2_{\mathrm{ML}}=\frac{1}{n}\sum_{i=1}^n(Y_i-\bar{X}_n)^2.$$

注意分母是 n（不是 $n-1$），故 $\mathbb{E}\big(\hat{\sigma}^2_{\mathrm{ML}}\big)=\frac{n-1}{n}\sigma^2\neq\sigma^2$：*MLE 的方差估计是有偏的*。这是偏差–方差分解（Thm 5.2）的活样本——MLE 用一点偏差换来了别的好性质。

5.3　MLE 的相合性：Jensen 与 KLIC 论证

[优先级 ●●● 高 High]　资格考足迹：2018, 2025

为什么“最大化似然”能命中真值？核心是一条信息论事实：在所有 θ 里，真值 θ_0 唯一地最小化“到真分布的 Kullback–Leibler 距离”。把这翻译成大样本语言，就是 MLE 相合的证明。

Theorem 5.7: MLE 相合的总体识别（Jensen / KLIC）

[STRUCTURE —know the steps, fudge details]

Let $\{Y_i\}$ be iid with true density $f(\cdot;\theta_0)$. Define the (normalized) population criterion

$$Q(\theta):=\mathbb{E}_{\theta_0}\left(\log f(Y_i;\theta)\right),\qquad Q_n(\theta):=\frac{1}{n}\sum_{i=1}^n\log f(Y_i;\theta).$$

Then θ_0 *maximizes* Q over Θ:

$$Q(\theta_0)-Q(\theta)=\mathbb{E}_{\theta_0}\left(\log\frac{f(Y_i;\theta_0)}{f(Y_i;\theta)}\right)=\mathrm{KL}(f_{\theta_0}\,\|\,f_\theta)\;\geq\;0,$$

with equality iff $f(\cdot;\theta)=f(\cdot;\theta_0)$ a.e. If in addition the model is *identified* ($\mathrm{P}_{\theta_0}(f(Y_i;\theta)\neq f(Y_i;\theta_0))>0$ for all $\theta\neq\theta_0$), the maximizer is *unique*: $\theta_0=\arg\max_\theta Q(\theta)$.

Proof for Theorem

[STRUCTURE —know the steps, fudge details]

Consider $Q(\theta) - Q(\theta_0) = \mathbb{E}_{\theta_0}\left(\log \frac{f(Y_i;\theta)}{f(Y_i;\theta_0)}\right)$. Since log is *concave*, Jensen's inequality gives

$$\begin{aligned}\mathbb{E}_{\theta_0}\left(\log \frac{f(Y_i;\theta)}{f(Y_i;\theta_0)}\right) &\le \log \mathbb{E}_{\theta_0}\left(\frac{f(Y_i;\theta)}{f(Y_i;\theta_0)}\right) = \log \int \frac{f(y;\theta)}{f(y;\theta_0)}\, f(y;\theta_0)\, d\mu(y) \\ &= \log \int f(y;\theta)\, d\mu(y) = \log 1 = 0.\end{aligned}$$

Hence $Q(\theta) \le Q(\theta_0)$, i.e. θ_0 maximizes Q. The inequality in Jensen is *strict* unless the ratio $f(Y_i;\theta)/f(Y_i;\theta_0)$ is a.s. constant ($= 1$ here, since its mean is 1); under identification this fails for $\theta \neq \theta_0$, so the maximizer is unique. ■

Theorem 5.8: MLE 相合（Consistency of the MLE）

[STRUCTURE —know the steps, fudge details]

Under the setup of Thm 5.7 plus standard regularity conditions (compact Θ, θ_0 identified, continuity of $\theta \mapsto \log f(y;\theta)$, and a dominating integrable envelope so that the convergence below is uniform), the MLE $\widehat{\theta}_n = \arg\max_\theta Q_n(\theta)$ satisfies $\widehat{\theta}_n \xrightarrow{p} \theta_0$.

Proof for Theorem

[STRUCTURE —know the steps, fudge details]

Four steps (the generic "consistency of extremum estimators"argument):

(i) $\widehat{\theta}_n$ maximizes the *random* criterion $Q_n(\theta) = \frac{1}{n}\sum_i \log f(Y_i;\theta)$.

(ii) By the WLLN, $Q_n(\theta) \xrightarrow{p} Q(\theta)$ pointwise for each θ (provided $\mathbb{E}_{\theta_0}(|\log f(Y_i;\theta)|) < \infty$); regularity upgrades this to *uniform* convergence $\sup_\theta |Q_n(\theta) - Q(\theta)| \xrightarrow{p} 0$.

(iii) Uniform convergence of the criterion transfers to convergence of the argmax: the maximizer $\widehat{\theta}_n$ of Q_n converges to the maximizer of Q.

(iv) By Thm 5.7 that maximizer is the unique point θ_0. Hence $\widehat{\theta}_n \xrightarrow{p} \theta_0$. ■

把骨架记牢：随机准则 Q_n 一致逼近确定准则 Q，而 Q 由 *Jensen* 唯一地在 θ_0 取最大，于是 Q_n 的最大点逼近 θ_0。Jensen 的方向之所以"恰好对"：log 凹 ⇒ 期望的对数 ≥ 对数的期望，让 θ_0 坐上最大值的宝座。这里没有用到任何一阶条件——相合性是全局识别的结论，与下一节靠 FOC 展开的渐近正态是两套独立的论证。

Remark (KLIC 是什么).

$\mathrm{KL}(f_{\theta_0}\|f_\theta)=\mathbb{E}_{\theta_0}\left(\log\frac{f_{\theta_0}}{f_\theta}\right)\geq 0$ 是真分布 f_{θ_0} 到候选 f_θ 的 *Kullback–Leibler* 散度，由上证 ≥ 0 且仅 $f_\theta=f_{\theta_0}$ 时为 0。MLE 的总体目标 $-Q(\theta)$ 等价于最小化 KLIC——“极大似然 = 在参数族里找离真分布 KL 最近的那个分布”。模型设定错误（真分布不在族内）时这套照样跑，得到的 $\theta^*=\arg\min_\theta \mathrm{KL}$ 称为伪真值，这就是拟极大似然（QMLE）的根。

5.4　MLE 渐近正态与信息矩阵等式

[优先级 ●●● 高 High]　资格考足迹：2019, 2024, 2025

相合只说“最终落到 θ_0”，不说以多快、按什么分布。把 FOC 在 θ_0 处做一阶 Taylor 展开，就榨出 $\sqrt{n}(\widehat{\theta}-\theta_0)$ 的极限分布。结论既漂亮又是必背模板。

Definition 5.9: Score、Hessian、Fisher 信息（一个观测）

For one observation, define the *score* $s(Y;\theta):=\dfrac{\partial}{\partial\theta}\log f(Y;\theta)$ and the *Fisher information*

$$I(\theta_0)\;:=\;\mathbb{E}_{\theta_0}\Big(s(Y;\theta_0)\,s(Y;\theta_0)^\top\Big)\;=\;\mathrm{Var}_{\theta_0}(s(Y;\theta_0))\overset{(\star)}{=}-\mathbb{E}_{\theta_0}\left(\frac{\partial^2}{\partial\theta\,\partial\theta^\top}\log f(Y;\theta_0)\right).$$

Lemma 5.10: Score 均值为零 + 信息矩阵等式（Information matrix equality）

[REPRODUCE —memorize this proof]

Under regularity conditions permitting differentiation under the integral sign:

$$\textbf{(零均值)}\quad \mathbb{E}_{\theta_0}(s(Y;\theta_0))=0;$$

$$\textbf{(IME)}\quad \underbrace{\mathrm{Var}_{\theta_0}(s(Y;\theta_0))}_{\text{score 方差= 外积}\mathbb{E}(ss^\top)}=\underbrace{-\,\mathbb{E}_{\theta_0}\left(\frac{\partial^2}{\partial\theta\,\partial\theta^\top}\log f(Y;\theta_0)\right)}_{\text{负 Hessian 形式}}=I(\theta_0).$$

Proof for Theorem

[REPRODUCE —memorize this proof]

Start from the identity $\int f(y;\theta)\,d\mu(y)=1$ for *all* θ. Differentiate w.r.t. θ under the integral sign:

$$0=\frac{\partial}{\partial\theta}\int f(y;\theta)\,d\mu=\int\frac{\partial f}{\partial\theta}\,d\mu=\int\underbrace{\frac{\partial\log f}{\partial\theta}}_{s(y;\theta)}f(y;\theta)\,d\mu=\mathbb{E}_\theta(s(Y;\theta)),$$

using $\partial_\theta f=(\partial_\theta\log f)\,f$. This gives the zero-mean property, so $\mathrm{Var}(s)=\mathbb{E}\big(ss^\top\big)$.

Differentiate *again*:

$$0 = \frac{\partial}{\partial \theta^\top} \int s(y;\theta) f(y;\theta)\, d\mu = \int \Big[\frac{\partial s}{\partial \theta^\top} f + s\, \frac{\partial f}{\partial \theta^\top} \Big] d\mu = \mathbb{E}_\theta\Big(\frac{\partial^2 \log f}{\partial \theta \partial \theta^\top} \Big) + \mathbb{E}_\theta\big(s\, s^\top \big),$$

where the second term used $\partial_\theta f = s\, f$ again. Rearranging yields the IME. ■

信息矩阵等式说“一阶信息（score 的方差）= 二阶信息（对数似然曲率的期望）”。直觉：对数似然在 θ_0 处越尖锐弯曲（负 Hessian 大），数据对参数的定位能力越强（score 抖得越厉害），估出的 θ_0 就越准。两种算 $I(\theta_0)$ 的方式（外积 vs 负 Hessian）在设定正确时相等；设定错误时不等，二者之差正是“三明治”稳健方差的来源。

Theorem 5.11: MLE 渐近正态（Asymptotic normality of the MLE）

[STRUCTURE —know the steps, fudge details]

Suppose $\theta_0 \in \mathrm{int}(\Theta)$, $\log f(y;\theta)$ has continuous second derivatives, $\widehat{\theta}_n \xrightarrow{p} \theta_0$, the IME holds, and $I(\theta_0)$ is nonsingular. Then

$$\boxed{\sqrt{n}\,(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N\big(0,\ I(\theta_0)^{-1}\big).}$$

Proof for Theorem

[STRUCTURE —know the steps, fudge details]

Since $\theta_0 \in \mathrm{int}(\Theta)$ and $\widehat{\theta}_n \xrightarrow{p} \theta_0$, the FOC $\frac{1}{n}\sum_i s(Y_i;\widehat{\theta}_n) = 0$ holds w.p. $\to 1$. Mean-value expand the score about θ_0:

$$0 = \frac{1}{n}\sum_{i=1}^n s(Y_i;\widehat{\theta}_n) = \frac{1}{n}\sum_{i=1}^n s(Y_i;\theta_0) + \Big[\frac{1}{n}\sum_{i=1}^n \frac{\partial s}{\partial \theta^\top}(Y_i;\tilde{\theta}_n) \Big] (\widehat{\theta}_n - \theta_0),$$

with $\tilde{\theta}_n$ on the segment between $\widehat{\theta}_n$ and θ_0 (so $\tilde{\theta}_n \xrightarrow{p} \theta_0$). 向量参数下，单个中值 $\tilde{\theta}_n$ 只是启发式写法——严格版本对 score 的每一行各取一个中值点，并对 Hessian 用一致大数律 (uniform LLN) 把它们一并收敛到 θ_0；这里按 STRUCTURE 标记略去。 Rearrange:

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = \Big[-\frac{1}{n}\sum_{i=1}^n \frac{\partial s}{\partial \theta^\top}(Y_i;\tilde{\theta}_n) \Big]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n s(Y_i;\theta_0).$$

Two limits. (i) By the WLLN and $\tilde{\theta}_n \xrightarrow{p} \theta_0$, the bracketed matrix $\xrightarrow{p} -\mathbb{E}_{\theta_0}\big(\partial s/\partial \theta^\top\big) = I(\theta_0)$ (negative-Hessian form of IME). (ii) The scores $s(Y_i;\theta_0)$ are iid, mean 0 (Lem 5.10), variance $I(\theta_0)$; by the CLT,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n s(Y_i;\theta_0) \xrightarrow{d} N\big(0, I(\theta_0)\big).$$

By Slutsky, $\sqrt{n}(\widehat{\theta}_n-\theta_0)\xrightarrow{d} I(\theta_0)^{-1}N(0,I(\theta_0))=N\big(0,I(\theta_0)^{-1}I(\theta_0)I(\theta_0)^{-1}\big)=N\big(0,I(\theta_0)^{-1}\big)$. ■

答“求 MLE 渐近分布”题（背这套）

For a scalar parameter: (1) compute the score $s(Y;\theta)=\partial_\theta \log f$; (2) Fisher information $I(\theta)=-\mathbb{E}\big(\partial_\theta^2\log f\big)=\mathrm{Var}(s)$ (either form, IME); (3) state $\sqrt{n}(\widehat{\theta}-\theta_0)\xrightarrow{d}N(0,1/I(\theta_0))$; (4) (optional) consistent variance estimate $\widehat{\mathrm{Var}(\widehat{\theta})}=1/(n\,I(\widehat{\theta}))$ for a CI $\widehat{\theta}\pm z_{1-\alpha/2}/\sqrt{n\,I(\widehat{\theta})}$.

