1 内容补充

本次内容补充主要是第二周课堂知识的进一步讨论，没有什么升级的东西，所以放在前面

1.1 Theorem (【2】定理 1.9)

设 $X_1, X_2, \cdots$ i.i.d. $\sim X, \mu_4 = \mathbb{E}(X - \mathbb{E}X)^4$ 存在有限, 记 $\mathbb{E}(X) = \mu$, $\mathrm{Var}(X) = \sigma^2$, 若函数 $h(x)$ 的四阶导数存在且有界, 则有

$$\begin{aligned}\mathbb{E}\left[h\left(\bar{X}_n\right)\right] =& h(\mu) + \frac{1}{2n}h''(\mu)\sigma^2 + O\left(n^{-2}\right)\\ \mathrm{Var}\left[h\left(\bar{X}_n\right)\right] =& \frac{1}{n}\left[h'(\mu)\right]^2\sigma^2 + \frac{1}{2n^2}\left\{h'(\mu)h''(\mu)\mu_3 + \left[h''(\mu)\right]^2\sigma^4\right.\\ &\left.+h'(\mu)h'''(\mu)\sigma^4\right\} + O\left(n^{-3}\right)\end{aligned}$$

评注：首先是这个定理，不太容易记住，对吧？我的建议是：与其记住这个定理的条件与结论，不如记住它的思想方法与推导过程。

其大致的证明过程就是：**对 $h(x)$ 在 $x = EX$ 处进行带 Lagrange 余项类型的 Taylor 展开，然后利用余项有界就可以了**。那怎么保证余项有界呢？一种自然的充分条件就是导数有界与四阶中心矩有界。所以**请一定要掌握其中的方法与思想**。即使忘记了它的结果的具体形式，自己推导一遍，也是花不了太多时间的。这比死记硬背条件与结果要有意义（当然，假如你记忆力很强，就当我没说）。

另外，既然对于一元函数可以 Taylor 展开，自然可以对于多元函数进行 Taylor 展开，这就是本次的 10.4 题。

还有，这个命题与 Delta 方法、Slutsky 定理、CLT、(S)LLN，都可以说是最最 fundamental 的结论与方法。这一点也请留意一下。

另外，请思考一下：**这个命题与 DELTA 方法的本质是什么？**就是 Taylor 展开，对吧？那 Taylor 展开的本质是什么？多项式逼近光滑函数，对吧？所以，从这个角度来看，就能理解为什么这些定理或者理论会这样发展了。因为我们不仅想估计随机变量 X 的期望、方差、极限分布/近似分布，还想估计 $f(X)$ 的。但是我们一般只知道 X 的矩的信息，那么就知道了 $P(X)$ 的信息 (P 为多项式)，所以自然就会用 Taylor 展开去寻找 $f(x)$ 的相关信息。

1.2 Theorem(【0】定理 3.2.2,【2】定理 2.10)

设 $X_1, X_2, \cdots$ i.i.d. $\sim X$, 总体 X 的 α_{2k} 有限。记 $\vec{\theta} = (\theta_1, \cdots, \theta_d)^T$, 其中 $\theta_i = g_i(\alpha_1, \cdots, \alpha_k)$, 若 g_i 关于 α_j 有连续偏导数, $i = 1, \cdots, d$, $j = 1, \cdots, s$, 则对 $\vec{\theta}$ 的矩估计 $\vec{\theta}_{MOM} = (g_1(\hat{\boldsymbol{\alpha}}), \cdots, g_d(\hat{\boldsymbol{\alpha}}))^T$, 其中 $\hat{\alpha} = (\hat{\alpha}_1, \cdots, \hat{\alpha}_k)^T$, 有

$$\sqrt{n}\left(\hat{\vec{\theta}}_{MOM} - \vec{\theta}\right) \xrightarrow{L} N_d\left(\overrightarrow{0}, G\Sigma G^T\right)$$

其中 $G = \left(\frac{\partial g_i}{\partial \alpha_j}\right)_{d\times k}$, Σ 是 $k \times k$ 阶矩阵, 其 (i, j) 元素为 $\alpha_{i+j} - \alpha_i\alpha_j$.

