前两节课的主要内容是为了让我们熟悉统计里所运用的数学知识，在一个虽然思维上很快就能得到渐近分布的过程中，严谨的使用定理 (例如 Continuous Mapping Theorem, CMT) 来得到最后的渐近分布 (eg. convergence of sample variance)。

1 各种收敛

首先我们需要知道一些收敛概念以及其定义，首先是统计里最关心的收敛，依概率收敛 (convergence in probablity)，主要关心的是远离期望的尾部概率(这点在高维和集中不等式中被大量讨论 (concentratation inequality))：

Def 1.1 依概率收敛：

$$\lim_{n\to\infty} P\left(|X_n - X| < \epsilon\right) = 1 \Leftrightarrow X_n \xrightarrow{p} X$$

由此可以引申出两个概念，第一是依概率有界，

Def 1.2 *A sequence of random variables X_n is said to be bounded in probability if, for any $\epsilon > 0$, there exists a constant k such that $P\left(|X_n| > k\right) \leq \epsilon$ for all n*

请注意这里课件上有一条注释，任何随机变量（向量）都是依概率有界的，这个在统计上是可以接受的，只要你不去构建那些奇怪的随机变量（可见 *localized conformal prediction*）。

第二个概念是随机小项 (stochastic o(),O())

Def 1.3 *More generally, for a given sequence of random variables R_n,*

$$\begin{aligned} X_n = o_p\left(R_n\right) \quad &\Leftarrow \quad X_n = Y_n R_n \text{ and } Y_n \xrightarrow{p} 0 \\ X_n = O_p\left(R_n\right) \quad &\Leftarrow \quad X_n = Y_n R_n \text{ and } Y_n = O_p(1) \end{aligned}$$

这个就是我们说的以 XX 速率收敛到 0，或者以 XX 速率有界 (at the rate R_n)

接下来介绍一个更强的收敛，几乎处处收敛 (convergence with probablity 1, 以概率 1 收敛，强收敛，几乎处处收敛)，这个收敛因为需要的条件太强了，所以我们一般也不用，事实上想想统计上依概率收敛就够了，甚至我们后面看到一般都是用以 L 阶矩收敛。

Def 1.4 几乎处处收敛 $P\left(\lim_{n\to\infty} X_n = X\right) = 1 \Leftrightarrow X_n \overset{wp1}{\to} X$ *or* $X_n \overset{a.s}{\to} X$

这个收敛实在是太强了，我们一般也不用，所以我们在下面只给出一个经常用来证明他的定理 (等价定义)，以及 Borel-Cantelli 引理：

Theorem 1.1 我们可以证明下面两个定义等价：

$$\lim_{n\to\infty} P\left(|X_m - X| < \epsilon, \forall m \geq n\right) = 1, \forall \epsilon > 0 \Leftrightarrow X_n \overset{a.s}{\to} X$$

这个等价定义事实上集合的上下极限与概率测度的连续性的共同结果，有时候会让证明变得简单。

Lemma 1.2 (Borel-Cantelli 引理)

$$\textit{If, for every } \epsilon > 0, \sum_{n=1}^{\infty} P\left(|X_n - X| > \epsilon\right) < \infty, \textit{ then } X_n \overset{wp1}{\to} X$$

这个引理在证明几乎处处收敛中也常用到

接下来我们介绍了以 r 阶矩收敛 (convergence in rth mean)，

Def 1.5 以 r 阶矩收敛

$$\lim_{n\to\infty} E\left|X_n - X\right|^r = 0 \Leftrightarrow X_n \overset{rth}{\to} X$$

很容易我们得到高阶矩收敛能够推出低阶矩收敛

Theorem 1.3 $X_n \overset{rth}{\to} X \Rightarrow X_n \overset{sth}{\to} X, 0 < s < r$

实际上以 r 阶矩收敛可以推出依概率收敛(高维里讨论的 Markov，chebyshev 不等式)，这在统计上证明依概率收敛几乎都是这么证明的，这是因为以 r 阶矩收敛是可以计算的

最后我们介绍依分布收敛 (convergence in law)，其定义是这样的

Def 1.6 如果在分布函数的每个连续点上我们有

$$\lim_{n\to\infty} F_{X_n}(t) = F_X(t)$$

我们则称 $X_n \overset{d}{\to} X$

这个收敛可以和特征函数很好的结合，但是我没学过复变函数，所以对于这个收敛很多东西都不懂，值得一提的是这个收敛是最弱的收敛，可以被依分布收敛推出，他们之前的具体关系如下：

$$\xi_n \xrightarrow{\text{a.e.}} \xi \text{ 或} \xi_n \xrightarrow{L_r} \xi$$
$$\Downarrow$$
$$\xi_n \xrightarrow{\mathbb{P}} \xi$$
$$\Downarrow$$
$$\xi_n \xrightarrow{w} \xi$$

2 随机变量的函数的收敛性

在这一章中，我们介绍那些我们通常认为成立的事情背后具体依托的是哪些定理，首先是最重要的定理，连续映射定理，Continuous Mapping Theorem(CMT)。

Theorem 2.1 (Continuous Mapping Theorem,CMT) *Let* $\mathbf{X}_1, \mathbf{X}_2, \ldots$ *and* $\mathbf{X}$ *be random p-vectors defined on a probability space, and let* $g(\cdot)$ *be a vector-valued (including real-valued) continuous function defined on* $\mathbb{R}^p$. *If* $\mathbf{X}_n$ *converges to* $\mathbf{X}$ *in probability, almost surely, or in law, then* $g(\mathbf{X}_n)$ *converges to* $g(\mathbf{X})$ *in probability, almost surely, or in law, respectively.*

这个定理简单来说就是连续函数在收敛性上具有传递性(当然连续可以是几乎处处连续)，

$$g \in C^0 \quad \text{and} \quad X_n \xrightarrow{\text{except rth}} X \Rightarrow g(\mathbf{X_n}) \xrightarrow{\text{except rth}} g(\mathbf{X})$$

Remark 2.1.1 之所以要 *except rth* 是显然的，可以考虑：$|X_n - X| \xrightarrow{rth} 0, g(x) = x^{\frac{s}{r}}$ *where* $s > r$。如果 *CMT* 对 $\xrightarrow{rth}$ 成立的话，那么低阶矩收敛可以推高阶矩收敛，这显然是错的。

Remark 2.1.2 这里如果在随机向量上运用连续映射定理的话，需要 *joint* 收敛，即需要 $\boldsymbol{X_n} \xrightarrow{joint} \boldsymbol{X}$，但是如果我们不考虑 $\xrightarrow{d}$ 而是只考虑 $\xrightarrow{p}$ *or* $\xrightarrow{w.p.1}$ 的话我们可以只让分量分别收敛。具体可见下面的定理

Theorem 2.2 (Σ and Π of r.v. $\xrightarrow{\boldsymbol{p}}$ or $\xrightarrow{\boldsymbol{a.s.}}$) *If* $X_n \xrightarrow{wp1} X$ *and* $Y_n \xrightarrow{wp1} Y$, *then* $X_n + Y_n \xrightarrow{wp1} X + Y$ *and* $X_n Y_n \xrightarrow{wp1} XY$. *Replacing the wp1 with in probability, the foregoing arguments also hold.*

我们接下来讨论一下为什么 $\xrightarrow{d}$ 时不成立，这是因为 $\xrightarrow{d}$ 实在是一个太弱的条件，如wiki所写，依分布收敛是最宽松的收敛方式之一。这种收敛不要求查看每个 ω，只要求序列的分布趋向于某个极限, 即只要求 $\mathbb{P}(X_n \leqslant a) \to \mathbb{P}(X \leqslant a)$。很容易理解的一点是，$\xrightarrow{p}$ or $\xrightarrow{a.s.}$ 你需要在意每个样本点 ω 被映射成了什么，但是 $\xrightarrow{d}$ 甚至不需要管这些，只需要管分布函数就行。甚至可以这么说，$\xrightarrow{p}$ or $\xrightarrow{a.s.}$ 定义的 X 是一个确定的，但是 $\xrightarrow{d}$ 确定的 X 甚至不是确定的。我们可以举一个例子来说明：

Example 2.1 若 $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{I})$，定义一个随机变量列 $\{\boldsymbol{X}_k\}$ 满足 $\boldsymbol{X}_i = \boldsymbol{X}, i = 1, 2, \cdots$，那么 $\{\boldsymbol{X}_k\}$ 依概率收敛或者几乎处处收敛到 $\boldsymbol{X}$, 但是如果考虑 $\xrightarrow{d}$ 的话，$\{\boldsymbol{X}_k\}$ 收敛到任意一个标准多元正态分布 $\boldsymbol{Y}$

这样我们很容易明白为什么 $\xrightarrow{d}$ 不能分量分别收敛，例如 $X_i = U, Y_i = 1 - U$，我们可以说 $X_i \xrightarrow{d} U, Y_i \xrightarrow{d} U$，但是 $X_i + Y_i = 1$ 而不是 $2U$。请注意这里不能说 Y_i 依概率或者几乎处处收敛到 U