5.4.1　Worked example：2019 Q5 ——二项 MLE 及其渐近分布

原题 (Comprehensive Exam 2019, Guggenberger, Q5(a)(b))

Suppose $X\sim\mathrm{bin}(n,p)$ with pmf $f(x,p)=\binom{n}{x}p^x(1-p)^{n-x}$ for $x=0,\dots,n$ and $p\in[0,1]$. One observes a single random variable X.
(a) What is the maximum likelihood estimator of p?
(b) What is the asymptotic distribution of the ML estimator of p as $n\to\infty$?

Solution.

(a) 似然 $L(p)=\binom{n}{x}p^x(1-p)^{n-x}$；$\binom{n}{x}$ 不含 p，可丢。对数似然 $\ell(p)=x\log p+(n-x)\log(1-p)+\text{const}$。score

$$\ell'(p)=\frac{x}{p}-\frac{n-x}{1-p}\overset{!}{=}0\implies x(1-p)=(n-x)p\implies\boxed{\hat{p}=\frac{X}{n}.}$$

（端点 $X=0$ 给 $\hat{p}=0$、$X=n$ 给 $\hat{p}=1$，与公式一致。）

(b) 把 $X\sim\mathrm{bin}(n,p)$ 看成 n 个 iid $\mathrm{Bern}(p)$ 之和 $X=\sum_{i=1}^n B_i$，则 $\hat{p}=\bar{B}_n$ 是样本均值。$\mathbb{E}(B_i)=p$，$\mathrm{Var}(B_i)=p(1-p)$，由 CLT

$$\sqrt{n}(\hat{p}-p)\xrightarrow{d}N\big(0,\,p(1-p)\big).$$

对照信息矩阵：单个 Bernoulli 的 score $s(B;p)=\frac{B}{p}-\frac{1-B}{1-p}$，$I(p)=-\mathbb{E}(\partial_p s)=\mathbb{E}\Big(\frac{B}{p^2}+\frac{1-B}{(1-p)^2}\Big)=\frac{1}{p}+\frac{1}{1-p}=\frac{1}{p(1-p)}$。于是 $I(p)^{-1}=p(1-p)$，与上面 CLT 给的渐近方差完全一致——正是 Thm 5.11 的实例。■

Remark (考点提炼).

“单个 $\mathrm{bin}(n,p)$，$n\to\infty$”就是“n 个 Bernoulli 的均值”，所以 $\hat{p}=X/n$ 的渐近分布走 CLT；不必真去对 Stirling 展开二项 pmf。两条路（直接 CLT vs 信息矩阵 $I(p)^{-1}$）必须对上，对不上就是哪步算错了——这是最好的自查。

5.5　Cramér–Rao 下界与效率

[优先级 ●●○ 中 Med]　资格考足迹：2024

Fisher 信息 $I(\theta)$ 还扮演另一个角色：它给任何无偏估计量的方差设了一个无法逾越的地板。这就是 Cramér–Rao 下界，也是“MLE 渐近有效”的根据。

Theorem 5.12: Cramér–Rao 下界（Cramér–Rao lower bound, CRLB）

[STRUCTURE —know the steps, fudge details]

Let $\widehat{\theta}_n$ be any unbiased estimator of a scalar θ based on iid $\{Y_i\}_{i=1}^n$, under the regularity conditions of Lem 5.10. Then

$$\operatorname{Var}_\theta\left(\widehat{\theta}_n\right) \;\geq\; \frac{1}{n\,I(\theta)},$$

where $I(\theta)$ is the one-observation Fisher information. An unbiased estimator attaining this bound is *efficient*.

Proof for Theorem

[STRUCTURE —know the steps, fudge details]

Let $S_n(\theta) = \sum_i s(Y_i;\theta)$ be the total score; $\mathbb{E}(S_n) = 0$ and $\operatorname{Var}(S_n) = nI(\theta)$. Differentiating the unbiasedness identity $\mathbb{E}_\theta\left(\widehat{\theta}_n\right) = \int \widehat{\theta}_n(y)\prod_i f(y_i;\theta)\,d\mu = \theta$ under the integral sign,

$$1 = \frac{\partial}{\partial\theta}\mathbb{E}_\theta\left(\widehat{\theta}_n\right) = \mathbb{E}_\theta\left(\widehat{\theta}_n\, S_n(\theta)\right) = \operatorname{Cov}\left(\widehat{\theta}_n, S_n(\theta)\right)$$

(the last step uses $\mathbb{E}(S_n) = 0$). By Cauchy–Schwarz, $1 = \operatorname{Cov}\left(\widehat{\theta}_n, S_n\right)^2 \leq \operatorname{Var}\left(\widehat{\theta}_n\right)\operatorname{Var}(S_n) = \operatorname{Var}\left(\widehat{\theta}_n\right)\cdot nI(\theta)$, i.e. $\operatorname{Var}\left(\widehat{\theta}_n\right) \geq 1/(nI(\theta))$. ■

Corollary 5.13: MLE 渐近有效（Asymptotic efficiency）

The MLE's asymptotic variance (for $\sqrt{n}(\widehat{\theta}_n - \theta_0)$) is $I(\theta_0)^{-1}$ (Thm 5.11), so $\operatorname{Var}\left(\widehat{\theta}_n\right) \approx I(\theta_0)^{-1}/n = \left(nI(\theta_0)\right)^{-1}$, exactly the Cramér–Rao bound. Hence the MLE *asymptotically attains* the lower bound: no regular (asymptotically normal) estimator can have smaller asymptotic variance. This is the Hájek–Le Cam convolution theorem: any such competitor $\bar{\theta}_n$ has $\sqrt{n}(\bar{\theta}_n - \theta_0) \xrightarrow{d} \tilde{Z} + M$ with $\tilde{Z} \sim N(0, I^{-1})$ and M independent of $\tilde{Z}$, so its asymptotic variance is $I^{-1} + \operatorname{Var}(M) \geq I^{-1}$.

把三句话串起来：CRLB 给“地板”$1/(nI)$；MLE 渐近方差恰是 I^{-1}（除以 n 后正好踩地板）；故 MLE 是渐近最有效的。注意 CRLB 是有限样本对无偏估计量的下界，

而“MLE 有效”是渐近陈述（MLE 有限样本常有偏，如上面正态的 $\hat{\sigma}^2$）。下面 2024 Q6(d) 正是用“I 块对角 $\Rightarrow$ 知道 σ 不降低 μ 的下界”来作判断。

5.5.1 Worked example：2024 Q6(d) ——已知方差帮不帮得上估均值?

原题 (Comprehensive Exam 2024, Q6(d))

True or false, explain: Suppose we have an i.i.d. sample $\{X_1, \ldots, X_n\}$ with $X_1 \sim N(\mu_0, \sigma_0^2)$. If we knew σ_0, then we would be able to estimate μ_0 more accurately than we could without knowing σ_0.

Solution.

False. 关键是 $N(\mu, \sigma^2)$ 的信息矩阵在 (μ, σ^2) 间块对角。一个观测的对数密度 $\log f = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma^2 - \frac{(X-\mu)^2}{2\sigma^2}$，score 为

$$s_\mu = \frac{X-\mu}{\sigma^2}, \qquad s_{\sigma^2} = -\frac{1}{2\sigma^2} + \frac{(X-\mu)^2}{2\sigma^4}.$$

交叉信息 $\mathbb{E}(s_\mu s_{\sigma^2}) = \frac{1}{2\sigma^6}\mathbb{E}\big((X-\mu)^3\big) - \frac{1}{2\sigma^4}\mathbb{E}(X-\mu) = 0$（正态三阶中心矩 $= 0$，一阶 $= 0$）。于是

$$I(\mu, \sigma^2) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{pmatrix}, \qquad I^{-1} = \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}.$$

对 μ 的渐近方差就是 I^{-1} 的 (μ,μ) 元 $= \sigma^2$，与是否已知 σ 无关。直白说：不论 σ 已知与否，$\hat{\mu} = \bar{X}_n$ 都是 μ_0 的 MLE/有效估计量，方差 σ_0^2/n。块对角 $\Rightarrow$ 估 μ 与估 σ^2 信息正交，知道一个不改善另一个的精度。故命题为假。■

Remark (术语联系).

> 这体现的是 μ 与 σ^2 的参数正交（parameter orthogonality）：Fisher 信息块对角，故对一块参数的 CRLB 不因另一块已知/未知而改变——把一个未知参数“钉死”到真值，不会放松正交参数的下界。注意它与统计量的辅助性（ancillarity，指抽样分布不依赖参数的统计量）是两个不同概念——$\bar{X}_n$ 的分布依赖 μ，并非 ancillary；这里成立的只是参数间的信息正交。

5.6 MLE 的不变性

[优先级 ●●○ 中 Med] 资格考足迹：2017

不变性是 MLE 一条极好用又常考的性质：要估 θ 的一个函数 $\tau = g(\theta)$，不必重做优化，直接把 $\widehat{\theta}$ 代进 g 即可。

Theorem 5.14: MLE 的不变性（Invariance of the MLE）

[REPRODUCE —memorize this proof]

Let $\widehat{\theta}$ be an MLE of θ, and let $\tau = g(\theta)$ for any function g (*not* necessarily one-to-one). Then $g(\widehat{\theta})$ is an MLE of τ.

Proof for Theorem

[REPRODUCE —memorize this proof]

One-to-one g. If g is invertible, reparametrize $\theta = g^{-1}(\tau)$. The likelihood in the new parameter is $L^*(\tau) := L(g^{-1}(\tau))$. Then

$$\sup_{\tau} L^*(\tau) = \sup_{\tau} L(g^{-1}(\tau)) = \sup_{\theta} L(\theta) = L(\widehat{\theta}),$$

attained at τ with $g^{-1}(\tau) = \widehat{\theta}$, i.e. $\hat{\tau} = g(\widehat{\theta})$.

General g (induced / profile likelihood). When g need not be one-to-one, define the *induced (profile) likelihood* of τ as the largest likelihood consistent with that value of τ:

$$L^*(\tau) := \sup_{\{\theta:\, g(\theta)=\tau\}} L(\theta).$$

The MLE of τ maximizes L^*. Now

$$\sup_{\tau} L^*(\tau) = \sup_{\tau} \sup_{\{\theta: g(\theta)=\tau\}} L(\theta) = \sup_{\theta} L(\theta) = L(\widehat{\theta}),$$

since ranging over all τ and then over the preimage $\{g(\theta) = \tau\}$ is the same as ranging over all θ. The outer sup is attained at $\hat{\tau} = g(\widehat{\theta})$ (take the $\theta = \widehat{\theta}$ that achieves $\sup_\theta L$; it lies in the slice $\{g(\theta) = g(\widehat{\theta})\}$). Hence $g(\widehat{\theta})$ maximizes L^*, i.e. $g(\widehat{\theta})$ is an MLE of τ. ■

这就是 2017 Q6 的全部要求：陈述 + 证明。考场上 one-to-one 的情形（换元后 sup 不变）几乎人人会写；拿满分的关键是一般 g 的情形——必须先定义诱导似然（profile likelihood）$L^*(\tau) = \sup_{g(\theta)=\tau} L(\theta)$，否则“$g(\widehat{\theta})$ 是 τ 的 MLE”这句话在 g 不可逆时根本没有良定义的目标函数。给定定义后，把“先对 τ 取 sup、再对纤维取 sup”合并成“对所有 θ 取 sup”，结论一行落地。

Example (不变性的用法).