评注: 这个定理的证明过程群文件已经发出来了, 大致思路就是多元 CLT+ 多元 Delta 方法, 其实这在第一次习题课讲义的补充里面都有例子: 求 $(\bar{X}_n - \mu, S_n^2 - \sigma^2)$ 的极限分布.

那么该如何记这个结果呢? 可以这么来看:

首先, 看协方差的表示形式 $G\Sigma G^T$, 这种矩阵相合形式在概率里面哪里经常出现呢?

Proposition: 设 $X \in \mathbb{R}^p$ 为随机向量, $Cov(X) = \Sigma$ 为协方差矩阵, $A \in \mathbb{R}^{m\times p}$ 为实矩阵, 则 $Cov(AX) = A\Sigma A^T$.

命题的证明此处就省略了, 就是根据定义来写的。(另外, 剧透一下: 这个命题在兰老师下学期的回归分析的课程里面会用得特别多。)

那么为什么这个命题与这里的协方差矩阵长得这么相似呢?

前面说过, Delta 方法的本质就是 Taylor 展开, Taylor 的本质就是多项式逼近, 而这里只需要线性 (一次) 逼近, 所以就用相当于微分映射 dg 来代替 g, 而微分映射 dg 的具体表示就是 Jacboi 矩阵 G, 所以决定 $g(X)$ 极限分布协方差的自然就是 G 与 Σ 了。而 $\Sigma = Cov(Y)$, $Y = (X, X^2, \cdots, X^k)$, $Cov(X^i, X^j) = \alpha_{i+j} - \alpha_i\alpha_j$。

所以这么来看, 就比较容易记住这个命题了。

1.3 【例 5.5.23】

设随机变量 X 的期望为 $\mathbb{E}_\mu X = \mu \neq 0$, 令 $g(\mu) = \frac{1}{\mu}$, 请给出 $g(\mu)$ 的一个矩估计, 并给出其期望和方差的近似估计。

评注: 这个题有不少同学都有一些疑惑, 12/14 课后有些同学与兰老师讨论了一下。

首先, 这里的条件肯定是不够的。所以需要加一些条件。

其次，我们的估计量有两种选法，一种是 $\frac{n}{\sum X_i}$，一种是 $\frac{1}{n}\sum \frac{1}{X_i}$，也就是调和平均与倒数的算术平均，二者也有确定的大小关系，那哪个更好呢? 前者。有以下几条原因：

1. 前者相比后者更“容易”收敛。不妨假设 $X > 0$ a.s. 成立，那么，假如有 1 一个样本 $X_i = x$ 比较接近 0，此时 $\frac{1}{n}\sum \frac{1}{X_i}$ 就容易爆炸（意思是，数值会变得异常大），但是 $\frac{n}{\sum X_i}$ 不会，因为有前面的样本“撑着”的。

2. 即使假设一些条件，比如说 $P(X > \epsilon) = 1 - \delta$，则在 $1 - \delta$ 的概率下，我们可以对 $\frac{n}{\sum X_i}$ 与 $\frac{1}{n}\sum \frac{1}{X_i}$ 使用有界收敛定理，它们的期望与方差都会收敛，收敛序列必定有界，所以它们的期望与方差是有界的。但是后者的期望是 $E\frac{1}{X}$，这个量很难说就是 $1/EX$，而且由柯西不等式 $E\frac{1}{X}EX \geq 1$，所以 $E\frac{1}{X}$ 一般都会更大。

综上，假设 $P(X > \epsilon) = 1 - \delta$ 成立（$\epsilon > 0, 0 < \delta < 1$），那么在 $1 - \delta$ 概率下，我们是可以使用 Taylor 展开对估计量 $\frac{n}{\sum X_i}$ 进行均值与方差的估计的。

另外，不要觉得形如“在 $1 - \delta$ 概率下”的命题很奇怪，实际上，在很多统计与机器学习的著作与文章里面，是经常会出现形如"With probability at least $1 - \delta$, the following statement holds" 的论述的，比如很多算法的理论分析里面，就会加这么一句话作为大前提。

2 第 2 次作业

10.1

$X_1, \cdots, X_n$ 是抽自概率密度函数为

$$f(x \mid \theta) = \frac{1}{2}(1 + \theta x), -1 < x < 1, -1 < \theta < 1$$

的总体的随机样本. 求 θ 的一个相合估计量，证明其相合性，并讨论其极限方差与渐近方差.