接下来我们讨论一下对于 $\xrightarrow{d}$ 的情况怎么处理，如果有 joint 收敛，那必然是好的，但是如果只有依分布收敛，但是有一个分量依分布收敛到一个常数，我们也是可以处理的，这就是 **Slutsky's Theorem**

Theorem 2.3 (Slutsky's Theorem) *Let $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, where c is a finite constant. Then,*

1. $X_n + Y_n \xrightarrow{d} X + c$;
2. $X_n Y_n \xrightarrow{d} cX$;
3. $X_n / Y_n \xrightarrow{d} X/c$ *if* $c \neq 0$

请注意 $X_n \xrightarrow{d} c \Leftrightarrow X_n \xrightarrow{p} c$，所以上述 $\xrightarrow{p}$ 可以换为 $\xrightarrow{d}$，不过在统计里，依分布收敛到一个常数的证明方式一般还是用尾概率的方式证明依概率收敛。**Slutsky's Theorem** 的主要运用方式就是先用 **CLT** 证明和依分布收敛到某个分布，然后再证明剩下的东西是个小项，即收敛到常数 0，然后加起来

3 处理和的方式

冯龙老师和邹老师都提到说，统计里要得到最后的渐近分布，最主要的就是要找到其中的和 (这里说核可能会显得神秘一点，但是统计里大多都是

和)，只不过可能是加权和？condition 完之后出现和？anyway，处理这些和的方式就是接下来介绍的大数定律与中心极限定理。maybe 有人会说集中不等式，其实证明大数定律的方式多少就和集中不等式比较像了，所以我更愿意把集中不等式看成大数定律成立的速率，即 *non-asympotic view*

在开始之前，我们首先强调一点，作为统计的一门实用？的课程，我们不 care 这个成立条件是不是最弱的，也不关心这个成立条件是不是比那一个弱，我们只关心这个条件好不好验证，是不是容易让人理解，运用范围大不大。(personal view)→ 事实上，满足最弱的条件而不满足一个比较强的条件的分布已经比较难构造了，这在统计关心的现实世界里很难很难见到。

3.1 大数定律

首先我们介绍独立同分布下的大数定律

Theorem 3.1 *Let $X_1, X_2, \ldots,$ be iid random variables having a CDF F.*
(i) The WLLN The existence of constants a_n for which

$$\frac{1}{n}\sum_{i=1}^{n} X_i - a_n \xrightarrow{p} 0$$

holds iff $\lim_{x\to\infty} x[1 - F(x) + F(-x)] = 0$, in which case we may choose $a_n = \int_{-n}^{n} x dF(x)$.
(ii) The SLLN The existence of a constant c for which

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{wp1} c$$

holds iff $E[X_1]$ is finite and equals c.

当然我们在这里给出的是最弱的条件，例如 $\lim_{x\to\infty} x[1-F(x)+F(-x)] = 0$ 这个条件，统计真的 care 吗？统计到底要怎样验证这个条件？统计到底为什么要假设这么弱的条件？现实里假设方差有限貌似在大部分条件下也没有问题，这样我们可以很容易得到 iid 下的 WLLN。至于 SLLN 这个条件，虽然比 WLLN 的条件强，但是一眼就能明白而且好接受，并且是好验证的，(personal view)→ 这才是统计需要的条件，但是我确实也很少见到直接使用这个 WLLN。

接下来我们介绍允许方差和期望不同的大数定律

Theorem 3.2 *Let $X_1, X_2, \ldots$, be random variables with finite expectations.*
(i) The WLLN Let $X_1, X_2, \ldots$, be uncorrelated with means $\mu_1, \mu_2, \ldots$ and variances $\sigma_1^2, \sigma_2^2, \ldots$ If $\lim_{n\to\infty} \frac{1}{n^2} \sum_{i=1}^n \sigma_i^2 = 0$, then

$$\frac{1}{n}\sum_{i=1}^n X_i - \frac{1}{n}\sum_{i=1}^n \mu_i \xrightarrow{p} 0.$$

(ii) The SLLN Let $X_1, X_2, \ldots$, be independent with means $\mu_1, \mu_2, \ldots$ and variances $\sigma_1^2, \sigma_2^2, \ldots$. If $\sum_{i=1}^\infty \sigma_i^2/c_i^2 < \infty$ where c_n ultimately monotone and $c_n \to \infty$, then

$$c_n^{-1} \sum_{i=1}^n (X_i - \mu_i) \xrightarrow{wp1} 0.$$

(iii) The SLLN with common mean Let $X_1, X_2, \ldots$, be independent with common mean μ and variances $\sigma_1^2, \sigma_2^2, \ldots$ If $\sum_{i=1}^\infty \sigma_i^{-2} = \infty$, then

$$\sum_{i=1}^n \frac{X_i}{\sigma_i^2} / \sum_{i=1}^n \sigma_i^{-2} \xrightarrow{wp1} \mu$$

这两个大数律在方差的角度也比较好理解，也能解决实际问题，比如异方差这种问题。(iii) 相当于给了一个 c_n 具体的形式，可以解决异方差下 μ 的 BLUE 估计问题。

Example 3.1 *Suppose $X_i \overset{indep}{\sim} (\mu, \sigma_i^2)$. Then, by simple calculus, the BLUE (best linear unbiased estimate) of μ is $\sum_{i=1}^n \sigma_i^{-2} X_i / \sum_{i=1}^n \sigma_i^{-2}$. Suppose now that the σ_i^2 do not grow at a rate faster than i; i.e., for some constant $K, \sigma_i^2 \le iK$. Then, $\sum_{i=1}^n \sigma_i^{-2}$ clearly diverges as $n \to \infty$, and so by the theorem the BLUE of μ is strongly consistent.*

所以说大多数时候，$\sum X_i/n$ 直接就是收敛就完了，然后通过收敛，加上各种函数变换和加减乘除去弄那些奇奇怪怪的东西就好了。

接下来我们介绍一下依分布收敛的内容 $\xrightarrow{d}$, 首先是一些等价的形式

Theorem 3.3 *Let $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2, \ldots$ random p-vectors.*
(i) (The Portmanteau Theorem) $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ is equivalent to the following condition: $E[g(\mathbf{X}_n)] \to E[g(\mathbf{X})]$ for every bounded continuous function g.
(ii) (Levy-Cramer continuity theorem) Let $\Phi_{\mathbf{X}}, \Phi_{\mathbf{X}_1}, \Phi_{\mathbf{X}_2}, \ldots$ be the character functions of $\mathbf{X}, \mathbf{X_1}, \mathbf{X}_2, \ldots$, respectively. $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ iff $\lim_{n\to\infty} \Phi_{\mathbf{X}_n}(\mathbf{t}) = \Phi_{\mathbf{X}}(\mathbf{t})$ for all $\mathbf{t} \in \mathbb{R}^p$.

(iii) (Cramer-Wold device) $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ *iff* $\mathbf{c}^T\mathbf{X}_n \xrightarrow{d} \mathbf{c}^T\mathbf{X}$ *for every* $\mathbf{c} \in \mathbb{R}^p$.（这个在课件上居然给出了证明，我感觉可能会考证明）

对于咱们统计证明来说，前两个定理大多数时候只需要从左边到右边就行了，一般是你通过其他方式得到了 $X_n \xrightarrow{d} X$，然后再利用这些左边到右边去弄一些东西。不过，这些东西可以提供一些统计上的想法，例如你想检验这个是不是依分布收敛到什么什么（? 不知道有没有这样的问题，但是检验分布相同的问题是不少的，也可以到这里找答案），你完全可以随便找一些函数 g，或者向量 c，然后去检验右边这些东西成不成立，毕竟这些东西是一维的，听说（ii）也有人用，不过我真的是完全不懂特征函数。

接下来是由依分布收敛导出的关于分布函数和概率密度函数的收敛性，考虑这个主要是为了保证分位数 $(F_n^{-1}(\alpha) \to F^{-1}(\alpha))$ 和区间估计 $(\int_a^b g(x)f_n(x)dx \to \int_a^b g(x)f(x)dx = 1-\alpha)$ 的收敛性

Theorem 3.4 *(i)* ***(Prohorov's Theorem)*** *If* $X_n \xrightarrow{d} X$ *for some* X*, then* $X_n = O_p(1)$
(ii) ***(Polya's Theorem)*** *If* $F_{X_n} \Rightarrow F_X$ *and* F_X *is continuous, then as* $n \to \infty$

$$\sup_{-\infty<x<\infty} |F_{X_n} - F_X| \to 0$$

*(iii)****(Scheffe Theorem)*** *Let* f_n *be a sequence of densities of absolutely continuous functions, with* $\lim_n f_n(\mathbf{x}) = f(\mathbf{x})$*, each* $\mathbf{x} \in \mathbb{R}^p$*. If* f *is a density function, then* $\lim_n \int |f_n(\mathbf{x}) - f(\mathbf{x})|\, d\mathbf{x} = 0$.*(这个定理也给出了证明，复习的时候可以看一下)*

下面这个定理说明了函数之间小项的传递性也可以传递到随机变量上

Theorem 3.5 *Let* g *be a function defined on* $\mathbb{R}^p$ *such that* $g(\mathbf{0}) = 0$*. Let* $\mathbf{X}_n$ *be a sequence of random vectors with values on* $\mathbb{R}$ *that converges in probability to zero. Then, for every* $r > 0$*,*
(i) if $g(\mathbf{t}) = o\left(\|\mathbf{t}\|^r\right)$ *as* $t \to 0$*, then* $g\left(\mathbf{X}_n\right) = o_p\left(\|\mathbf{X}_n\|^r\right)$*;*
(ii) if $g(\mathbf{t}) = O\left(\|\mathbf{t}\|^r\right)$ *as* $t \to 0$*, then* $g\left(\mathbf{X}_n\right) = O_p\left(\|\mathbf{X}_n\|^r\right)$.