$Y_i \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$。求 $\sigma = \sqrt{\sigma^2}$ 的 MLE 与 $\mathrm{P}(Y_1 \le c) = \Phi(\frac{c-\mu}{\sigma})$ 的 MLE。

Solution.

由 $\hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{n}\sum(Y_i - \bar{Y})^2$ 及不变性，$\hat{\sigma}_{\mathrm{ML}} = \sqrt{\hat{\sigma}^2_{\mathrm{ML}}}\,(g(x) = \sqrt{x})$。又 $g(\mu, \sigma) = \Phi(\frac{c-\mu}{\sigma})$，故该概率的 MLE 是 $\Phi\big(\frac{c-\hat{\mu}}{\hat{\sigma}}\big)$——直接代入，无需重新优化。

5.7 边界参数的 MLE：FOC 失效的两类陷阱

[优先级 ●●● 高 High] 资格考足迹：2018

当参数把分布的支撑卡住时，似然在最优点处不可导，“求导置零”给出错误答案。资格考 2018 Q4 一次性考了两种典型：（a）似然在边界处单调、最大值落在约束端点；（b）似然在一段区间上完全平坦、MLE 不唯一（集值）。

5.7.1 Worked example：2018 Q4(a) ——$U[-\theta,\theta]$ 的 MLE

原题 (Comprehensive Exam 2018, Guggenberger, Q4(a))

Assume X_i are i.i.d. uniform $[-\theta,\theta]$, $i=1,\dots,n$, for some $\theta>0$. Find the MLE of θ.

Solution.

单个密度 $f(x;\theta)=\frac{1}{2\theta}\mathbb{1}\{-\theta\le x\le\theta\}=\frac{1}{2\theta}\mathbb{1}\{|x|\le\theta\}$。似然

$$L_n(\theta)=\prod_{i=1}^{n}\frac{1}{2\theta}\mathbb{1}\{|X_i|\le\theta\}=\left(\frac{1}{2\theta}\right)^n\mathbb{1}\{\theta\ge\max_i|X_i|\},$$

因为“所有 $|X_i|\le\theta$” $\iff$ “$\theta\ge\max_i|X_i|$”。**别求导**：在可行域 $\theta\ge\max_i|X_i|$ 上 $(1/2\theta)^n$ 关于 θ 严格递减，故 L_n 在最左端点最大：

$$\boxed{\widehat{\theta}=\max_{1\le i\le n}|X_i|.}$$

若取 $\theta<\max_i|X_i|$，某个 $|X_i|>\theta$ 落在支撑外，似然直接 $=0$（不可行）；取 $\theta>\max_i|X_i|$ 又把 $1/\theta$ 压小。最大点恰卡在边界。■

Remark (考点提炼).

支撑依赖参数 $\Rightarrow$ FOC 失效。这里 score $\partial_\theta\log L=-n/\theta<0$ 处处为负、永不为零，盲目置零会一无所获。正确机器：(1) 把指示函数收拢成对 θ 的单个约束 $\theta\ge\max_i|X_i|$；(2) 在约束集上看 $L_n(\theta)$ 的单调方向，最大点落在端点。顺带：$\widehat{\theta}=\max_i|X_i|$ 略低估 θ（样本极值进不到真边界），是有偏的，且收敛速率是超-$\sqrt{n}$ 的 n（$n(\theta-\widehat{\theta})\xrightarrow{d}$ 指数型），不走 Thm 5.11 的正态——边界参数的渐近论自成一派。

5.7.2　Worked example：2018 Q4(b) ——平坦似然、集值 MLE

原题 (Comprehensive Exam 2018, Guggenberger, Q4(b))

Suppose $Y_i = \theta + \varepsilon_i$, $i = 1, \dots, n$, where ε_i are i.i.d. uniform $[-1, 1]$ and $\theta \in (0, 1)$. What can you say about the MLE of θ?

Solution.

$\varepsilon_i = Y_i - \theta \sim U[-1, 1]$，故单个密度 $f(y;\theta) = \frac{1}{2}\mathbb{1}\{-1 \le y - \theta \le 1\} = \frac{1}{2}\mathbb{1}\{\theta - 1 \le y \le \theta + 1\}$。似然

$$L_n(\theta) = \left(\tfrac{1}{2}\right)^n \mathbb{1}\{\theta - 1 \le Y_i \le \theta + 1 \ \forall i\} = \left(\tfrac{1}{2}\right)^n \mathbb{1}\{\max_i Y_i - 1 \le \theta \le \min_i Y_i + 1\}.$$

约束 $\theta \ge Y_i - 1 \ \forall i \iff \theta \ge \max_i Y_i - 1$，且 $\theta \le Y_i + 1 \ \forall i \iff \theta \le \min_i Y_i + 1$。在这段区间上 $L_n \equiv (1/2)^n$ 完全平坦（不含 θ!）。结论：

> MLE 不唯一——区间 $\big[\max_i Y_i - 1,\ \min_i Y_i + 1\big] \cap (0, 1)$ 中任何一点都是 MLE。

似然没有唯一最大点，是集值的。■

Remark (考点提炼).

> 当误差密度平坦（uniform）且参数只是平移时，似然在一整段上恒定，MLE 退化为“与数据相容的整个区间”。这道题考的就是识别这一现象、说清“MLE 不唯一 / 集值”，并写出那个相容区间。注意区间非空（要 $\max_i Y_i - 1 \le \min_i Y_i + 1$，即 $\max_i Y_i - \min_i Y_i \le 2$，恒成立，因所有 Y_i 都落在某长度-2 区间 $[\theta_0 - 1, \theta_0 + 1]$ 内）。对照 (a)：(a) 似然单调、最大点是端点；(b) 似然平坦、最大点是一整段。两种都不是 FOC 内点解。Q4(c) 接着让你换用 GMM（拿 $\mathbb{E}(|Y_i|)$ 当矩条件）绕开这个识别困难，那是 GMM 章的内容。

5.8　条件 MLE（给定协变量）

[优先级 ●●● 高 High]　资格考足迹：2025

回归型数据 (Y_i, X_i) 里，我们关心 Y 给定 X 的条件分布 $f(y \mid x;\theta)$。条件 *MLE* 最大化条件对数似然——只要协变量的边际分布 $g(x)$ 不含 θ，它就会从似然里分离出去、对优化毫无影响。

Theorem 5.15: 条件 MLE：协变量密度分离（Conditional MLE）

[STRUCTURE —know the steps, fudge details]

Suppose (Y_i, X_i) are iid, $Y_i \mid X_i = x \sim f(y \mid x;\theta_0)$, and X_i has marginal density

$g(x)$ *not depending on* θ. The full log-likelihood factors as

$$\ell_n(\theta) = \sum_{i=1}^{n} \log f(Y_i \mid X_i; \theta) + \underbrace{\sum_{i=1}^{n} \log g(X_i)}_{\text{free of } \theta}.$$

Hence the (full-data) MLE of θ equals the *conditional* MLE $\widehat{\theta}_n = \arg\max_\theta \sum_i \log f(Y_i \mid X_i; \theta)$: the regressor distribution is irrelevant to the θ-argmax.

这就是"正态回归的 MLE = OLS"背后的机制：$Y_i \mid X_i \sim N(\alpha + \beta X_i, \sigma^2)$ 时，条件对数似然 $\propto -\frac{1}{2\sigma^2}\sum(Y_i - \alpha - \beta X_i)^2$，最大化它就是最小化残差平方和，得 $\hat{\alpha}, \hat{\beta}$ 即 OLS。$g(x)$ 怎么样都不影响——所以我们做回归时从不去建模 X 的分布。下面 2025 Q5(b) 把这套用到指数误差的乘法模型上。

Remark (假设可以放宽多远).

定理里"(Y_i, X_i) iid、X_i 有共同边际 g"只是为了把整份似然干净地写成上面那个连加的分离形式。真正驱动结论的其实更弱：只要给定 $\{X_i\}$ 时各 Y_i 独立、条件密度 $f(y \mid x_i; \theta)$ 设定正确，且 $\{X_i\}$ 的（联合）分布不含 θ，那么条件对数似然 $\sum_i \log f(Y_i \mid X_i; \theta)$ 关于 θ 的 argmax 就不受 X 的规律影响——此时 X_i 允许非同分布、甚至相依，估计量不变。换言之"共同 g、iid"是为陈述方便，可以退到"X 的律不带 θ 信息"这一条真正的前提。

5.8.1 Worked example：2025 Q5(b) ——指数乘法模型的条件 MLE

原题 (Comprehensive Exam 2025, Pinkse, Q5(b); cf. Q5(a))

Let $\{u_i\}$ be i.i.d. standard exponential (density $e^{-u}\mathbb{1}\{u \geq 0\}$). Suppose $y_i = \theta_0 \, x_i \, u_i$ where x_i has bounded support on the positive halfline and $\theta_0 > 0$. Determine the (conditional) maximum likelihood estimator $\widehat{\theta}$ of θ_0 if the u_i's are independent of the x_i's.

Solution.

Step 1：写出条件密度。给定 x_i（且 $u_i \perp x_i$），$y_i = (\theta_0 x_i)\, u_i$ 是"正常数 $\theta_0 x_i$ 乘标准指数"，即指数分布、均值 $\theta_0 x_i$（率 $1/(\theta_0 x_i)$）。用变量替换 $u = y/(\theta x)$、$du/dy = 1/(\theta x)$：

$$f(y_i \mid x_i; \theta) = \frac{1}{\theta x_i} \exp\Big(-\frac{y_i}{\theta x_i}\Big), \qquad y_i > 0.$$

Step 2：条件对数似然。 x_i 的边际分布不含 θ（且 $u \perp x$），由 Thm 5.15 只需最

大化

$$\ell_n(\theta)=\sum_{i=1}^{n}\Big[-\log\theta-\log x_i-\frac{y_i}{\theta x_i}\Big]=-n\log\theta-\sum_{i=1}^{n}\log x_i-\frac{1}{\theta}\sum_{i=1}^{n}\frac{y_i}{x_i}.$$

Step 3：FOC（这里是内点解，可以求导）。

$$\ell_n'(\theta)=-\frac{n}{\theta}+\frac{1}{\theta^2}\sum_{i=1}^{n}\frac{y_i}{x_i}\overset{!}{=}0\implies\boxed{\widehat{\theta}=\frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{x_i}.}$$

二阶导 $\ell_n''(\theta)=\frac{n}{\theta^2}-\frac{2}{\theta^3}\sum\frac{y_i}{x_i}$，在 $\widehat{\theta}$ 处 $=\frac{n}{\widehat{\theta}^2}-\frac{2n}{\widehat{\theta}^2}=-\frac{n}{\widehat{\theta}^2}<0$，确为极大。

合理性自查。 $\mathbb{E}(y_i/x_i\mid x_i)=\mathbb{E}(\theta_0 u_i)=\theta_0$（标准指数均值 1），故 $\widehat{\theta}=\overline{(y/x)}_n$ 无偏且由 WLLN 相合：$\widehat{\theta}\xrightarrow{p}\theta_0$。又 $\mathrm{Var}(y_i/x_i\mid x_i)=\theta_0^2\,\mathrm{Var}(u_i)=\theta_0^2$（标准指数方差 1，即 Q5(a) 的二阶累积量），故 $\sqrt{n}(\widehat{\theta}-\theta_0)\xrightarrow{d}N(0,\theta_0^2)$。■

Remark (考点提炼 + 与 Q5(a) 的联系).