解答：

考虑一阶矩估计（期望）：

$$EX = \int x f(x|\theta) dx = \int_{-1}^{1} \frac{1}{2}(1 + \theta x) x dx = \frac{\theta}{3}$$

因此，考虑估计量 $Y = 3\bar{X}$，由（强/弱）大数律可得 Y 的（强/弱）相合性.
为求其极限方差，先求 X 的方差

$$EX^2 = \int x^2 f(x|\theta)dx = \int_{-1}^{1} \frac{1}{2}(1+\theta x)x^2 dx = 1/3 \implies Var(X) = 1/3 - \theta^2/9$$

所以

$$Var(Y) = \frac{9}{n}Var(X) = \frac{1}{n}(3-\theta^2)$$

故 Y 的极限方差为 $3-\theta^2$.
注意 Y 的渐近正态性可由 CLT 导出，故其渐近方差与极限方差相同.
注：极限方差与渐进方差在相差一个常数因子意义下是唯一的. 例如，将定义中 k_n 的取成 $2k_n$，渐近方差与极限方差就会差一个常数因子.

批改反馈：有些同学把“极限方差”把“相合估计”的定义弄错了。

10.4

令随机变量 $Y_1, \cdots, Y_n$ 满足 $Y_i = \beta X_i + \epsilon_i, i = 1, \cdots, n$, 其中 $X_1, \cdots, X_n$ 为独立 $N(\mu, \tau^2)$ 随机变量, $\epsilon_1, \cdots, \epsilon_n$ 为 i.i.d. 的 $N(0, \sigma^2)$ 的, 且各个 X 与各个 ϵ 独立. 精确的方差计算很困难, 因而我们可以采取近似的方法. 求出下列各量的近似均值和方差并用 μ, τ^2 和 σ^2 表示:
(a) $\sum X_i Y_i / \sum X_i^2$. (b) $\sum Y_i / \sum X_i$. (c) $\sum (Y_i/X_i)/n$.

解法 1：

首先，以下三问均应该把 Y 用 X 与 ϵ 表示出来.
(a) $Z = \beta + W, W = \sum_{i=1}^n \epsilon_i X_i / \sum_{i=1}^n X_i^2 = \overline{\epsilon X}/\overline{X^2}$
(b) $Z = \beta + W, W = \sum_{i=1}^n \epsilon_i / \sum_{i=1}^n X_i = \bar{\epsilon}/\bar{X}$
(c) $Z = \beta + \overline{\epsilon/X}$.
故 (c) 中的 Z 为独立和的平均，容易算，先算 (c)：注意 ϵ 与 X 的独立性，所以，$EZ = \beta$.
$Var(Z) = \frac{1}{n}Var(\epsilon/X) = \frac{1}{n}E(\epsilon^2/X^2) = \frac{1}{n}E(\epsilon^2)E\frac{1}{X^2} = \frac{\sigma^2}{n}E\frac{1}{X^2}$.
将正态分布密度函数带入可知，$E\frac{1}{X^2} = \infty$.
故 $Var(Z) = \infty$.
(a)(b) 中的 W 均为 x/y 的形式，所以考虑二元函数 $h(x,y) = \frac{x}{y}$ 在 $(x,y) =$

(x_0, y_0) 处 Taylor 展开;

$$\begin{aligned}h(x,y) =& h(x_0,y_0) + (x-x_0)h_x(x_0,y_0) + (y-y_0)h_y(x_0,y_0) + \\ & \frac{1}{2}[(x-x_0)^2 h_{xx}(x_0,y_0) + (y-y_0)^2 h_{yy}(x_0,y_0) + 2(x-x_0)(y-y_0)h_{xy}(x_0,y_0)] + O(r^3) \\ =& \frac{x_0}{y_0} + \frac{x-x_0}{y_0} - \frac{(y-y_0)x_0}{y_0^2} + \frac{(y-y_0)^2 x_0}{y_0^3} - \frac{(x-x_0)(y-y_0)}{y_0^2} + O(r^3),\end{aligned}$$

其中，$r = \sqrt{(x-x_0)^2 + (y-y_0)^2}$.