Proof 可以考虑函数 $f(\mathbf{t}) = g(\mathbf{t})/||\mathbf{t}||^r$,*(i)* 可以直接用 *CMT* 就可以得到结论，*(ii)* 只需要利用 $||t|| \le \delta \Rightarrow |f(\mathbf{t})| \le M$，所以

$$P\left(|f\left(\mathbf{X}_n\right)| > M\right) \le P\left(\|\mathbf{X}_n\| > \delta\right) \to 0$$

3.2 中心极限定理

首先是最经典的中心极限定理，**Lindeberg-Levy 中心极限定理**

Theorem 3.6 (Lindeberg-Levy) *Let X_i be iid with mean μ and finite variance σ^2. Then*

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

喔！iid，方差有限，多么经典的条件，(personal view)→ 要我说，中心极限定理到这里完了就行了，who care 那些稀奇古怪的条件呢？

在这里，我们详细地讨论一个例子，样本的方差的极限分布，我们将把用到的每一个定理都写出来，来说明一个在脑子很快就能进行的过程实际上是有坚实的定理支撑的，当然之后我们还是要更快更好，不过我们至少说明一下统计不是一门纯艺术好吧（bushi

Example 3.2 (Sample variance) *Suppose $X_1,\ldots,X_n$ are iid with mean μ, variance σ^2 and $E(X_1^4)<\infty$. Consider the asymptotic distribution of $S_n^2=\frac{1}{n-1}\sum_{i=1}^n\left(X_i-\bar{X}_n\right)^2$*

首先我们得到分解

$$\sqrt{n}\left(S_n^2-\sigma^2\right)=\sqrt{n}\left(\frac{1}{n-1}\sum_{i=1}^n\left(X_i-\mu\right)^2-\sigma^2\right)-\sqrt{n}\frac{n}{n-1}\left(\bar{X}_n-\mu\right)^2$$

因为第二项是个小项，我们首先来分析这一项，这里实际上还不能 $\bar{X}_n \xrightarrow{a.s.} \mu$ 来得到这一项趋于零,因为前面还有一个 $\sqrt{n}$,这里合适的方式是 $\frac{\left(\sqrt{n}\left(\bar{X}_n-\mu\right)\right)^2}{\sqrt{n}}$，我们需要通过三个步骤来说明他是一个小项。

Step 1 $\sqrt{n}\left(\bar{X}_n-\mu\right)\xrightarrow{d}N(0,\sigma^2)$ *(Lindeberg-Levy 中心极限定理，CLT)*

Step 2 $\left(\sqrt{n}\left(\bar{X}_n-\mu\right)\right)^2\xrightarrow{d}\sigma^2\chi^2$ *(连续映射定理，CMT)*

Step 3 $\frac{\left(\sqrt{n}\left(\bar{X}_n-\mu\right)\right)^2}{\sqrt{n}}\xrightarrow{d}\frac{\sigma^2}{\infty}=0$ *(CMT or Slusky's Theorem)*

这三个步骤之后，我们 *claim*，$\frac{\left(\sqrt{n}\left(\bar{X}_n-\mu\right)\right)^2}{\sqrt{n}}=o(1)$ ***or*** $\frac{\left(\sqrt{n}\left(\bar{X}_n-\mu\right)\right)^2}{\sqrt{n}}\xrightarrow{d}0$

接下来我们分析第一项，第一项直接就是由中心极限定理 *(CLT)*得到

$$\sqrt{n}\left(\frac{1}{n-1}\sum_{i=1}^n\left(X_i-\mu\right)^2-\sigma^2\right)\xrightarrow{d}N(0,\widetilde{\sigma}^2)$$

接下来我们用Slusky's Theorem可以得到两块加起来也是正态分布。

(personal view)→ 其实我感觉处理核的技巧在这里就可以结束了，毕竟在统计里接触中心极限定理一般也就是这样的条件，不过为了完整性，我们下面还是简单摘录一下其他的中心极限定理。

Theorem 3.7 (Multivariate CLT for iid case) *Let* $\mathbf{X}_i$ *be iid random p-vectors with mean* $\boldsymbol{\mu}$ *and and covariance matrix* $\boldsymbol{\Sigma}$*. Then*

$$\sqrt{n}(\overline{\mathbf{X}} - \boldsymbol{\mu}) \xrightarrow{d} N_p(\mathbf{0}, \boldsymbol{\Sigma}).$$

推广到高维的中心极限定理是自然的。为了介绍下面一个定理，我们还得介绍一个这样的概念

Def 3.1 *A function* $g : \mathbb{R} \to \mathbb{R}$ *is called slowly varying at* ∞ *if, for every* $t > 0, \lim_{x\to\infty} g(tx)/g(x) = 1$.

这个东西就是说函数在尾部变化的很慢，其实就是尾部比较薄，毕竟太薄了就没地方给他变了，接下来就是包含这种定义的中心极限定理

Theorem 3.8 *Let* $X_1, X_2, \ldots$ *be iid from a CDF* F *on* $\mathbb{R}$*. Let* $v(x) = \int_{-x}^{x} y^2 dF(y)$*. Then, there exist constants* $\{a_n\}, \{b_n\}$ *such that*

$$\frac{\sum_{i=1}^{n} X_i - a_n}{b_n} \xrightarrow{d} N(0, 1),$$

if and only if $v(x)$ *is slowly varying at* ∞.

当然这种中心极限定理不用要求方差存在，不过他也没给出 a_n 和 b_n 的构造方式，所以用处也不大。接下来这个中心极限定理是一个条件还比较像人话的中心极限定理，**Lindeberg-Feller** 中心极限定理

Theorem 3.9 (Lindeberg-Feller CLT) *Suppose* X_n *is a sequence of independent variables with means* μ_n *and variances* $\sigma_n^2 < \infty$*. Let* $s_n^2 = \sum_{i=1}^{n} \sigma_i^2$*. If for any* $\epsilon > 0$

$$\frac{1}{s_n^2} \sum_{j=1}^{n} \int_{|x-\mu_j|>\epsilon s_n} (x - \mu_j)^2 \, dF_j(x) \to 0,$$

where F_i *is the CDF of* X_i*, then*

$$\frac{\sum_{i=1}^{n} (X_i - \mu_i)}{s_n} \xrightarrow{d} N(0, 1)$$

Remark 3.9.1 这个定理说明了可以使用具有不同的均值 μ_n 和方差 σ_n 的中心极限定理，并且条件还是比较接近人话的，不过这个积分区间还是比较搞，还是比较难验证。

为了进一步让这个条件说人话，可以使用 **Liapounov 中心极限定理**

Theorem 3.10 (Liapounov CLT) *Suppose X_n is a sequence of independent variables with means μ_n and variances $\sigma_n^2 < \infty$. Let $s_n^2 = \sum_{i=1}^n \sigma_i^2$. If for some $\delta > 0$*

$$\frac{1}{s_n^{2+\delta}} \sum_{j=1}^{n} E\left|X_j - \mu_j\right|^{2+\delta} \to 0$$

as $n \to \infty$, then

$$\frac{\sum_{i=1}^n (X_i - \mu_i)}{s_n} \xrightarrow{d} N(0,1)$$

Remark 3.10.1 在课件上有一个点说的是，如果 X_i 是 *uniformly bounded*，那么 ***Liapounov CLT*** 中的条件就被满足了,这个主要是因为 $\mathbb{E}\left(\left|X_i - \mu_i\right|^3\right) \leq C\mathbb{E}\left(\left|X_i - \mu_i\right|^2\right)$，把这个带进去就可以了。
这个倒还提醒了一件事，某种意义上来看，上面那个分子是 n 的速度，下面这个分母有点像 $n^{\frac{3}{2}}$ 的速度

这个定理就比较好了，不仅条件比较好验证，而且所有元素都给出了构造方法。

再接下来我们考虑 double array 和 triangular array 这两种形式的随机变量列，这个还是比较宽泛地拓展了中心极限定理的使用范围的，浅浅收回之前的暴论。

- Doubule array 指的是 $X_{ij}, j \leq i \overset{iid}{\sim} F_i$, $(i \leq j)$ 代表这是一个三角形，就是一行随机变量内部都独立同分布，但每行之间的分布可以不同。
- Triangular array 指的是 $X_{ij} \sim F_{ij}$，就是每个随机变量都独立服从不同的分布

首先是 double array 的中心极限定理

Theorem 3.11 *Let the X_{ij} be distributed as a double array. Then*

$$P\left(\frac{\sqrt{n}\left(\bar{X}_n - \mu_n\right)}{\sigma_n} \leq x\right) \to \Phi(x)$$

as $n \to \infty$ *for any sequence* F_n *with mean* μ_n *and variance* σ_n^2 *for which*

$$E_n \left| X_{n1} - \mu_n \right|^3 / \sigma_n^3 = o(\sqrt{n})$$

Here E_n *denotes the expectation under* F_n.