> 关键转化：$y=\theta x\cdot u$ 中 u 标准指数 ⇒ 给定 x 时 y 是均值 θx 的指数，写出其密度即可。Q5(a) 先让你算标准指数的累积量 $\kappa_k=(k-1)!$（来自 CGF $K(t)=-\log(1-t)$）——$\kappa_1=1$（均值）、$\kappa_2=1$（方差）正好用于上面的自查。本题虽叫"条件 MLE"，但因 θ 不卡支撑边界，FOC 照常有效，最终 $\widehat{\theta}=\frac{1}{n}\sum y_i/x_i$ 就是 y_i/x_i 的样本均值。

5.9 自测题 Self-Test

Example (Self-Test 1：泊松 MLE 与渐近分布).

$Y_i\overset{i.i.d.}{\sim}\mathrm{Poisson}(\lambda),\lambda>0$，pmf $f(y;\lambda)=e^{-\lambda}\lambda^y/y!$。求 $\hat{\lambda}$ 及 $\sqrt{n}(\hat{\lambda}-\lambda)$ 的渐近分布。

Solution.

$\ell_n(\lambda)=\sum_i(-\lambda+Y_i\log\lambda-\log Y_i!)$，score $\ell_n'=-n+\frac{1}{\lambda}\sum Y_i=0\Rightarrow\hat{\lambda}=\bar{Y}_n$。信息：$\ell_n''=-\frac{1}{\lambda^2}\sum Y_i$，$I(\lambda)=-\mathbb{E}\big(\partial_\lambda^2\log f\big)=\mathbb{E}(Y)/\lambda^2=\lambda/\lambda^2=1/\lambda$。故 $\sqrt{n}(\hat{\lambda}-\lambda)\xrightarrow{d}N(0,\lambda)$。自查：$\mathrm{Var}(Y_i)=\lambda$，直接 CLT 同样给 $N(0,\lambda)$，对上了。

Example (Self-Test 2：$U[0,\theta]$ 的 MLE（边界）).

$X_i\overset{i.i.d.}{\sim}U[0,\theta]$，$\theta>0$。求 MLE，并说明为何 FOC 失效。

Solution.

$f(x;\theta)=\frac{1}{\theta}\mathbb{1}\{0\le x\le\theta\}$，$L_n(\theta)=\theta^{-n}\mathbb{1}\{\theta\ge\max_i X_i\}$。在 $\theta\ge\max_i X_i$ 上 θ^{-n} 递减，故 $\widehat{\theta}=\max_i X_i=X_{(n)}$。FOC 失效：$\partial_\theta\log L=-n/\theta<0$ 永不为零，因为支撑端点依赖 θ。（与 2018 Q4(a) 同理，$|X_i|$ 换成 X_i。）$\widehat{\theta}$ 有偏：$\mathbb{E}\big(X_{(n)}\big)=\frac{n}{n+1}\theta<\theta$。

Example (Self-Test 3：不变性陈述 + 一句证明).

陈述 MLE 不变性，并用一句话说明一般 g（不必单射）情形的证明骨架。

Solution.

陈述：若 $\widehat{\theta}$ 是 θ 的 MLE，$\tau = g(\theta)$，则 $g(\widehat{\theta})$ 是 τ 的 MLE。骨架：定义诱导（profile）似然 $L^*(\tau) = \sup_{\{\theta: g(\theta)=\tau\}} L(\theta)$，则 $\sup_\tau L^*(\tau) = \sup_\theta L(\theta) = L(\widehat{\theta})$，在 $\hat{\tau} = g(\widehat{\theta})$ 处取得，故 $g(\widehat{\theta})$ 最大化 L^*。（即 2017 Q6。）

Example (Self-Test 4：信息正交 / parameter orthogonality).

$X_i \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$。若 μ_0 已知，估 σ^2 的渐近方差会比 μ_0 未知时更小吗?

Solution.

不会。$I(\mu, \sigma^2)$ 块对角（见 5.5.1），σ^2 方向的渐近方差为 I^{-1} 的 (σ^2, σ^2) 元 $= 2\sigma^4$，与 μ 已知否无关。这是 2024 Q6(d) 的对偶版——参数正交，知道一个不改善另一个。

Example (Self-Test 5：相合性靠 MSE).

设 $\widehat{\theta}_n$ 满足 $\mathbb{E}\left(\widehat{\theta}_n\right) = \theta + \frac{c}{n}$（$c$ 常数）且 $\mathrm{Var}\left(\widehat{\theta}_n\right) = \frac{v}{n}$。证 $\widehat{\theta}_n \xrightarrow{p} \theta$。

Solution.

$\mathrm{Bias} = c/n \to 0$，$\mathrm{Var} = v/n \to 0$，故 $\mathrm{MSE} = \mathrm{Var} + \mathrm{Bias}^2 = \frac{v}{n} + \frac{c^2}{n^2} \to 0$。由 Prop 5.5（MSE$\to 0$ + Markov）得 $\widehat{\theta}_n \xrightarrow{p} \theta$。这复刻了“$L^2$ 收敛 $\Rightarrow$ 依概率收敛”的标准论证。

Chapter 6　经典假设检验（Classical Hypothesis Testing）

导读 · Roadmap（先读这里，不含公式）

假设检验回答的是一个判决问题：手里有一份数据，要在两个互斥的命题（零假设 H_0 与备择假设 H_1）之间做出取舍。本章把这套判决形式化——检验就是一条把数据映到「拒绝 / 不拒绝」的规则（等价于指定一个拒绝域），它会犯两类错误：弃真（第一类）与存伪（第二类）。我们先把 size / level、功效函数、p-value 这些度量讲清，再进入核心：在「简单零 vs 简单备」这个最干净的情形下，*Neyman–Pearson* 引理给出唯一的最优检验——似然比检验；把它推广到复合备择，靠的是单调似然比（MLR）与 *Karlin–Rubin* 定理，得到一侧检验的一致最优性（UMP）。当 UMP 不存在时（多维均值、双侧备择），退而求其次的统一武器是似然比 *LR*、*Wald*、*score* 三检验，三者在零假设下都收敛到 χ^2_r。

资格考在这一章几乎**年年**设题，且偏好两类招牌题：（1）Bernoulli / Binomial 上把 NP 引理、simple-vs-simple UMP、再用 MLR 升级到 one-sided 复合 UMP 一路做到底（2017、2019）；（2）「五检验」问题——$X_i \sim N(\mu_i, 1)$ 独立，检验全体 $\mu_i = 0$，给五个拒绝域定临界值、算 $n = 2$ 的功效、并证不存在 UMP 检验（2018、2025）。本章对这两条主线给足完整、可在考场复现的解答。

6.1　检验的基本构件：拒绝域、两类错误、size 与功效

[优先级 ●●● 高 High]　资格考足迹：2018, 2023, 2025

数据 $X = (X_1, \dots, X_n)$ 的分布由参数 $\theta \in \Theta$ 刻画。我们把参数空间劈成两块 $\Theta = \Theta_0 \sqcup \Theta_1$，要在

$$H_0 : \theta \in \Theta_0 \qquad \text{vs.} \qquad H_1 : \theta \in \Theta_1$$

之间做判决。若 Θ_0（或 Θ_1）是单点，称该假设简单（simple）；否则称复合（composite）。一个检验就是一条规则，等价于在样本空间里指定一个拒绝域 Φ（也叫临界域 critical region）：观测落入 Φ 就拒绝 H_0，否则不拒绝。Φ 的补集 Φ^c 叫接受域（acceptance region）。

Definition 6.1: 两类错误、size、level（Type I/II error, size, level）

For a test with critical region Φ:

- *Type I error* (弃真): rejecting H_0 when it is true; its probability is $\mathrm{P}_\theta(X \in \Phi)$ for $\theta \in \Theta_0$.
- *Type II error* (存伪): failing to reject H_0 when H_1 is true; probability $\mathrm{P}_\theta(X \notin \Phi) = 1 - \mathrm{P}_\theta(X \in \Phi)$ for $\theta \in \Theta_1$.
- The *size* of the test is the worst-case Type I error rate,

$$\alpha^* \;=\; \sup_{\theta \in \Theta_0} \mathrm{P}_\theta(X \in \Phi)\,.$$

- The test has *level* α if $\alpha^* \le \alpha$ (a level is any upper bound on the size).

size 和 level 的区别考场上务必分清：*size* 是真实的最大弃真概率（一个由检验唯一确定的数），*level* 是你声称控制住的上界（0.05、0.10 这种名义值）。一个 size = 0.03 的检验同时是 level-0.05 也是 level-0.10 的检验。复合零假设下 size 取 sup，这个 sup 常在 Θ_0 的边界取得（见 Karlin–Rubin），这是 2017 Q2(e) 的全部要点。

Definition 6.2: 功效函数 Power function

The *power function* of a test with critical region Φ is

$$K(\theta) \;=\; \mathrm{P}_\theta(X \in \Phi) \;=\; \mathrm{P}_\theta(\text{reject } H_0)\,, \qquad \theta \in \Theta.$$

The size is $\sup_{\theta \in \Theta_0} K(\theta)$; the probability of Type II error at $\theta \in \Theta_1$ is $1 - K(\theta)$; and *power* at $\theta \in \Theta_1$ means $K(\theta)$ itself.

功效函数 $K(\theta)$ 是同一个对象在 Θ_0 上和 Θ_1 上的两副面孔：在 Θ_0 上它是「错误拒绝」的概率（越小越好，其上确界就是 size）；在 Θ_1 上它是「正确拒绝」的概率（越大越好，这才叫功效）。理想检验是 Θ_0 上压到 $\le \alpha$、Θ_1 上尽量顶到 1。整章的优化目标——UMP——就是「在 size $\le \alpha$ 的约束下，Θ_1 上处处把 $K(\theta)$ 顶到最高」。

Definition 6.3: p-value

[REPRODUCE —memorize this proof]

Suppose tests are indexed by level $\alpha \in (0, 1)$, with rejection region Φ_α nested ($\Phi_\alpha \subseteq \Phi_{\alpha'}$ for $\alpha \le \alpha'$). The *p-value* at observed data x is

$$p(x) \;=\; \inf\{\alpha : x \in \Phi_\alpha\},$$

the smallest level at which H_0 would be rejected given the data. Equivalently, for a test that rejects when a statistic $T(x)$ is large, $p(x) \;=$

$\sup_{\theta\in\Theta_0} \mathrm{P}_\theta(T(X) \geq T(x))$.