对于 (a):

$(x_0, y_0) = (E[\epsilon X], E[X^2]) = (0, \mu^2 + \tau^2) \implies h(x,y) = \frac{x}{y_0} - \frac{x(y-y_0)}{y_0^2} + O(r^3)$

所以，$E(W) = 0 \implies EZ = \beta$.

$$\begin{aligned}EW^2 &\approx E[\frac{x}{y_0} - \frac{x(y-y_0)}{y_0^2}]^2 \\ &= E[\overline{\epsilon X}(\frac{2}{y_0} - \frac{\overline{X^2}}{y_0^2})]^2 \\ &\approx E[\frac{\overline{\epsilon X}}{y_0}]^2 \text{(这里实在太难算了，所以把二阶项也忽略了)} \\ &= \frac{E(\sum_{i=1}^n \epsilon_i X_i)^2}{n^2 y_0^2} \\ &= \frac{1}{n^2 y_0^2}(\sum_i E\epsilon_i^2 X_i^2 + 2\sum_{i<j} E\epsilon_i X_i \epsilon_j X_j) \\ &= \frac{1}{n^2 y_0^2}(n\sigma^2 y_0^2 + 2\sum_{i<j} E\epsilon_i EX_i E\epsilon_j EX_j) \\ &= \frac{\sigma^2}{n}.\end{aligned}$$

所以，$Var(Z) = Var(W) = EW^2 = \frac{\sigma^2}{n}$

对于 (b):

$(x_0, y_0) = (E[\epsilon], E[X]) = (0, \mu) \implies h(x,y) = \frac{x}{y_0} - \frac{x(y-y_0)}{y_0^2} + O(r^3)$

所以，$E(W) = 0 \implies EZ = \beta$.

$$\begin{aligned}
EW^2 &\approx E[\frac{x}{y_0} - \frac{x(y-y_0)}{y_0^2}]^2 \\
&= E[\bar{\epsilon}(\frac{2}{y_0} - \frac{\overline{X}}{y_0^2})]^2 \\
&= E\bar{\epsilon}^2 \cdot E(\frac{2}{y_0} - \frac{\overline{X}}{y_0^2})^2 \\
&= \frac{\sigma^2}{n} \cdot E(\frac{2}{\mu} - \frac{\overline{X}}{\mu^2})^2 \\
&= \frac{\sigma^2}{n} \cdot \frac{E\bar{X}^2}{\mu^2} \\
&= \frac{\sigma^2}{n}(1 + \frac{\tau^2}{n\mu^2})
\end{aligned}$$

所以，$Var(Z) = Var(W) = \frac{\sigma^2}{n}(1 + \frac{\tau^2}{n\mu^2})$.

评注 1: 关于三个小问的期望，都是可以严格证明等于 β（后面会证），关于方差，(c) 可以严格证明为无穷大（证明），因此，就不需要用 Taylor 展开估计了，主要是 (a),(b) 两问的方差估计可以有方法。

评注 2：这题有不少伪证，比如最常见的伪证就是

$$E(X/Y) = EX/EY \; Var(X/Y) = Var(X)/Var(Y),$$

注意期望算子只有对于独立的随机变量才有 $EXY = EXEY$, 而方差算子则根本没有这个性质，所以请不要乱写证明!

虽然这道题是让你估计，**但是不能天马行空、为所欲为地乱估计**，总要有一些理由，好吧?