接下来是 triangular array 的中心极限定理

Theorem 3.12 *Let the* X_{ij} *be distributed as a triangular array and let* $E(X_{ij}) = \mu_{ij}$, $\operatorname{var}(X_{ij}) = \sigma_{ij}^2 < \infty$, *and* $s_n^2 = \sum_{j=1}^n \sigma_{nj}^2$. *Then,*

$$\frac{\sum_{j=1}^n (X_{nj} - \mu_{nj})}{s_n} \xrightarrow{d} N(0,1),$$

provided that

$$\frac{1}{s_n^{2+\delta}} \sum_{j=1}^n E \left| X_{nj} - \mu_{nj} \right|^{2+\delta} \to 0$$

通过两个定理我们可以看到，其实中心极限定理可以跟你是不是同分布没啥关系，只要你的尾概率（maybe）被控制住，就不会出现哪些远离中心值太离谱的数据，用邹老师的话来说，就是不会出现 domainant 的数据点，这样他们远离中心的距离就是会出现正态性，很神奇的一个结论。

当然我们有时候也不会太关注这种条件这么弱的结论，我们统计整个加权就已经很够用了，这就是 **Hajek-Sidak CLT**

Theorem 3.13 *(Hajek-Sidak) Suppose* $X_1, X_2, \ldots$ *are iid random variables with mean* μ *and variance* $\sigma^2 < \infty$. *Let* $c_n = (c_{n1}, c_{n2}, \ldots, c_{nn})$ *be a vector of constants such that*

$$\max_{1 \le i \le n} \frac{c_{ni}^2}{\sum_{j=1}^n c_{nj}^2} \to 0$$

as $n \to \infty$. *Then*

$$\frac{\sum_{i=1}^n c_{ni} (X_i - \mu)}{\sigma \sqrt{\sum_{j=1}^n c_{nj}^2}} \xrightarrow{d} N(0,1)$$

这个定理的中心思想还是一样的，只要不出现 dominant 的数据点，就还是满足中心极限定理。

在接下里介绍**多元 Lindeberg-Feller 中心极限定理**

Theorem 3.14 (Lindeberg-Feller multivariate) *Suppose* $\mathbf{X}_i$ *is a sequence of independent vectors with means* $\boldsymbol{\mu}_i$*, covariances* $\boldsymbol{\Sigma}_i$ *and distribution function* F_i*. Suppose that* $\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Sigma}_i \to \boldsymbol{\Sigma}$ *as* $n \to \infty$*, and that for any* $\epsilon > 0$

$$\frac{1}{n}\sum_{j=1}^{n}\int_{\|\mathbf{x}-\boldsymbol{\mu}_j\|>\epsilon\sqrt{n}} \|\mathbf{x}-\boldsymbol{\mu}_j\|^2 \, dF_j(\mathbf{x}) \to 0$$

then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\mathbf{X}_i - \boldsymbol{\mu}_i) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma})$$

当然这个条件还是比较搞的，我们在写论文的时候一般就不会使用这样的条件，我们一般会给一点简化，一下面这个为例

Example 3.3 ***(multiple regression)*** *In the linear regression problem, we observe a vector* $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ *for a fixed or random matrix* $\mathbf{X}$ *of full rank, and an error vector* ε *with iid components with mean zero and variance* σ^2*. The least squares estimator of* $\boldsymbol{\beta}$ *is* $\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$*. This estimator is unbiased and has covariance matrix* $\sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$*. If the error vector* ε *is normally distributed, then* $\widehat{\boldsymbol{\beta}}$ *is exactly normally distributed. Under reasonable conditions on the design matrix,* $\widehat{\boldsymbol{\beta}}$ *is asymptotically normally distributed for a large range of error distributions.*

如果我们把 p 固定然后让 n 趋于 ∞, 如果我们 *define* $\mathbf{A} := \left(\mathbf{X}^T\mathbf{X}\right)^{-1/2}\mathbf{X}^T$, $\mathbf{a}_{n1}, \cdots, \mathbf{a}_{nn}$ 是 $\mathbf{A}$ 的列向量, 我们将估计量和真值之间的差乘上一个 $\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{\frac{1}{2}}$, 我们可以得到下面这个表达式

$$\left(\mathbf{X}^T\mathbf{X}\right)^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \left(\mathbf{X}^T\mathbf{X}\right)^{-1/2}\mathbf{X}^T\varepsilon = \sum_{i=1}^{n}\mathbf{a}_{ni}\varepsilon_i$$

那么多元的中心极限定理就是要满足下面这个条件

$$\sum_{i=1}^{n}\|\mathbf{a}_{ni}\|^2 E\varepsilon_i^2 I_{\{\|\mathbf{a}_{ni}\|\|\varepsilon_i|>\epsilon\}} \to 0$$

这个示性函数的存在让这个条件的验证变得比较复杂，也不够自然，如果我们注意到了

$$\sum\|\mathbf{a}_{ni}\|^2 = \operatorname{tr}\left(\mathbf{A}\mathbf{A}^T\right) = p$$

我们可以只让后面乘的这一项的最大值趋于零就行，即

$$\max_i E\varepsilon_i^2 I_{\{\|\mathbf{a}_{ni}\|\|\varepsilon_i|>\epsilon\}} \to 0$$

或者我们注意到 $\max_i E\varepsilon_i^2 I_{\{\|\mathbf{a}_{ni}\||\varepsilon_i|>\epsilon\}} \to 0$ 可以被 $\epsilon^{-k}E\left|\varepsilon_i\right|^{k+2}\left\|\mathbf{a}_{ni}\right\|^k$ 控制住，所以我们可以提出另一个条件

$$\sum_{i=1}^{n}\left\|\mathbf{a}_{ni}\right\|^k \to 0; \quad E\left|\varepsilon_1\right|^k < \infty, k > 2$$

Remark 3.14.1 因为 $\sum\left\|\mathbf{a}_{ni}\right\|^2 = \operatorname{tr}\left(\mathbf{A}\mathbf{A}^T\right) = p$，所以整体上 $\left\|\mathbf{a}_{ni}\right\|^2$ 趋势上还是趋于 *0* 的，所以这个 $\sum_{i=1}^{n}\left\|\mathbf{a}_{ni}\right\|^k \to 0$ 也是可以理解的

接下来我们需要处理一些 dependent 的情况，我们可以先介绍一下下面这个例子，照冯龙老师和邹老师的意思，所有 dependent 的情况都是转换成 independent 的情况

Example 3.4 *Suppose $X_1, X_2, \ldots$ is a stationary Gaussian sequence with $E(X_i) = \mu, \operatorname{var}(X_i) = \sigma^2 < \infty$. Then, for each n, $\sqrt{n}\left(\bar{X}_n - \mu\right)$ is normally distributed and so $\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{d} N(0, \tau^2)$, provided $\tau^2 = \lim_{n\to\infty}\operatorname{var}\left(\sqrt{n}\left(\bar{X}_n - \mu\right)\right) < \infty$. But*

$$\operatorname{var}\left(\sqrt{n}\left(\bar{X}_n - \mu\right)\right) = \sigma^2 + \frac{1}{n}\sum_{i\neq j}\operatorname{cov}(X_i, X_j) = \sigma^2 + \frac{2}{n}\sum_{i=1}^{n}(n-i)\gamma_i,$$

where $\gamma_i = \operatorname{cov}(X_1, X_{i+1})$. Therefore, $\tau^2 < \infty$ if and only if $\frac{1}{n}\sum_{i=1}^{n}(n-i)\gamma_i$ has a finite limit, say ρ, in which case $\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{d} N(0, \sigma^2 + \rho)$.