Proof for Theorem

只需证「T 大就拒绝」这类检验的两种写法一致。把零假设下 T 的尾函数（survival function）记为

$$G(t) := \sup_{\theta\in\Theta_0} \mathrm{P}_\theta(T(X) \geq t),$$

它关于 t 非增。视 T 在 H_0 下的分布为连续（本讲义处理临界值的一贯做法），则 G 连续且严格递减，于是 level-α 检验取临界值 $c_\alpha = G^{-1}(\alpha)$——即唯一使 $\sup_{\theta\in\Theta_0} \mathrm{P}_\theta(T(X) \geq c_\alpha) = \alpha$ 成立的值——故 $\Phi_\alpha = \{x : T(x) \geq c_\alpha\}$。因 $\alpha \mapsto c_\alpha = G^{-1}(\alpha)$ 递减，α 变大 c_α 变小、拒绝域变大，这正是嵌套性 $\Phi_\alpha \subseteq \Phi_{\alpha'}$ $(\alpha \leq \alpha')$ 的来源。对固定的观测 x，把严格递减的 G 作用到不等式两边（方向翻转）：

$$x \in \Phi_\alpha \iff T(x) \geq c_\alpha = G^{-1}(\alpha) \iff G(T(x)) \leq \alpha.$$

所以「会拒绝 x 的那些 level」恰构成区间 $\{\alpha : x \in \Phi_\alpha\} = [G(T(x)), 1)$，取下确界即得

$$p(x) = \inf\{\alpha : x \in \Phi_\alpha\} = G(T(x)) = \sup_{\theta\in\Theta_0} \mathrm{P}_\theta(T(X) \geq T(x)).$$

（T 有原子的离散情形：改为「T 严格大于临界值才拒绝」或做随机化，末式仍成立，只是此时 p 在 H_0 下随机优于均匀分布，即 $\mathrm{P}_\theta(p \leq \alpha) \leq \alpha$，不再恰为 $\mathrm{Unif}(0,1)$。）

p-value 是「让你恰好被拒绝的那个 level」——它把「观测到的证据有多极端」压缩成一个 $[0,1]$ 的数。报告 p-value 比报告「在 5% 下拒绝」信息更全，因为读者可自选 α：$p \leq \alpha$ 就拒绝。一个常见误区：p-value 不是「H_0 为真的概率」，而是「若 H_0 为真，看到这么极端或更极端数据的概率」。

6.2 Neyman–Pearson 引理：simple vs simple 的最优检验

[优先级 ●●● 高 High] 资格考足迹：2017, 2019

最干净的情形是简单零 *vs* 简单备：$H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$。这里 Θ_0, Θ_1 都是单点，size 就是 $K(\theta_0)$，功效就是 $K(\theta_1)$。我们要在 $K(\theta_0) = \alpha$ 的约束下把 $K(\theta_1)$ 顶到最大。NP 引理说：答案唯一，就是似然比检验。

Definition 6.4: 似然函数 Likelihood

Let $X = (X_1, \ldots, X_n) \overset{i.i.d.}{\sim} f(x;\theta)$ w.r.t. a dominating measure μ. The *likelihood* is

$$L(\theta; x) = L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i;\theta).$$

Theorem 6.5: Neyman–Pearson 引理 (Neyman–Pearson Lemma)

[REPRODUCE —memorize this proof]

Consider $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$. Among all tests of level α, the *most powerful* (MP) test rejects H_0 on the critical region

$$\Phi \;=\; \Big\{x :\; \frac{L(\theta_0; x)}{L(\theta_1; x)} < k\Big\} \quad\Longleftrightarrow\quad \Big\{x :\; \frac{L(\theta_1; x)}{L(\theta_0; x)} > 1/k\Big\},$$

where $k > 0$ is chosen so that the size equals α, i.e. $\mathrm{P}_{\theta_0}(X \in \Phi) = \alpha$.

- **(存在 Existence & Sufficiency)** A test of this form with $\mathrm{P}_{\theta_0}(X \in \Phi) = \alpha$ exists (continuous case) and is most powerful.
- **(必要 Necessity)** Conversely, any MP level-α test must (up to the boundary $\{L(\theta_0)/L(\theta_1) = k\}$, which has probability 0 in the continuous case) coincide with a likelihood-ratio region of this form.

直觉：似然比 $L(\theta_0; x)/L(\theta_1; x)$ 小，意味着「数据在 θ_1 下比在 θ_0 下更像」，那当然应该拒绝 H_0。NP 引理把这条朴素直觉提升成一条最优性定理：在所有 size $\le \alpha$ 的检验里，没有谁的功效能超过这个似然比检验。注意 Andrews 讲义把拒绝域写成 $L(\theta_0)/L(\theta_1) < k$（小比值拒绝），与 Casella–Berger 写成 $L(\theta_1)/L(\theta_0) > k'$（大比值拒绝）等价，只是 k 的方向相反，考场上随便用哪一种、把 k 的符号说清即可。

Proof for Theorem

[REPRODUCE —memorize this proof]

设 A 是任一 size $\alpha_A \le \alpha$ 的检验的拒绝域，Φ 是上述似然比检验的拒绝域、size 恰为 α。要证 Φ 至少和 A 一样有力，即

$$\mathrm{P}_{\theta_1}(X \in \Phi) - \mathrm{P}_{\theta_1}(X \in A) \;\ge\; 0.$$

用 $d\mathrm{P} = L\, d\mu$，把两个概率拆到公共与互补的部分上：

$$\mathrm{P}_{\theta_1}(X \in \Phi) - \mathrm{P}_{\theta_1}(X \in A) = \int_{\Phi \cap A^c} L(\theta_1; x)\, d\mu - \int_{\Phi^c \cap A} L(\theta_1; x)\, d\mu,$$

因为公共部分 $\Phi \cap A$ 在两边相消。现在用 Φ 的定义比较 $L(\theta_1)$ 与 $L(\theta_0)$：

- 在 Φ 上有 $L(\theta_0; x) < k\, L(\theta_1; x)$，即 $L(\theta_1; x) > \frac{1}{k} L(\theta_0; x)$；故第一项 $\int_{\Phi \cap A^c} L(\theta_1)\, d\mu \;\ge\; \frac{1}{k} \int_{\Phi \cap A^c} L(\theta_0)\, d\mu$。
- 在 Φ^c 上有 $L(\theta_0; x) \ge k\, L(\theta_1; x)$，即 $L(\theta_1; x) \le \frac{1}{k} L(\theta_0; x)$；故第二项 $\int_{\Phi^c \cap A} L(\theta_1)\, d\mu \;\le\; \frac{1}{k} \int_{\Phi^c \cap A} L(\theta_0)\, d\mu$。

两式合并：

$$\begin{aligned}\mathrm{P}_{\theta_1}(X\in\Phi)-\mathrm{P}_{\theta_1}(X\in A) &\geq \frac{1}{k}\Big[\int_{\Phi\cap A^c} L(\theta_0)\,d\mu-\int_{\Phi^c\cap A} L(\theta_0)\,d\mu\Big]\\ &=\frac{1}{k}\big[\mathrm{P}_{\theta_0}(X\in\Phi)-\mathrm{P}_{\theta_0}(X\in A)\big].\end{aligned}$$

最后括号里 $=\alpha-\alpha_A\geq 0$（再次用公共部分相消、把 $\Phi\cap A^c$ 与 $\Phi^c\cap A$ 拼回 Φ 与 A 的 size）。由 $k>0$ 得整体 ≥ 0。■

Remark (考场答题模板).

NP 引理题的标准流程：(1)写出似然 $L(\theta;x)=\prod f(x_i;\theta)$；(2)写似然比 $L(\theta_0)/L(\theta_1)$ 并化简成一个充分统计量 $T(x)$ 的单调函数；（3）把「$L(\theta_0)/L(\theta_1)<k$」翻译成「$T(x)>c$」或「$T(x)<c$」——**注意不等号方向取决于 θ_1 与 θ_0 的大小关系**，常需分情形；（4）由 T 在 θ_0 下的分布定 c 使 size $=\alpha$。许多题（指数族）化简后 $T=\sum x_i$，临界值用其零分布分位数。

6.2.1 Worked example：2017 Part-501 Q2（Bernoulli，NP → UMP → Karlin–Rubin）

原题 (Comprehensive Exam 2017, Part 501, Q2, 25 pts)

Suppose $X_1,\ldots,X_n$ are i.i.d. Bern(p) random variables.

(a) State the lemma by Neyman and Pearson.

(b) What is the likelihood function in this case?

(c) Find a UMP test for $H_0:p=p_0$ versus $H_1:p=p_1$, where $p_0\neq p_1$.

(d) Show that the test from part (c) is UMP for $H_0:p=p_0$ versus $H_1:p>p_0$.

(e) Is the test from part (c) UMP for $H_0:p\leq p_0$ versus $H_1:p>p_0$? Why, why not?

Solution.

(a) 见 Theorem 6.5（陈述拒绝域 $L(\theta_0)/L(\theta_1)<k$、$k$ 由 size $=\alpha$ 定、且该检验在 simple-vs-simple 下最有力）。

(b) 记 $T=\sum_{i=1}^n x_i$。由 $f(x;p)=p^x(1-p)^{1-x}$，

$$L(p;x)=\prod_{i=1}^n p^{x_i}(1-p)^{1-x_i}=p^T(1-p)^{n-T}.$$

(c) simple vs simple，用 NP 引理。似然比

$$\frac{L(p_1;x)}{L(p_0;x)} = \Big(\frac{p_1}{p_0}\Big)^T \Big(\frac{1-p_1}{1-p_0}\Big)^{n-T} = \Big(\frac{1-p_1}{1-p_0}\Big)^n \Big(\frac{p_1(1-p_0)}{p_0(1-p_1)}\Big)^T.$$

关键观察：括号里的「比值的比值」$\dfrac{p_1(1-p_0)}{p_0(1-p_1)}$ 当 $p_1 > p_0$ 时 > 1，故似然比是 T 的严格增函数；当 $p_1 < p_0$ 时 < 1，是 T 的严格减函数。于是

$$\text{若} p_1 > p_0:\ \text{Reject } H_0 \iff T > c; \qquad \text{若} p_1 < p_0:\ \text{Reject } H_0 \iff T < c,$$

其中 c 由 $\mathrm{P}_{p_0}(T > c) = \alpha$（或 $\mathrm{P}_{p_0}(T < c) = \alpha$）确定，$T \sim \mathrm{Bin}(n, p_0)$ 在零下。按题目说明，临界值当作连续情形处理（不必管离散跳跃 / 随机化）。这个似然比检验就是 simple-vs-simple 的 MP（即 UMP）检验。

(d) 把备择换成复合的 $H_1 : p > p_0$。关键在 (c) 中拒绝域 $\{T > c\}$ 及其临界值 c（由 $\mathrm{P}_{p_0}(T > c) = \alpha$ 定）**完全不依赖具体的** p_1——只要 $p_1 > p_0$，最优拒绝域都是同一个 $\{T > c\}$。因此对复合备择 $H_1 : p > p_0$ 里的每一个 $p_1 > p_0$，这同一个检验都同时是对该 p_1 的 MP 检验。「对每一个备择参数都最有力的同一个检验」正是 UMP 的定义，故它是 $H_0 : p = p_0$ vs. $H_1 : p > p_0$ 的 UMP 检验。（背后是 Bernoulli 族关于 $T = \sum x_i$ 有 MLR，见下节 Karlin–Rubin。）

(e) 是，仍是 UMP。把零假设也扩成复合的 $H_0 : p \le p_0$。检验不变（拒绝 $T > c$，c 仍由 $\mathrm{P}_{p_0}(T > c) = \alpha$ 定）。需证两点：(i) 它是 level-α 的，即 size $\le \alpha$；(ii) 它仍 UMP。功效函数

$$K(p) = \mathrm{P}_p(T > c)$$

关于 p 单调不减（$T \sim \mathrm{Bin}(n,p)$ 随机增大于 p，这正是 MLR 的推论，见 Lemma 6.8）。因此

$$\sup_{p \le p_0} K(p) = K(p_0) = \alpha,$$

即 size 在边界 $p = p_0$ 取得、恰等于 α，故 size $= \alpha \le \alpha$，level 合格。又因为它在更小的零假设 $\{p = p_0\}$ vs. $\{p > p_0\}$ 上已是 UMP（由 (d)），而扩大零假设只会缩小「可竞争对手」的集合（任何对 $\{p \le p_0\}$ 是 level-α 的检验，自动对 $\{p = p_0\}$ 也是 level-α），故在更大的 H_0 下它依然在所有 level-α 检验中功效最大。结论：对 $H_0 : p \le p_0$ vs. $H_1 : p > p_0$，仍是 UMP。

6.3 单调似然比 MLR 与 Karlin–Rubin 定理

[优先级 ●●● 高 High] 资格考足迹：2017, 2019, 2023

上一节里「同一个 $\{T > c\}$ 对所有 $p_1 > p_0$ 都最优」并非巧合，而是因为 Bernoulli 族有单调似然比。这条结构性质是把 NP 引理从 simple-vs-simple 一举推到一侧复合假设的桥梁。

Definition 6.6: 单调似然比 Monotone Likelihood Ratio (MLR)

A family $\{f(x;\theta):\theta\in\Theta\subseteq\mathbb{R}\}$ has *monotone likelihood ratio* in a statistic $T(x)$ if for every $\theta_1>\theta_0$ the ratio

$$\frac{L(\theta_1;x)}{L(\theta_0;x)}$$

is a nondecreasing function of $T(x)$ (on the set where it is defined).

Fact 6.7: 一参数指数族都有 MLR

若 $f(x;\theta)=h(x)\exp\{\eta(\theta)T(x)-B(\theta)\}$ 是一参数指数族且自然参数 $\eta(\theta)$ 关于 θ 严格增，则 $\dfrac{L(\theta_1;x)}{L(\theta_0;x)}=\exp\{[\eta(\theta_1)-\eta(\theta_0)]\,T(x)+\text{const}\}$ 是 $T(x)$ 的严格增函数，故该族关于 T 有 MLR。Bernoulli、Binomial、Poisson、$N(\mu,1)$（关于 $\bar{X}$）、指数分布等都属此类。

Lemma 6.8: MLR ⇒ 功效函数单调

[STRUCTURE —know the steps, fudge details]

若该族关于 T 有 MLR，则对任何形如 $\{T>c\}$ 的拒绝域，功效函数 $K(\theta)=\mathrm{P}_\theta(T>c)$ 关于 θ 单调不减。

Proof for Lemma

取 $\theta_1>\theta_0$。$\{T>c\}$ 是一个仅依赖 T 的事件。由 NP 引理的方向性结论（似然比检验最有力）施于 simple θ_0 vs. simple θ_1：在 MLR 下，似然比 $L(\theta_1)/L(\theta_0)$ 是 T 的增函数，故「似然比 $>k$」恰好等价于「$T>c$」某个 c。这意味着 $\{T>c\}$ 就是自身 *size* $=K(\theta_0)$ 的 MP 检验（连续情形边界 $\{T=c\}$ 概率为 0，无需随机化）。把它和「平凡检验：以常概率 $K(\theta_0)$ 随机拒绝、不看数据」相比——后者功效恒为 $K(\theta_0)$、size 也是 $K(\theta_0)$。MP 性给出 $K(\theta_1)\geq K(\theta_0)$。故 K 不减。■

Theorem 6.9: Karlin–Rubin 定理

[STRUCTURE —know the steps, fudge details]

设 $\{f(x;\theta):\theta\in\Theta\subseteq\mathbb{R}\}$ 关于统计量 $T(x)$ 有 MLR。考虑一侧假设

$$H_0:\theta\leq\theta_0\qquad\text{vs.}\qquad H_1:\theta>\theta_0.$$

则「Reject H_0 iff $T(x)>c$」（其中 c 由 $\mathrm{P}_{\theta_0}(T>c)=\alpha$ 确定）是 level-α 的 *UMP* 检验。对称地，$H_0:\theta\geq\theta_0$ vs. $H_1:\theta<\theta_0$ 的 UMP 检验是「Reject iff $T(x)<c$」。

Proof for Theorem

用 Lemma 6.8 的单调性，对拒绝域 $\Phi^* = \{T > c\}$、c 由 $\mathrm{P}_{\theta_0}(T > c) = \alpha$ 定：

$$\sup_{\theta \le \theta_0} K(\theta) = K(\theta_0) = \alpha,$$

故 Φ^* 是 level-α，且 size 在边界 θ_0 取得。再证 UMP：取备择里任一 $\theta_1 > \theta_0$。任何对复合零 $\{\theta \le \theta_0\}$ 是 level-α 的检验，特别对 simple θ_0 也 level-α；而由 NP 引理，对 simple θ_0 vs. simple θ_1，$\Phi^* = \{T > c\}$（其形状不随 θ_1 变）是 MP 的。于是 Φ^* 在每个 $\theta_1 > \theta_0$ 上都不被任何 level-α 对手超过，即 UMP。■

Remark (为什么 one-sided 有 UMP、two-sided 没有).

MLR 的威力在于「最优拒绝域 $\{T > c\}$ 的形状不随备择参数改变」——这正是 UMP 存在的充要直觉。一旦备择是双侧 $H_1 : \theta \neq \theta_0$，朝右的最优域是 $\{T > c\}$、朝左的最优域是 $\{T < c'\}$，两者方向冲突、没有单一检验能同时最优，UMP 就**不存在**（见第 6.5 节与 2025 Q3）。这条「形状是否随备择而变」的判则，正是 2023 Guggenberger Q1(b) 要的答案。

6.3.1 Worked example：2019 Comprehensive Q5（单次 Binomial 观测）

原题 (Comprehensive Exam 2019, Guggenberger, Q5, 12 pts)

Suppose $X \sim \mathrm{Bin}(n, p)$ with pmf $f(x; p) = \binom{n}{x} p^x (1-p)^{n-x}$, $x = 0, \ldots, n$, $p \in [0, 1]$. One observes a *single* random variable X.

(a) What is the MLE of p?

(b) What is the asymptotic distribution of the ML estimator of p as $n \to \infty$?

(c) State the Neyman–Pearson Lemma.

(d) Find the level-α UMP test of $H_0 : p = p_0$ versus $H_1 : p = p_1$ for $p_1 > p_0$. Simplify the form of the test as much as possible. (Treat the critical value as if the distribution were continuous.)

(e) Find the level-α UMP test of $H_0 : p \le p_0$ versus $H_1 : p > p_0$.

Solution.

(a) $\log f = \text{const} + X \log p + (n - X) \log(1 - p)$；一阶条件 $\frac{X}{p} - \frac{n-X}{1-p} = 0$ 给出 $\hat{p} = X/n$。

(b) Fisher 信息 $I(p) = 1/[p(1-p)]$，故 $\sqrt{n}(\hat{p} - p) \xrightarrow{d} N\big(0, p(1-p)\big)$。（等价地由 CLT：$X/n$ 是 n 个 Bern(p) 的均值。）

(c) 见 Theorem 6.5。

(d) simple vs simple，$p_1 > p_0$。似然比

$$\frac{L(p_1;X)}{L(p_0;X)} = \Big(\frac{p_1}{p_0}\Big)^X \Big(\frac{1-p_1}{1-p_0}\Big)^{n-X} = \Big(\frac{1-p_1}{1-p_0}\Big)^n \Big(\underbrace{\frac{p_1(1-p_0)}{p_0(1-p_1)}}_{>1 \text{ since } p_1>p_0} \Big)^X,$$

是 X 的严格增函数。故 NP 引理给出：**Reject H_0 iff $X > c$**，c 由 $\mathrm{P}_{p_0}(X > c) = \alpha$ 定（连续化处理）。

(e) 由 (d)，最优拒绝域 $\{X > c\}$ 及其 c 不依赖 p_1，故对整个 $H_1 : p > p_0$ 一致最优。Binomial 族关于 X 有 MLR（$\frac{p_1(1-p_0)}{p_0(1-p_1)} > 1$），由 Karlin–Rubin（Theorem 6.9）：同一检验「Reject iff $X > c$」对 $H_0 : p \le p_0$ vs. $H_1 : p > p_0$ 是 UMP。size 在边界取得：$\sup_{p \le p_0} \mathrm{P}_p(X > c) = \mathrm{P}_{p_0}(X > c) = \alpha$（功效 $K(p) = \mathrm{P}_p(X > c)$ 关于 p 不减，Lemma 6.8）。

6.4 UMP 不存在时的统一武器：LR / Wald / Score 三检验

[优先级 ●●● 高 High] *资格考足迹：2022*

当备择是双侧、或参数是多维、或有讨厌参数（nuisance parameters）时，UMP 往往*不存在*。这时实务上用「三大经典检验」——似然比（LR）、Wald、score（又名 Lagrange multiplier / Rao）。它们不保证最优，但有「合理的好功效」，且在零假设下都收敛到同一个 χ^2 分布，临界值取卡方分位数即可。设

$$H_0 : h(\theta) = 0_r \qquad \text{vs.} \qquad H_1 : h(\theta) \neq 0_r,$$

$h : \mathbb{R}^d \to \mathbb{R}^r$ 连续可微、$r \le d$，雅可比 $H(\theta) = \partial h(\theta)/\partial \theta'$ 行满秩 r。$\widehat{\theta}$ 为（无约束）MLE，$\widetilde{\theta}$ 为约束 MLE（在 $h(\theta) = 0$ 下最大化），$I(\theta)$ 为 Fisher 信息，$S_n(\theta) = \sum_i \partial \log f(X_i;\theta)/\partial\theta$ 为 score。

Definition 6.10: LR 统计量与广义似然比检验

The *(generalized) likelihood ratio* is

$$\Lambda(x) = \frac{\max_{\theta:\, h(\theta)=0} L(\theta;x)}{\max_{\theta} L(\theta;x)} = \frac{L(\widetilde{\theta};x)}{L(\widehat{\theta};x)} \in (0,1],$$

and the LR test *rejects H_0 when Λ is small*, i.e. when $-2\log\Lambda$ is large.

Theorem 6.11: $-2\log\Lambda$ 的零极限分布（LR 检验）

[STRUCTURE —know the steps, fudge details]

Under the ML regularity conditions and H_0,

$$-2\log\Lambda(X_1,\dots,X_n) \;=\; 2\big[\log L(\widehat{\theta}) - \log L(\widetilde{\theta})\big] \;\xrightarrow{d}\; \chi^2_r \quad \text{as } n \to \infty,$$

where r is the number of restrictions. The level-α test rejects H_0 if $-2\log\Lambda > \chi^2_{r,1-\alpha}$.

Proof for Theorem

[REPRODUCE —memorize this proof]

下面完整证 $h(\theta)=\theta-\theta_0$（即 $H_0:\theta=\theta_0$，$r=d$）这一标志性特例——这正是考场要求能独立复现的部分。一般 h（$r<d$，需处理讨厌方向 / 约束 MLE $\widetilde{\theta}$、并用秩为 r 的投影把二次型降成 χ^2_r 而非 χ^2_d）是标准的引用结果，思路同此特例、只是记号更重，本处不展开（见 van der Vaart, *Asymptotic Statistics*, Ch. 16）。在 H_0 下真值为 θ_0。写

$$-2\log\Lambda = 2\sum_{i=1}^{n}\big[\log f(X_i;\widehat{\theta})-\log f(X_i;\theta_0)\big].$$

对 $\sum_i \log f(X_i;\theta_0)$ 在 $\widehat{\theta}$ 处做二阶 Taylor 展开（$\theta^\dagger$ 在 θ_0 与 $\widehat{\theta}$ 之间）：

$$\begin{aligned}\sum_i \log f(X_i;\theta_0) &= \sum_i \log f(X_i;\widehat{\theta}) + \underbrace{S_n(\widehat{\theta})'}_{=0}(\theta_0-\widehat{\theta}) \\ &\quad + \tfrac{1}{2}(\widehat{\theta}-\theta_0)'\Big[\sum_i \tfrac{\partial^2}{\partial\theta\partial\theta'}\log f(X_i;\theta^\dagger)\Big](\widehat{\theta}-\theta_0).\end{aligned}$$

其中一阶项 $S_n(\widehat{\theta})=0$ 是无约束 MLE 的一阶条件（以趋于 1 的概率成立，因 $\theta_0\in\operatorname{int}(\Theta)$）。代回：

$$\begin{aligned}-2\log\Lambda &= (\widehat{\theta}-\theta_0)'\Big[-\sum_i \tfrac{\partial^2}{\partial\theta\partial\theta'}\log f(X_i;\theta^\dagger)\Big](\widehat{\theta}-\theta_0) \\ &= \sqrt{n}(\widehat{\theta}-\theta_0)'\Big[-\tfrac{1}{n}\sum_i \tfrac{\partial^2}{\partial\theta\partial\theta'}\log f(X_i;\theta^\dagger)\Big]\sqrt{n}(\widehat{\theta}-\theta_0).\end{aligned}$$