这个例子里虽然没有转换成独立的随机变量，但是要求了 $\frac{1}{n}\sum_{i=1}^{n}(n-i)\gamma_i$ 有极限，这其实要求 γ_i 要以一个比较大的速率趋于 0，至少 $\frac{1}{i}$ 是不行的，这其实说明了，X_i 在以一个比较打的速度趋向无关。

接下来我们可以介绍一个定义以及由这个定义引出来的中心极限定理

Def 3.2 *A stationary sequence $\{X_n\}$ is called m-dependent for a given fixed m if $(X_1, \ldots, X_i)$ and $(X_j, X_{j+1}, \ldots)$ are independent whenever $j - i > m$.*

这个定义是说在 m 步之后，大家没有关系了，这样的话也可以分段来做

Theorem 3.15 (m-dependent sequence) *Let $\{X_i\}$ be a stationary m-dependent sequence. Let $E(X_i) = \mu$ and $\operatorname{var}(X_i) = \sigma^2 < \infty$. Then $\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{d} N(0, \tau^2)$, where $\tau^2 = \sigma^2 + 2\sum_{i=2}^{m+1}\operatorname{cov}(X_1, X_i)$.*

这个定理的证明也很简单，就是 m 个为一段分段，剩下的那些不到 m 的肯定是 negligible，具体可见下面的例子

Example 3.5 *Suppose Z_i are i.i.d. with a finite variance σ^2, and let $X_i = (Z_i + Z_{i+1})/2$. Then, obviously $\sum_{i=1}^n X_i = \frac{Z_1+Z_{n+1}}{2} + \sum_{i=2}^n Z_i$ Then, by Slutsky's theorem, $\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{d} N(0, \sigma^2)$. Notice we write $\sqrt{n}\left(\bar{X}_n - \mu\right)$ into two parts in which one part is dominant and produces the CLT, and the other part is asymptotically negligible. This is essentially the method of proof of the CLT for more general m-dependent sequences.*

3.3 中心极限定理的误差

接下来介绍一个我之前从来没有想过的问题，用中心极限定理的误差，这对我们做分位数的估计和中心极限定理的估计都十分关键，那就是我们需要关心下面这一项的大小，

$$\Delta_n = \sup_x \left| P\left(\frac{\bar{X}_n - E\left(\bar{X}_n\right)}{\sqrt{\operatorname{var}\left(\bar{X}_n\right)}} \leq x \right) - \Phi(x) \right|$$

当然，我们知道在中心极限定理的作用下，这一项是趋于 0 的，但是我们更想要一种 non-asymptotic 的速率，这就是Berry-Esseen 定理

Theorem 3.16 (Berry-Esseen Theorem) *(i) (iid case) Let $X_1, \ldots, X_n$ be iid with $E(X_1) = \mu$, $\operatorname{var}(X_1) = \sigma^2$, and $\beta_3 = E|X_1 - \mu|^3 < \infty$. Then there exists a universal constant C, not depending on n or the distribution of the X_i, such that*

$$\sup_x \left| P\left(\frac{\sqrt{n}\left(\bar{X}_n - \mu\right)}{\sigma} \leq x \right) - \Phi(x) \right| \leq \frac{C\beta_3}{\sigma^3\sqrt{n}}.$$

(ii) (independent but not iid case) Let $X_1, \ldots, X_n$ be independent with $E(X_i) = \mu_i, \operatorname{var}(X_i) = \sigma_i^2$, and $\beta_{3i} = E|X_i - \mu_i|^3 < \infty$. Then there exists a universal constant C^*, not depending on n or the distribution of the X_i, such that

$$\sup_x \left| P\left(\frac{\bar{X}_n - E\left(\bar{X}_n\right)}{\sqrt{\operatorname{var}\left(\bar{X}_n\right)}} \leq x \right) - \Phi(x) \right| \leq \frac{C^* \sum_{i=1}^n \beta_{3i}}{\left(\sum_{i=1}^n \sigma_i^2\right)^{3/2}}$$

Remark 3.16.1 对这个定理做如下解释

- 如果只关心速率的话，可以不管 C，在邹老师给出的例子里，C 的取值是 *0.8*

- 这个速率可以看成两部分，一部分是 $\sqrt{n}$，n 越大，这个东西越小，这很合理，例外一部分你不能单独看 σ^3，不然你会认为这个 σ 越大，这个估计越准确，这和直观不符，一般是 σ 越小，分布越集中，这个东西越好，你可以吧 $\frac{\beta_3}{\sigma^3}$ 看成一个整体，直观上来看，这个越分散应该越大，所以这个界还是分布越集中越好

这个定理说明了了这个误差和方差的和的发散速度有关，如果是独立同分布的情况下，这个速度就至少是 $n^{\frac{3}{2}}$ 了。

但是这个误差控制有个问题，就是这个东西控制的是误差最大的地方，但是有的时候我们并不关心这个误差最大的地方，有的时候我们只关心比较尾部的概率，比如上 0.975 分位数，这里的尾概率本来就比较小了，你给我一个两个比较小的值的差的绝对的 bound 是没啥用的，所以我们需要下面这个 **Berry-Essen Theorem**

Theorem 3.17 (Berry-Esseen Theorem) *Let $X_1, \ldots, X_n$ be independent with $E(X_i) = \mu_i, \operatorname{var}(X_i) = \sigma_i^2$, and $E|X_i - \mu_i|^{2+\delta} < \infty$ for some $0 < \delta \leq 1$. Then*

$$\left| P\left(\frac{\bar{X}_n - E\left(\bar{X}_n\right)}{\sqrt{\operatorname{var}\left(\bar{X}_n\right)}} \leq x \right) - \Phi(x) \right| \leq \frac{D}{1+|x|^{2+\delta}} \frac{\sum_{i=1}^n E|X_i - \mu_i|^{2+\delta}}{\left(\sum_{i=1}^n \sigma_i^2\right)^{1+\frac{\delta}{2}}}$$

for some universal constant $0 < D < \infty$.

有了这个定理之后我们就有了关于相对误差的信息了。

接下来我们关注的问题是如何将这一套关于误差的东西应用在推断上，上面这些东西看上去在理论证明方面大有可为之处，能告诉我们很多很多关于统计量的性质的问题，但是下面这个东西就更妙了，他能直接帮我们提高推断的效果，这个东西就是 Edgeworth expansions

Theorem 3.18 (Two-term Edgeworth expansion) *Suppose F is absolutely continuous distributions and $E_F(X^4) < \infty$. Then*

$$F_{Z_n}(x) = \Phi(x) + \frac{C_1(F)p_1(x)\phi(x)}{\sqrt{n}} + \frac{C_2(F)p_2(x) + C_3(F)p_3(x)}{n} + O\left(n^{-3/2}\right),$$

uniformly in x, where

$$C_1(F) = \frac{E(X-\mu)^3}{6\sigma^3}, C_2(F) = \frac{\frac{E(x-\mu)^4}{\sigma^4} - 3}{24}, C_3(F) = \frac{C_1^2(F)}{72},$$
$$p_1(x) = 1 - x^2, p_2(x) = 3x - x^3, p_3(x) = 10x^3 - 15x - x^5.$$

这个 **Berry-Essen Theorem** 告诉我们

$$|F_{Z_n}(x) - \Phi(x)| = O(n^{-\frac{1}{2}})$$

这个 **Edgeworth expansion** 就是在告诉我们 $O(n^{-1/2})$ 里面有什么，然后你把这些东西用样本一估代进去，就可以提高他的收敛速度，这种思路还是常见的，比如 debias lasso 就有这种思路在里面。

Remark 3.18.1 为什么现在没人用了？

- 只能处理样本均值，缺少泛化性
- 缺少的要数据估计，但是数据量大才能估的好，数据量大本身效果就好
- *Bootstrap* 的出现，泛化性很强，直接把这个打败了

与 *Edgeworth expansion* 对应的还有一个 *Cornish-Fisher* 展开，处理的是逆函数，分位数函数的展开

接下来我们还要介绍一个定理，这个定理可以把速率控制在现在集齐 popular 的指数阶上，这个在Multiple Test上很有帮助，因为这个时候 x 可能会和样本量 n 挂钩。邹老师提到的一个最经典的例子是这个 x 会和样本量挂上钩，也就是说，这个 x 趋于极端的速度也很快，这样的话有个一致的界反而也没什么用，就是绝对值里面这两个相减的东西本来就是以样本的指数阶趋于 0 了，那你给我 bound 住一个一致的 $\sqrt{n}$ 的速率就没用了。（Maybe 想一下 $F(x) \sim 1 - e^x$ wil help。）

Theorem 3.19 (Large deviation for the mean) *Suppose that $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.) random variables with mean zero and variance σ^2, satisfying $E\{\exp(u|X|)\} < \infty$ with some $u > 0$. Then for any $0 \le x \le cn^{1/6}$ and $c > 0$,*

$$\frac{\Pr(\sqrt{n}\bar{X}/\sigma > x)}{1 - \Phi(x)} = \exp\left\{\frac{x^3\kappa}{6\sqrt{n}}\right\}\{1 + o(1)\}$$