两个收敛：(i) 由 WLLN（$\theta^\dagger \xrightarrow{p}\theta_0$）$-\frac{1}{n}\sum_i \partial^2\log f(X_i;\theta^\dagger)/\partial\theta\partial\theta' \xrightarrow{p} I(\theta_0)$；(ii) MLE 渐近正态 $\sqrt{n}(\widehat{\theta}-\theta_0)\xrightarrow{d} Z\sim N(0,I^{-1}(\theta_0))$。由连续映射定理，

$$-2\log\Lambda \xrightarrow{d} Z'I(\theta_0)Z.$$

令 $S=I^{1/2}(\theta_0)Z\sim N(0,\,I^{1/2}I^{-1}I^{1/2})=N(0,I_r)$，则 $Z'I(\theta_0)Z=S'S\sim\chi^2_r$。■

Theorem 6.12: Wald 检验

[STRUCTURE —know the steps, fudge details]

设估计量满足 $\sqrt{n}(\widehat{\theta}-\theta)\xrightarrow{d}N(0,V(\theta))$，$\hat{V}_n\xrightarrow{p}V(\theta)$。*Wald* 统计量

$$W_n = n\,h(\widehat{\theta})'\big[H(\widehat{\theta})\hat{V}_n H(\widehat{\theta})'\big]^{-1}h(\widehat{\theta}).$$

在 H_0 下 $W_n\xrightarrow{d}\chi^2_r$（由 delta 方法：$\sqrt{n}\,h(\widehat{\theta})\xrightarrow{d}N(0,HVH')$，再 CMT）。Reject H_0 if $W_n>\chi^2_{r,1-\alpha}$。

Definition 6.13: Score（LM / Rao）检验

Score statistic（在约束 MLE $\widetilde{\theta}$ 处评估，只需算约束估计）：

$$LM_n = \tfrac{1}{n}\,S_n(\widetilde{\theta})'\,I(\widetilde{\theta})^{-1}\,S_n(\widetilde{\theta}),$$

where $S_n(\theta)=\sum_{i=1}^n \partial\log f(X_i;\theta)/\partial\theta$. Under H_0, $LM_n\xrightarrow{d}\chi^2_r$; reject if $LM_n>\chi^2_{r,1-\alpha}$.

三检验是同一渐近内容的三副面孔，区别在「需要算哪个估计量」：**Wald** 只要无约束 MLE $\widehat{\theta}$（看 $h(\widehat{\theta})$ 离 0 多远）；**score/LM** 只要约束 MLE $\widetilde{\theta}$（看在约束点上 score 离 0 多远——若 H_0 真，约束点处梯度应近 0）；**LR** 两个都要算（比两处的对数似然高度差）。三者在零下都 $\to\chi^2_r$，局部备择下渐近功效也相同（一阶等价）；有限样本谁好谁坏依模型而定。考场记法：「Wald 算无约束、LM 算约束、LR 都算」。注意 Wald 不具参数化不变性（对 h 的非线性重写敏感），LR 具不变性——这是常考的辨析点。

Remark（*p*-value（三检验通用））.

设观测到统计量值 t_n（$=-2\log\Lambda$、W_n 或 LM_n），则 p-value $=\mathrm{P}(Q\ge t_n)=1-F_{\chi^2_r}(t_n)$，$Q\sim\chi^2_r$。$p\le\alpha\iff t_n>\chi^2_{r,1-\alpha}\iff$ 拒绝。

6.5 招牌题：「五检验」问题与 UMP 不存在性

[优先级 ●●● 高 High] 资格考足迹：2018, 2025

这是资格考反复出现的招牌题（2018 Econ-501 Q2、2025 Part-510 Q3）。它把本章每件武器都用上：定临界值（size）、算功效函数、并证明 UMP 不存在。下面给 2025 的完整解（2018 是同题，仅 n 一般化、答案结构一致）。

原题 (Comprehensive Exam 2025, Part 510, Q3)

Suppose $X_i\sim N(\mu_i,1)$, $i=1,\dots,n$ are independent. Consider the following five tests of $H_0:\mu_1=\cdots=\mu_n=0$ versus H_1 : "not H_0" with *acceptance*

regions

$$A_1 = \{(x_1,\dots,x_n) : \textstyle\sum_{i=1}^n x_i \le k_1\}, \qquad A_2 = \{(x_1,\dots,x_n) : |\textstyle\sum_{i=1}^n x_i| \le k_2\},$$
$$A_3 = \{(x_1,\dots,x_n) : x_i^2 \le k_3 \text{ for all } i\}, \quad A_4 = \{(x_1,\dots,x_n) : x_1^2 \le k_4\},$$
$$A_5 = \{(x_1,\dots,x_n) : \textstyle\sum_{i=1}^n x_i^2 \le k_5\},$$

where k_j are constants (possibly depending on n) chosen so the tests have size α for some $\alpha \in (0,1)$.

(a) Determine $k_j,\ j = 1,\dots,5$.

(b) Let $n = 2$ from now on. Determine the power functions of the first and the fourth test.

(c) Show that none of the five tests is (weakly) more powerful than all the other tests considered here uniformly over the alternative space $\{(\mu_1,\mu_2) : \mu_1 < \infty, \mu_2 < \infty\}$.

Remark (读题：拒绝域 = 接受域的补).

题给的是接受域 A_j，拒绝域是 $\Phi_j = A_j^c$。例如 $\Phi_1 = \{\sum x_i > k_1\}$（一侧），$\Phi_2 = \{|\sum x_i| > k_2\}$（双侧），$\Phi_3 = \{\max_i x_i^2 > k_3\}$，$\Phi_4 = \{x_1^2 > k_4\}$，$\Phi_5 = \{\sum x_i^2 > k_5\}$。size $= \mathrm{P}_{H_0}(\text{reject}) = \mathrm{P}_{H_0}\left(X \in A_j^c\right) = 1 - \mathrm{P}_{H_0}(X \in A_j) = \alpha$，即在 H_0 下要求 $\mathrm{P}_{H_0}(X \in A_j) = 1-\alpha$。

Solution.

(a) 临界值。在 H_0 下所有 $\mu_i = 0$，故 $X_i \overset{i.i.d.}{\sim} N(0,1)$，$X_i^2 \overset{i.i.d.}{\sim} \chi_1^2$。

- A_1（**一侧**）：$\sum_i X_i \sim N(0,n)$。$\mathrm{P}(\sum X_i \le k_1) = 1-\alpha \Rightarrow \dfrac{k_1}{\sqrt{n}} = z_{1-\alpha}$，即

$$\boxed{k_1 = \sqrt{n}\, z_{1-\alpha}.}$$

- A_2（**双侧**）：$\mathrm{P}(|\sum X_i| \le k_2) = 1-\alpha \Rightarrow \dfrac{k_2}{\sqrt{n}} = z_{1-\alpha/2}$，即 $\boxed{k_2 = \sqrt{n}\, z_{1-\alpha/2}.}$
- A_3（**全部 $\le k_3$，Šidák**）：因 X_i 独立，

$$\mathrm{P}\big(X_i^2 \le k_3\ \forall i\big) = \big[\mathrm{P}\big(\chi_1^2 \le k_3\big)\big]^n = 1-\alpha \ \Rightarrow\ \mathrm{P}\big(\chi_1^2 \le k_3\big) = (1-\alpha)^{1/n},$$

即 $\boxed{k_3 = \chi^2_{1,\,(1-\alpha)^{1/n}}}$（$\chi_1^2$ 的 $(1-\alpha)^{1/n}$ 分位数）；等价地 $k_3 = \big(z_{(1+(1-\alpha)^{1/n})/2}\big)^2$。

- A_4（**单坐标**）：$X_1^2 \sim \chi_1^2$，$\mathrm{P}\big(X_1^2 \le k_4\big) = 1-\alpha \Rightarrow \boxed{k_4 = \chi^2_{1,1-\alpha}.}$
- A_5（**omnibus**）：$\sum_i X_i^2 \sim \chi_n^2$，$\mathrm{P}\big(\sum X_i^2 \le k_5\big) = 1-\alpha \Rightarrow \boxed{k_5 = \chi^2_{n,1-\alpha}.}$

(b) 功效函数（$n=2$）。现在 $X_1 \sim N(\mu_1,1)$、$X_2 \sim N(\mu_2,1)$ 独立。

Test 1（$\Phi_1=\{X_1+X_2>k_1\}$，$k_1=\sqrt{2}\,z_{1-\alpha}$）：$X_1+X_2 \sim N(\mu_1+\mu_2,2)$，故

$$K_1(\mu_1,\mu_2)=\mathrm{P}(X_1+X_2>k_1)=1-\Phi\Big(\frac{k_1-(\mu_1+\mu_2)}{\sqrt{2}}\Big)=1-\Phi\Big(z_{1-\alpha}-\frac{\mu_1+\mu_2}{\sqrt{2}}\Big).$$

它只依赖 $\mu_1+\mu_2$，关于 $\mu_1+\mu_2$ 单调增；当 $\mu_1+\mu_2<0$ 时功效低于 α（biased，向负方向几乎无力）。

Test 4（$\Phi_4=\{X_1^2>k_4\}$，$k_4=\chi^2_{1,1-\alpha}=z^2_{1-\alpha/2}$）：$X_1^2$ 服从非中心 $\chi^2_1(\mu_1^2)$。直接用 $\Phi_4=\{|X_1|>\sqrt{k_4}\}$：

$$\begin{aligned}K_4(\mu_1,\mu_2)&=\mathrm{P}\Big(|X_1|>\sqrt{k_4}\Big)=1-\mathrm{P}\Big(-\sqrt{k_4}\le X_1\le\sqrt{k_4}\Big)\\&=1-\Phi(\sqrt{k_4}-\mu_1)+\Phi(-\sqrt{k_4}-\mu_1),\end{aligned}$$

其中 $\sqrt{k_4}=z_{1-\alpha/2}$。它只依赖 μ_1，完全忽略 μ_2（当 $\mu_1=0$ 时 $K_4\equiv\alpha$，无论 μ_2 多大）。

(c) 五个里无人占优。要证「五个检验里没有任何一个在整个备择空间上（弱）优于其余全部」。策略：**对每个检验，找出一个备择点 (μ_1,μ_2)，使另一个检验在该点功效严格更高。**这就直接回答了题目（五个里无人占优）。

注意：单凭「五个里无人占优」还不能推出「不存在 UMP」——假想中的 UMP 未必是这五个之一，它可以落在这族之外、却同时弱优于这五个，与「五个里无人占优」并不矛盾。真正证明「不存在 UMP」的，是下方关于最优拒绝域形状随备择方向改变的论证（见**结论**段）。

- **Test 1（一侧和）在 $\mu_1+\mu_2<0$ 方向失效。**取 $\mu_1=\mu_2=-t$（$t>0$ 大）。则 $K_1=1-\Phi(z_{1-\alpha}+\frac{2t}{\sqrt{2}})\to 0$，远低于 α。而 Test 5（$\Phi_5=\{X_1^2+X_2^2>k_5\}$，非中心 $\chi^2_2(\mu_1^2+\mu_2^2)$）的功效随 $\mu_1^2+\mu_2^2=2t^2\to\infty$ 而 $\to 1$。故 Test 5 在此点严格优于 Test 1。

- **Test 4（只看坐标 1）忽略 μ_2。**取 $\mu_1=0$，$\mu_2=t$（t 大）。则 $K_4\equiv\alpha$（与 μ_2 无关）。而 Test 5 的功效随 $\mu_2^2=t^2\to\infty$ 而 $\to 1$，严格更高。（Test 2、Test 3 在此点也都 $>\alpha$。）

- **Test 2 / Test 5（全向 omnibus）在「单一方向」上被 Test 1 击败。**取 $\mu_1=\mu_2=t>0$（沿 $(+1,+1)$ 方向），记 $S=\dfrac{X_1+X_2}{\sqrt{2}}\sim N(\sqrt{2}\,t,1)$。三检验此时的拒绝域都可用 S 表出：$\Phi_1=\{S>z_{1-\alpha}\}$、$\Phi_2=\{|S|>z_{1-\alpha/2}\}$，而 $\Phi_5=\{X_1^2+X_2^2>k_5\}$ 是圆盘外部。功效

  $$K_1=1-\Phi\big(z_{1-\alpha}-\sqrt{2}\,t\big),\qquad K_2=1-\Phi\big(z_{1-\alpha/2}-\sqrt{2}\,t\big)+\Phi\big(-z_{1-\alpha/2}-\sqrt{2}\,t\big).$$

  关键（*NP* 引理）：在简单对 $H_0:(0,0)$ vs. $H_1:(t,t)$ 上，似然比 $\dfrac{L(t,t)}{L(0,0)}=\exp\{t(X_1+X_2)-t^2\}$ 关于 X_1+X_2（即 S）单调增，故 Test 1 的拒绝域 $\{S>z_{1-\alpha}\}$ 恰好是这对假设的似然比检验——由 NP 引理它在 size $=\alpha$ 下最有力（MP），

且（似然比连续、无原子）在 a.e. 意义下**唯一**。于是任何 level-α 检验在 (t,t) 处功效都 $\le K_1$，凡拒绝域与 $\{S > z_{1-\alpha}\}$ 在正概率集上不同者**严格** $< K_1$。这一击同时打倒 Test 2 与 Test 5：

- **打败 *Test 2***：Φ_2 与 $\{S > z_{1-\alpha}\}$ 在正概率集上不同（$\{z_{1-\alpha} < S < z_{1-\alpha/2}\}$ 只在 Test 1 里、$\{S < -z_{1-\alpha/2}\}$ 只在 Test 2 里，两者概率均 >0），故

$$K_2 < K_1 \qquad (\forall\, t > 0).$$

  这对**所有** $t>0$ 都成立，不必「t 大」。**易错点（务必避开）**：切勿这样证——「$K_2 = \big[1-\Phi(z_{1-\alpha/2}-\sqrt{2}\,t)\big] + \Phi(-z_{1-\alpha/2}-\sqrt{2}\,t)$，丢掉非负的下尾项得 $K_2 < 1-\Phi(z_{1-\alpha/2}-\sqrt{2}\,t)$」。丢掉一个**非负**项，得到的是 K_2 的**下界**而非上界：$K_2 \ge 1-\Phi(z_{1-\alpha/2}-\sqrt{2}\,t) =: A$，写成 $<$ 方向就反了。后半段「因 $z_{1-\alpha/2} > z_{1-\alpha}$ 故 $A < K_1$」本身对，但你手上只有 $K_2 \ge A$ 与 $A < K_1$——这两条根本推不出 K_2 与 K_1 的大小。$K_2 < K_1$ 结论是真的，却只能靠上面的 NP 论证，不能靠「丢尾项」。

- **打败 *Test 5***：Φ_5 是圆盘外部、并非半平面，与 $\{S > z_{1-\alpha}\}$ 亦在正概率集上不同，故同理 $K_5 < K_1$（$\forall\, t>0$）。直觉：Test 5 把同样的信号能量摊进二维非中心 $\chi^2_2(2t^2)$、抬高了临界值 k_5，而 Test 1 把 α 全押在正确的一维方向上。数值印证：$\alpha = 0.05$、$t = 1.5$（即 $\mu_1+\mu_2 = 3$）时 $K_1 = 1-\Phi(1.645-2.121) = \Phi(0.476) \approx 0.683$，而 $K_5 = \mathrm{P}\big(\chi^2_2(4.5) > 5.99\big) \approx 0.460 < K_1$。

- **Test 3（max 型）与 Test 5（sum 型）互不占优——给出两个反向点。**

  - **信号集中在单坐标**（$\mu_1 = c$ 大、$\mu_2 = 0$）：Test 3 的拒绝靠 $\max_i X_i^2$，能直接抓住放大的 X_1^2；Test 5 把 X_2^2（纯噪声）也加进统计量、抬高了临界值 k_5，反而稀释信号。具体地，取 $\alpha = 0.05$，$\mu_1 = 3$，$\mu_2 = 0$：Test 3 临界 $k_3 = \chi^2_{1,(0.95)^{1/2}} \approx 5.00$，其拒绝靠 $\max_i X_i^2$，故 $K_3 = 1-\mathrm{P}\big(\chi^2_1(9) \le 5.00\big)\ \mathrm{P}\big(\chi^2_1 \le 5.00\big) \approx 0.783$（坐标 1 服从非中心 $\chi^2_1(9)$、坐标 2 是纯噪声 χ^2_1）；Test 5 临界 $k_5 = \chi^2_{2,0.95} \approx 5.99$，$K_5 = \mathrm{P}\big(\chi^2_2(9) > 5.99\big) \approx 0.771 < K_3$。故此点 Test 3 严格优于 Test 5。

  - **信号均摊到两坐标**（$\mu_1 = \mu_2 = c$ 中等）：能量分散，Test 5「看坐标平方和」抓得住、Test 3「只看最大坐标」抓不全。取 $\alpha = 0.05$，$\mu_1 = \mu_2 = 2$（$\mu_1^2+\mu_2^2 = 8$）：$K_5 = \mathrm{P}\big(\chi^2_2(8) > 5.99\big) \approx 0.718$，而 Test 3 每坐标非中心参数仅 $\mu_i^2 = 4$、$K_3 = 1-\big[\mathrm{P}\big(\chi^2_1(4) \le 5.00\big)\big]^2 \approx 1-(0.594)^2 \approx 0.648 < K_5$。故此点 Test 5 严格优于 Test 3。

  这两个反向点合起来证明 Test 3 与 Test 5 互不占优。

结论：每个检验都「在某个方向最好、在另一个方向被超越」。**根本原因**：备择 H_1 是**多维双侧的**（(μ_1,μ_2) 可朝任意方向偏离原点），而 NP 引理给出的最优拒绝域**形状随备择方向改变**——朝 $(+1,+1)$ 最优的是 $\{X_1+X_2 > c\}$，朝单坐标最优的是 $\{|X_1| > c\}$，朝任意方向各不相同，没有单一拒绝域能同时朝所有方向最优。故**不存在 UMP 检验**。■

Remark (考点提炼：什么时候 UMP 不存在).

背一句判则：**当最优拒绝域的形状随备择参数（方向）改变时，UMP 不存在**。三种典型情形：(i) 双侧备择 $H_1:\theta\neq\theta_0$（左右两侧最优域冲突）；(ii) 多维 / 多参数均值的偏离（本题，各方向冲突）；(iii) 有讨厌参数。补救方向（2023 Guggenberger Q1(b) 的答案）：限制在无偏检验里求 UMPU（uniformly most powerful unbiased）；或用不变性求 UMP invariant；或换准则——加权平均功效（weighted-average power）、最不利分布（least-favorable）、minimax。其中 omnibus 的 $\sum X_i^2$ 检验（Test 5）正是「旋转不变」准则下的 UMP invariant 检验。

6.6 自测题 Self-Test

Example (Self-Test 1：NP 引理（正态均值，方差已知）).

$X_1,\dots,X_n\overset{i.i.d.}{\sim}N(\mu,\sigma^2)$，$\sigma^2$ 已知。检验 $H_0:\mu=\mu_0$ vs. $H_1:\mu=\mu_1$（$\mu_1>\mu_0$）。求 level-α MP 检验，并说明它对 $H_1:\mu>\mu_0$ 是否 UMP。

Solution.

似然比 $\dfrac{L(\mu_0)}{L(\mu_1)}=\exp\left\{\frac{1}{2\sigma^2}[2n\bar{x}(\mu_1-\mu_0)-n(\mu_1^2-\mu_0^2)]\right\}^{-1}$ 化简后是 $\bar{x}$ 的单调函数。因 $\mu_1>\mu_0$，「似然比小」$\iff$ $\bar{x}$ 大，故 **Reject H_0 iff** $\bar{X}>\mu_0+\frac{\sigma}{\sqrt{n}}z_{1-\alpha}$，等价 $T_n=\frac{\sqrt{n}(\bar{X}-\mu_0)}{\sigma}>z_{1-\alpha}$。临界值与 μ_1 无关，且 $N(\mu,\sigma^2)$ 关于 $\bar{X}$ 有 MLR，故由 Karlin–Rubin，该检验对 $H_1:\mu>\mu_0$（乃至 $H_0:\mu\leq\mu_0$）是 UMP。功效 $K(\mu)=1-\Phi\big(z_{1-\alpha}-\frac{\sqrt{n}(\mu-\mu_0)}{\sigma}\big)$。

Example (Self-Test 2：size 在边界取得).

设功效函数 $K(\theta)$ 关于 θ 单调不减（如 MLR 下的一侧检验）。证明对复合零 $H_0:\theta\leq\theta_0$，检验的 size 等于 $K(\theta_0)$。

Solution.

size $=\sup_{\theta\leq\theta_0}K(\theta)$。$K$ 不减 $\Rightarrow$ 对一切 $\theta\leq\theta_0$ 有 $K(\theta)\leq K(\theta_0)$，故 $\sup_{\theta\leq\theta_0}K(\theta)=K(\theta_0)$，在边界 θ_0 取得。这正是「复合零下 size 在边界达到」的一般机制，也是 2017 Q2(e)、2019 Q5(e) 把 one-sided 检验从 simple-null 升级到 composite-null 的关键步骤。

Example (Self-Test 3：双侧无 UMP).

$X_1,\dots,X_n\overset{i.i.d.}{\sim}N(\mu,1)$。说明对 $H_0:\mu=0$ vs. $H_1:\mu\neq0$ 不存在 UMP 检验。

Solution.

对 simple 备择 $\mu=\mu_1>0$，NP 给出的 MP 拒绝域是 $\{\bar{X}>c_+\}$；对 $\mu=\mu_1<0$，MP 拒绝域是 $\{\bar{X}<c_-\}$。两者形状相反、互相在对方方向上功效不足（$\{\bar{X}>c_+\}$ 在 $\mu<0$ 时功效 $<\alpha$）。没有单一检验能同时对所有 $\mu\neq0$ 最优，故 UMP 不存在。实务上改用双侧 $|\bar{X}|>c$（这是 LR 检验、也是 UMP 无偏检验）。

Example (Self-Test 4：LR / Wald / Score 谁算哪个估计量).

检验 $H_0 : \theta = \theta_0$（$d = r = 1$）。简述三检验各需要哪个估计量、统计量形式与零极限分布。

Solution.

LR：需无约束 MLE $\widehat{\theta}$ 与约束 MLE $\widetilde{\theta} = \theta_0$；$-2\log\Lambda = 2[\log L(\widehat{\theta}) - \log L(\theta_0)] \xrightarrow{d} \chi^2_1$。**Wald**：只需 $\widehat{\theta}$；$W_n = n(\widehat{\theta} - \theta_0)^2/\hat{V}_n \xrightarrow{d} \chi^2_1$（$\hat{V}_n \xrightarrow{p} I^{-1}$）。**Score/LM**：只需 $\widetilde{\theta} = \theta_0$（无需拟合无约束模型）；$LM_n = \frac{1}{n} S_n(\theta_0)^2 I(\theta_0)^{-1} \xrightarrow{d} \chi^2_1$。三者局部备择下渐近等价；Wald 不具参数化不变性，LR 具不变性。

Example (Self-Test 5：五检验之 Test 5 的功效（$n = 2$）).

在「五检验」设定下（$n = 2$），写出 omnibus 检验 Test 5（$\Phi_5 = \{X_1^2 + X_2^2 > \chi^2_{2,1-\alpha}\}$）的功效函数，并指出它依赖 (μ_1, μ_2) 的哪个量。

Solution.

$X_1^2 + X_2^2$ 服从非中心卡方 $\chi^2_2(\lambda)$，非中心参数 $\lambda = \mu_1^2 + \mu_2^2$。故

$$K_5(\mu_1, \mu_2) = \mathrm{P}\big(\chi^2_2(\mu_1^2 + \mu_2^2) > \chi^2_{2,1-\alpha}\big)\,,$$

只依赖 $\mu_1^2 + \mu_2^2$（到原点的欧氏距离平方），关于该量单调增、且旋转不变。这解释了为何 Test 5 是「全方向」检验：它对所有方向一视同仁，因而在任一具体方向上被对准该方向的定向检验（如 Test 1）超越——再次印证无 UMP。它本身是旋转不变准则下的 UMP invariant 检验。