4 工具：统计量函数的分布-delta 方法

4.1 delta 方法和多元 delta 方法

接下里我们将介绍统计量函数的分布，请注意，如果我们利用前面的定理，我们只能得到如果 X_n 收敛的 T 的话，那么 $g(X_n)$ 也收敛到 $g(T)$，

但是大多数情况下我们都是 $X_n - \theta$ 收敛到某个分布，这个时候我们如果去考虑 $g(X_n - \theta)$ 的收敛的话意义不是很大，更多时候我们我们想要得到 $g(X_n) - g(\theta)$ 的信息收敛来得到关于 θ 的 inference，这个时候我们就需要 delta 方法

Theorem 4.1 (Delta Theorem) *Let T_n be a sequence of statistics such that*

$$\sqrt{n}\left(T_n - \theta\right) \xrightarrow{d} N\left(0, \sigma^2(\theta)\right).$$

Let $g : \mathbb{R} \to \mathbb{R}$ be once differentiable at θ with $g'(\theta) \neq 0$. Then

$$\sqrt{n}\left[g\left(T_n\right) - g(\theta)\right] \xrightarrow{d} N\left(0, \left[g'(\theta)\right]^2 \sigma^2(\theta)\right).$$

Proof 证明只用到了泰勒展开，首先

$$g\left(x_0 + h\right) = g\left(x_0\right) + hg'\left(x_0\right) + o(h)$$

如果我们有 $T_n - \theta = o_p(1)$ 的话，我们代进去就有

$$g\left(T_n\right) = g(\theta) + \left(T_n - \theta\right) g'(\theta) + o_p\left(T_n - \theta\right)$$

然后我们再弄一下

$$\sqrt{n}\left[g\left(T_n\right) - g(\theta)\right] = \sqrt{n}\left(T_n - \theta\right) g'(\theta) + \sqrt{n} o_p\left(T_n - \theta\right)$$

之后由 *slusky* 定理我们就得到了我们想要的结论。当然这里还用到了小项的传递性 *(Theorem 3.5)*，所以这也解释了为什么这个要在 θ 处展开，要不然这东西根本就不是一个小项嘛，要不然就嵌套不进去了嘛

Remark 4.1.1 我们对这个小项的转换做几点说明：

- 小项的传递性是让你把 $o(h)$ 换成 $o_p(T_n - \theta)$ 这样弄没问题的，你把这个 o 看成一个函数 g 就行了，这在函数里的解释就是 $g(x) = o(x^m), f(x) = o(x^n) \Rightarrow g(f(x)) = o(x^{m+n})$
- 之所以 $\sqrt{n} o_p\left(T_n - \theta\right)$ 还是 $o_p(1)$,你可以从形式上直接转化 $\sqrt{n} o_p\left(\frac{1}{\sqrt{n}}\right) = o_p(1)$，当然你令个函数出来看一下也行是吧

  $$g(x) = o(h), \quad \sqrt{n} g(T_n - \theta) = \sqrt{n}(T_n - \theta)\frac{g(T_n - \theta)}{T_n - \theta}$$

  两边令 n 趋于 ∞，就可以得到 $\sqrt{n} g(T_n - \theta)$ 趋于 0

当然了，我们也不一定要规定原来的渐近分布是正态的，我们只需要有个渐近分布就行了，例如 $a_n(T_n-\theta)\xrightarrow{d}Y$，那么我们就可以得到

$$a_n(T_n-\theta)\xrightarrow{d}[g'(\theta)]Y$$

当然我们可能遇到 $g'(\theta)=0$ 这一种情况，我们只需要再往下展一阶就够了，就是下面这个定理

Theorem 4.2 *Let T_n be a sequence of statistics such that*

$$\sqrt{n}(T_n-\theta)\xrightarrow{d}N\left(0,\sigma^2(\theta)\right).$$

Let g be a real-valued function differentiable $k(\geq 1)$ at θ with $g^{(k)}(\theta)\neq 0$ but $g^{(j)}(\theta)=0$ for $j<k$. Then

$$(\sqrt{n})^k\left[g(T_n)-g(\theta)\right]\xrightarrow{d}\frac{1}{k!}\left[g^{(k)}(\theta)\right]\left[N\left(0,\sigma^2(\theta)\right)\right]^k.$$

当然，一元的不过是小孩子的游戏罢了，真男人还得看多元的，毕竟这才是 research 中常用的，多元的 Delta 方法叙述如下

Theorem 4.3 *Suppose $\{\mathbf{T}_n\}$ is a sequence of k-dimensional random vectors such that $\sqrt{n}(\mathbf{T}_n-\boldsymbol{\theta})\xrightarrow{d}N_k(\mathbf{0},\boldsymbol{\Sigma}(\boldsymbol{\theta}))$. Let $g:\mathbb{R}^k\to\mathbb{R}^m$ be once differentiable at $\boldsymbol{\theta}$ with the gradient matrix $\nabla g(\boldsymbol{\theta})$. Then*

$$\sqrt{n}\left(g(\mathbf{T}_n)-g(\boldsymbol{\theta})\right)\xrightarrow{d}N_m\left(\mathbf{0},\nabla^T g(\boldsymbol{\theta})\boldsymbol{\Sigma}(\boldsymbol{\theta})\nabla g(\boldsymbol{\theta})\right)$$

provided $\nabla^T g(\boldsymbol{\theta})\boldsymbol{\Sigma}(\boldsymbol{\theta})\nabla g(\boldsymbol{\theta})$ is positive definite.

对于这一套方法，我没什么好 comment 的，只能说这确实给这种多元的函数找到了一种通用的方法来处理这一类问题，当然 maybe 计算过程是十分复杂的，但是起码这是一定能做的。

4.2 Variance-stabilizing transformations

这个东西的主要思想就是一般我们用来做置信区间和假设检验的式子是 $T_n-\theta\sim N(0,\sigma^2(\theta))$，如果你用这个东西的话，我们还要去估计一下 $\sigma(\theta)$，但这个东西你不一定能估好，所以我们想找一个函数 g，使得 $g(T_n)-g(\theta)$ 的方差与 θ 无关，这样的话我们就不用估计方差了，下面简述一下这个过程，假设我们找到的函数是 g，那么我们有

$$\sqrt{n}\left(g(T_n)-g(\theta)\right)\xrightarrow{d}N\left(0,\left[g'(\theta)\right]^2\sigma^2(\theta)\right)$$

如果我们想要方差是常数，那么这就等价于

$$[g'(\theta)]^2 \sigma^2(\theta) = k^2$$

那么我们就可以反解出 g

$$g(\theta) = k \int \frac{1}{\sigma(\theta)} d\theta$$

下面我们可以给出一个例子

Example 4.1 *Suppose $X_1, X_2, \cdots$, are iid Poisson(θ).*

首先我们有 $\sqrt{n}(\bar{X}_n - \theta) \xrightarrow{d} N(0, \theta)$，所以说 $\sigma(\theta) = \sqrt{\theta}$，所以 g 的选择应该是

$$g(\theta) = \int \frac{k}{\sqrt{\theta}} d\theta = 2k\sqrt{\theta}$$

所以我们可以得到 $\sqrt{n}\left(\sqrt{\bar{X}_n} - \sqrt{\theta}\right) \xrightarrow{d} N(0, 1/4)$，所以我们可以得到 θ 的一个渐进的置信区间为

$$\left\{ \left(\sqrt{\bar{X}_n} - \frac{z_\alpha}{2\sqrt{n}} \right)^2, \left(\sqrt{\bar{X}_n} + \frac{z_\alpha}{2\sqrt{n}} \right)^2 \right\}$$

当然如果左边平方里面这个东西小于 0 的话，我们可以直接把左边写成 0，但是其注意，这个东西你是没有估计方差这一步

一个更复杂的有趣的关于相关系数的例子可以在课件上找到。总的来说，这个方法还是蛮好的，但是有的时候这个（maybe 大多数时候），这个函数 g 没有显示表达式，这就没法弄了，所以这个局限性还是蛮大的

5 检验分布差异

5.1 经验分布函数 ECDF

这一小节主要在讨论经验函数，就是讨论经验分布函数 (Empirical Cumulative Distribution Function , ECDF)

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \le x\}}$$

当然我们在这里讨论的是一维的情况，高维的时候这个东西会出现维数祸根问题。下面是这个东西的一些性质，这些都是显然的，只要你看到了这个经验分布函数不过是一堆独立同分布的伯努利变量的 summation

Theorem 5.1 *For fixed $x, x \in (-\infty, \infty)$, (i) $F_n(x)$ is unbiased and has variance*

$$\text{var}\left[F_n(x)\right] = \frac{F(x)[1-F(x)]}{n};$$

(ii) $F_n(x)$ is consistent in mean square, i.e., $F_n(x) \xrightarrow{2nd} F(x)$;
(iii) $F_n(x) \xrightarrow{wp1} F(x)$;
(iv) $F_n(x)$ is AN $\left(F(x), \frac{F(x)[1-F(x)]}{n}\right)$.

不过这些东西都是对于一个 fixed 的 x 而言的，下面我们关注一个 global 的，就是全局刻画 F_n 和 F 的接近程度，即 Kolmogorov-Smirnov 距离，就是 K-S 距离。

$$D_n = \sup_{-\infty<x<\infty} |F_n(x) - F(x)|$$

有的时候我们还用无限维的距离（最大距离）来表示，就是记为 $||F_n(x) - F(x)||_\infty$，然后我们就有一个定理来给出这个东西的界

Theorem 5.2 (DKW's inequality) *Let F_n be the ECDF based on iid $X_1, \ldots, X_n$ from a CDF F defined on R. There exists a positive constant C (not depending on F) such that*

$$P\left(D_n > z\right) \leq Ce^{-2nz^2}, z > 0, \text{ for all } n = 1, 2, \ldots$$

Remark 5.2.1 首先这个 C 是 *global* 的，不随 x 变动的，然后这个接近的速度是指数的速率，这个东西就很好，比如你做 *Bonferri* 的时候，前面就算是乘上一个 n 还是收敛到了 *0*

当然了我们可以把这个东西写的形式上更好看一点，就是

$$P(\sqrt{n}D_n > z) \leq Ce^{-2z^2}$$

这样的话我们可以立即得出，$\sqrt{n}D_n = O_p(1)$，并且 $C = 2$ 是一个最小的参数了，不过如果你只关系速率的，这个 doesn't matter。我们还可以得到下面这个定理

Theorem 5.3 *Let F and C be as in DKW Theorem. Then for every $\epsilon > 0$,*

$$P\left(\sup_{m\geq n} D_m > \epsilon\right) \leq \frac{C}{1-h_\epsilon} h_\epsilon^n,$$

where $h_\epsilon = \exp(-2\epsilon^2)$.

这个东西的证明过程就是 Borel-Cantelli 引理

$$P\left(\sup_{m\geq n} D_m > \epsilon\right) \leq \sum_{m=n}^{\infty} P\left(D_m > \epsilon\right) \leq C \sum_{m=n}^{\infty} h_\epsilon^m = \frac{C}{1-h_\epsilon} h_\epsilon^n$$

当然你就可以马上得到

Theorem 5.4 (Clivenko-Cantelli)

$$D_n \overset{wp1}{\to} 0$$

当然写到这个我们看到 $\sqrt{n}D_n$ 是 O(1) 的，D_n 是小 o(1) 的，我们就可以猜了 $\sqrt{n}D_n$ 是不是有个渐近分布，(personal view)→ 虽然看到这里你觉得很有道理，不过我觉得这简直 bullshit，真正这个东西太复杂了，只有靠 Kolmogorov 这样的超级大神才能弄。anyway，最后我们得到了

Theorem 5.5 (Kolmogorov) *Let F be continuous. Then*

$$\lim_{n\to\infty} P\left(\sqrt{n}D_n \leq z\right) = 1 - 2\sum_{j=1}^{\infty}(-1)^{j+1} e^{-2j^2z^2}, z > 0$$

不过这个东西还是太复杂了，而且近似程度不太好（这个我不知道来源嗷），不过这个东西有个惊为天人的东西，就是这个渐近分布是和原分布无关的，就是是 distribution free 的，所我们可靠一个简单的分布模型来得到他对应分位数或者分布，这个意思就是

$$\sqrt{n}D_n \overset{d}{=} \sqrt{n} \max_{0\leq i\leq n} \max\left(\frac{i}{n} - U_{(i)}, U_{(i)} - \frac{i-1}{n}\right)$$

所以如果我们想要知道这个渐近的分位数，跑一堆均匀分布得到分位数就可以了。

Remark 5.5.1 *anyway，统计始终没有免费的午餐，这个东西得到了 distribution free，得到了 flexible，那么他就终将失去了一些东西*

- 计算量，这个好理解，如果你可以直接算出分位数，那就没有复杂度，但是这个你是要通过大量计算才能得到的

- 精确程度，这个就是说，如果你也得到准确的分位数，这个近似模拟得到的分位数肯定是不精确的，至于为什么说上面个那个东西近似程度不好，我猜应该是这个要是算到无限项就是精确的，但是不是无限性就收敛的比较慢吧。

我们可以利用这个搞一下分布函数（**CDF**）的渐近分布，当然这都建立在我们已经得到了这个恶心的分布的分位数的基础上。

Example 5.1 (Kolmogorov-Smirnov confidence intervals) *We know given $\alpha \in (0,1)$, there is a well-defined $d = d_{\alpha,n}$ such that, for any continuous* CDF $F, P_F\left(\sqrt{n}D_n > d\right) = \alpha$.

$$\begin{aligned} 1-\alpha &= P_F\left(\sqrt{n}D_n \le d\right) = P_F\left(\sqrt{n}\left\|F_n - F\right\|_\infty \le d\right) \\ &= P_F\left(|F_n - F| \le \frac{d}{\sqrt{n}}, \forall x\right) \\ &= P_F\left(F_n(x) - \frac{d}{\sqrt{n}} \le F(x) \le F_n(x) + \frac{d}{\sqrt{n}}, \forall x\right). \end{aligned}$$

最后我们就得到了，区间估计就是

$$KS_{n,\alpha} : \left\{\max\left(0, F_n(x) - \frac{d}{\sqrt{n}}\right) \le F(x) \le \min\left(1, F_n(x) + \frac{d}{\sqrt{n}}\right)\right\}$$

当然这个东西是和 x 无关的，这就可能出现我们之前说的问题，那就是在有些地方的估计效果会很差，比如两端。

5.2 卡方检验 Chi-square test

我们刚刚在上面介绍了利用 KS 统计量检验分布的原假设下的结果，接下来介绍非参数的方法 chi-square，作为一种非参数方法，他更加的 flexible，更加的 versatile，不过这是 efficiency 的 trade-off，他的主题思想就是，你分成几个区域，如果你真的服从原假设，你在这个区域内的点的分布应该是已知的，然后统一几个区域点的数量和期望的差距就可以得到 Chi-square 统计量，那么提出的统计量就是

$$K^2 = \sum_{i=1}^{k} \frac{\left(n_i - np_{0i}\right)^2}{np_{0i}}$$

这里的符号我就不定义了，the same as you think。那么下面这个定理说明了这个东西在渐近分布下的渐近分布

Theorem 5.6 *(The asymptotic null distribution) Suppose $X_1, X_2, \ldots X_n$ are iid observations from some distribution F. Consider testing $H_0 : F = F_0$ (specified). $K^2 \xrightarrow{d} \chi^2_{k-1}$ under H_0.*

Proof 我们首先还是要找到加和的形式，令 $\mathbf{n} = (n_1, \ldots, n_k)^T = \sum_{i=1}^n \mathbf{Z}_i$, 其中 $\mathbf{Z}_i = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ 代表 i 这个点最终落入了哪个区域，这样我们就得到了这个

$$\mathbf{Y} = (Y_1, \ldots, Y_k)^T = \left(\frac{n_1 - np_{01}}{\sqrt{np_{01}}}, \ldots, \frac{n_k - np_{0k}}{\sqrt{np_{0k}}} \right)^T$$

的渐近分布,即 $\mathbf{Y} \xrightarrow{d} N(0, \mathbf{\Sigma})$, 其中 $\mathbf{\Sigma} = \mathbf{I}_k - \boldsymbol{\mu}\boldsymbol{\mu}^T$ *and* $\boldsymbol{\mu} = \left(\sqrt{p_{01}}, \ldots, \sqrt{p_{0k}}\right)^T$，我们需要观察到，$\mathbf{Y}$ 的协方差矩阵是一个幂等矩阵，接下里就是常规的分解成正交阵的过程，因为是幂等矩阵，特征值就全是 *0* 或 *1*，我们把 Y^TY 就可以写成 $X^T\mathbf{\Sigma}X$ 的形式，然后再对 $\mathbf{\Sigma}$ 做奇异值分解就行了，自由度是多少，我们算下 *trace* 就行了，显然是 *k-1*

下面我们对这个东西做一点拓展，考虑如果 np_{0i} 非常小的情况，这表示这个统计量会稍微有点不稳定，我们想要搞一个转换 $\mathbf{g}(\mathbf{x})$, 把这个分母给弄掉，当然为了简单起见，我们使用这样的 $\mathbf{g}(\mathbf{x}) = (g_1(x_1), \ldots, g_k(x_k))^T$, 这样的梯度矩阵就是一个对角阵，我们可以知道 $\sqrt{n}\left(\mathbf{g}\left(\overline{\mathbf{Z}}_n\right) - \mathbf{g}(\mathbf{p}_0)\right)$ 和 $\sqrt{n}\nabla^{-1}\mathbf{g}(\mathbf{p}_0)\left(\mathbf{g}\left(\overline{\mathbf{Z}}_n\right) - g(\mathbf{p}_0)\right)$ 同分布 (delta 方法),这样我们就得到了 $\mathbf{g}\left(\overline{\mathbf{Z}}_n\right) - g(\mathbf{p}_0)$ 弄成卡方分布的过程

$$\begin{aligned}
\chi^2_g =& n\left(\mathbf{g}\left(\overline{\mathbf{Z}}_n\right) - \mathbf{g}(\mathbf{p}_0)\right)^T \nabla^{-1}\mathbf{g}(\mathbf{p}_0)\,\mathrm{diag}(\mathbf{p})\nabla^{-1}\mathbf{g}(\mathbf{p}_0) \\
& \left(\mathbf{g}\left(\overline{\mathbf{Z}}_n\right) - \mathbf{g}(\mathbf{p}_0)\right) \\
=& n\sum_{i=1}^k \frac{\left(g_i(n_i/n) - g_i(p_{0i})\right)^2}{p_{0i}\left[g_i'(p_{0i})\right]^2} \xrightarrow{d} \chi^2_{k-1}.
\end{aligned}$$

所以为了把分母弄掉，我们只要令 $\mathbf{g}(\mathbf{x}) = \left(\sqrt{x_1}, \ldots, \sqrt{x_k}\right)^T$，所以最后我们就得到统计量

$$\chi^2_H = 4n\sum_{i=1}^k \left(\sqrt{n_i/n} - \sqrt{p_{0i}}\right)^2$$

这个东西又叫 Hellinger χ^2，这是因为 Hellinger 距离的定义方式是

$$d^2(f, g) = \int (\sqrt{f(x)} - \sqrt{g(x)})^2 dx$$

接下来我们考虑备择假设下的渐近分布，就是在考虑功效，我们可以得到下面这个定理

Theorem 5.7 *Under F_1, (i) $\frac{K^2}{n} \xrightarrow{p} \sum_{i=1}^{k} \frac{(p_{1i}-p_{0i})^2}{p_{0i}}$.*
(ii) If $\sum_{i=1}^{k} \frac{(p_{1i}-p_{0i})^2}{p_{0i}} > 0$, then $K_P^2 \xrightarrow{p} \infty$ and hence the Pearson χ^2 test is consistent against F_1.

首先我们可以看到这个检验统计量就是在检验 p_{1i} 和 p_{0i} 的差异，这个意思就是说这个东西就是信号。其次我们可以看到在备择假设下，这个东西肯定是趋于无穷了，当然这对于功效来说是好的，但是如果我们要比较两个检验统计量的效率，这个东西看上去也没这么好是吧，毕竟如果大家都是趋于无穷，或者说大家在备择假设下 p 值都趋于 0 了，这个东西怎么比较呢？我们可以比较趋于 0 的速率是吧，但是我们可以等价？地考虑另一个方法，那就是让信号趋于 0，可以肯定的是，如果信号以一个比较快的速率趋于 0，所有的统计量都检验不到了，所以我们可以找到那个最大的速度就行了。我们可以给出下面这个定理：

Theorem 5.8 *(The asymptotic alternative distribution) Under H_1, say $\mathbf{p} = \mathbf{p}_1 = \mathbf{p}_0 + \delta n^{-1/2}$. Then $K^2 \xrightarrow{d} \chi^2_{k-1}(\lambda)$, where $\lambda = \sum_{i=1}^{k} \delta_i^2/p_{0i}$ is the noncentrality parameter.*

我们可以看到速率是 $n^{\frac{1}{2}}$ 时，备择假设下也有了渐近分布，这说明这个速率再大就趋于无穷了，再小，就和原假设没区别，所以这个最大能检验的速率就是 $\frac{1}{2}$。
最后我们总结一下，ECDF 是参数的方法，χ^2 是非参数的方法，前者更有效率后者更加 flexible。不过这种检验也没什用对吧，很少会让你检验这个数据是不是来自于一个特定的分布，更多的时候我们关注他们是不是来自于一个函数族，比如我们关心他们是不是来自于正态的函数族，而不是是不是来自于 $N(0,1)$。当然我们可以估计参数再使用，不过这样 ECDF 就会出问题，而 χ^2 不会。而且 χ^2 很容易就推广到了两样本是不是来自于同一分布的检验，所以看上去似乎 χ^2 更加优越一些。

6 样本矩 Sample Moment

首先我们定义一些符号，分别定义了总体形式和样本形式下的矩

$$\alpha_k = \int_{-\infty}^{\infty} x^k dF(x) = EX_1^k$$
$$\mu_k = \int_{-\infty}^{\infty} (x-\alpha_1)^k dF(x) = E\left[(X_1-\alpha_1)^k\right]$$
$$a_k = \int_{-\infty}^{\infty} x^k dF_n(x) = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \quad k = 1, 2, \ldots,$$
$$b_k = \frac{1}{n}\sum_{i=1}^{n} (X_i-\alpha_1)^k$$
$$m_k = \int_{-\infty}^{\infty} (x-a_1)^k dF_n(x) = \frac{1}{n}\sum (X_i-a_1)^k, k = 2, 3, \ldots =$$

当然对于 b_k 和 a_k 你都没什么好弄的，这都是些独立的 summation 嘛，这直接就是大数定律加上 CLT，主要是最后 m_k 这个不是独立的 summation 的形式，我们需要额外关心一下，我们有下面这个定理

Theorem 6.1 *Suppose that $\mu_{2k} < \infty$. (i) $m_k \overset{wp1}{\to} \mu_k$; (ii) The random vector $\sqrt{n}(m_2-\mu_2, \ldots, m_k-\mu_k)^T$ is $AN_{k-1}(\mathbf{0}, \mathbf{\Sigma}^*)$, where $\mathbf{\Sigma}^* = \left(\sigma_{ij}^*\right)_{(k-1)\times(k-1)}$ with $\sigma_{ij}^* = \mu_{i+j+2} - \mu_{i+1}\mu_{j+1} - (i+1)\mu_i\mu_{j+2} - (j+1)\mu_{i+2}\mu_j + (i+1)(j+1)\mu_i\mu_j\mu_2$*

在证明这个定理之前,我们首先找一下 b_k 的联合渐近分布很简单 $\sqrt{n}(b_1-\mu_1, \ldots, b_k-\mu_k)^T \overset{d}{\to} AN_k(\mathbf{0}, \widetilde{\mathbf{\Sigma}})$, where $\widetilde{\mathbf{\Sigma}} = (\widetilde{\sigma}_{ij})_{k\times k}$ with $\widetilde{\sigma}_{ij} = \mu_{i+j} - \mu_i\mu_j$.

Proof 首先我们要把它拆解开

$$m_k = \frac{1}{n}\sum_{i=1}^{n} (X_i-a_1)^k = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{k} C_k^j (X_i-\alpha_1)^j (\alpha_1-a_1)^{k-j}$$

这样我们就得到了

$$m_k = \sum_{j=0}^{k} C_k^j (-1)^{k-j} b_j b_1^{k-j}$$

其中 $b_0 = 1$，这样的话就被我们弄到 b_k 上去了，但是 b_k 的渐近分布我们是已知的，所以我们可以直接利用 *Delta* 方法，令

$$g(t_1, \ldots, t_k) = \left(\sum_{j=0}^{2} C_2^j (-1)^{2-j} t_j t_1^{2-j}, \ldots, \sum_{j=0}^{k} C_k^j (-1)^{k-j} t_j t_1^{k-j}\right)^T$$

然后令 $\boldsymbol{\theta} = (0, \mu_2, \ldots, \mu_k)^T$，那么 $g(\boldsymbol{\theta}) = (\mu_2, \ldots, \mu_k)^T$. 计算 ∇g

$$\nabla^T g_{|\theta} = \begin{pmatrix} -2\mu_1 & 1 & 0 & \cdots & 0 \\ \vdots & & \ddots & & \\ -(i+1)\mu_i & 0 & \cdots & 1 & \cdots \\ \vdots & & \vdots & & \\ -k\mu_{k-1} & 0 & \cdots & & 1 \end{pmatrix}$$

$\boldsymbol{\theta}$ 第一个是 0 是因为 $\mu_1 = 0$，这也是计算梯度的时候的关键，然后我们就得到了

$$\boldsymbol{\Sigma}^* = \nabla^T g_{|\theta} \widetilde{\boldsymbol{\Sigma}} \nabla g_{|\theta}$$

关于这个矩阵的计算，我们如果把 $\widetilde{\boldsymbol{\Sigma}}$ 分解成为 $\Sigma - \mu\mu^\top$ 的话会好算一点，如果进一步分块的话，还可以写成

$$\nabla^\top \tilde{\Sigma} \nabla = \left(\nabla_1^\top, I\right) \left(\hat{\Sigma} - \mu\mu^\top\right) \begin{pmatrix} \nabla_1 \\ I \end{pmatrix}$$

不过这样貌似更难算了，比如就只对 $\widetilde{\boldsymbol{\Sigma}}$ 分解然后算