不过，还是有些同学给出了除了上述做法以外比较好的解法的，针对 (a) 问，写一下。

解法 2

注意 $EW = \sum_{i=1}^n E[\epsilon_i X_i / \sum_{i=1}^n X_i^2]$, 又由于每个 ϵ_i 均与所有的 $X_1, \cdots, X_n$ 独立，因此 $EW = \sum_{i=1}^n E[\epsilon_i]E[X_i / \sum_{i=1}^n X_i^2] = 0$.
所以 $Var(W) = EW^2 = E[(\sum_{i=1}^n \epsilon_i X_i / \sum_{i=1}^n X_i^2)^2] = E[\sum_{i=1}^n \epsilon_i^2 X_i^2 / S(X)^2 + \sum_{i=1}^n \epsilon_i \epsilon_j X_i X_j / S(X)^2] = \sum_{i=1}^n E[\epsilon_i^2 X_i^2 / S(X)^2] + \sum_{i \neq j} E[\epsilon_i \epsilon_j X_i X_j / S(X)^2] = \sum_{i=1}^n E[\epsilon_i^2] E[X_i^2 / S(X)^2] = \sigma^2 E[1/S(X)]$，其中 $S(X) = \sum_{i=1}^n X_i^2$. 严格来说，$S(X)$ 服从为非中心化的卡方分布，先考虑 $\mu = 0$，即中心化卡方分布

的情形，注意到卡方分布的 p.d.f 的 kernel 为 $x^{n/2-1}\exp -x/2$，由伽马函数的性质可知，只要 $n>2$，那么 $E[1/S(X)]$ 就是存在的，所以就可以对它进行 taylor 展开然后估计。(**注意，这也说明了，为什么 (c) 的方差是无穷大，但是 (a)(b) 的方差不会是无穷大!!!**)
所以此时就可以套用定理 1.9 了，该怎么用呢? 1/x 的四阶导数不是无界吗? 那怎么办呢? 注意在本讲义的 Section 1.3 里面已经讨论这个问题了，先假设 $\mathbb{P}(S(X)>\epsilon)=1-\delta$，那么在概率 $1-\delta$ 下，可以使用定理 1.9，所以

$$E(\frac{1}{S(X)/n})=\frac{1}{EX_1^2}+\frac{Var(X_1^2)}{n(EX_1^2)^3}+O(\frac{1}{n^2}) \implies$$
$$E(\frac{1}{S(X)})=\frac{1}{n(\mu^2+\tau^2)}+\frac{4\mu^2\tau^2+2\tau^4}{n^2(\mu^2+\tau^2)^3}+O(\frac{1}{n^3})$$

所以在概率 $1-\delta$ 下，

$$Var(Z)=Var(W)=\sigma^2\left[\frac{1}{n(\mu^2+\tau^2)}+\frac{4\mu^2\tau^2+2\tau^4}{n^2(\mu^2+\tau^2)^3}\right]+O(\frac{1}{n^3})$$

另外，再注意到由于 $S(X)$ 是卡方分布，所以我们的假设 $\mathbb{P}(S(X)>\epsilon)=1-\delta$ 中，δ 可以任意接近 0（相应的 ϵ 也会比较接近 0，但是始终是固定的正数）. 因此我们严格的证明了下述结论：
Proposition: 对于任意的 $\delta>0$，在概率 $1-\delta$ 下，下述估计成立：

$$Var(Z)=\sigma^2\left[\frac{1}{n(\mu^2+\tau^2)}+\frac{4\mu^2\tau^2+2\tau^4}{n^2(\mu^2+\tau^2)^3}\right]+O(\frac{1}{n^3}).$$

注： 表述“对于任意的 $\delta>0$，在概率 $1-\delta$ 下”与在"Almost surely,..." 意义下，前者要比后者“弱”那么一点点。但是都是表述“一个命题很大概率意义上成立”的表述方法.

解法 3

也有用 CLT 方法、考虑极限分布来估计的，但是这个显然没有上面的估计精确，而且也不太严谨，所以这里就不写了。

[0]3.1

设 $X_1,\cdots,X_n$ i.i.d. $\sim N(a,\sigma^2)$, 样本方差 $S^2=\sum_{i=1}^n\left(X_i-\bar{X}\right)^2/(n-1)$ 是 σ^2 的无偏估计. 证明：S^2 是 σ^2 的强相合估计和均方相合估计，并讨

论其渐近正态性.

解法一：

这道题完全可以放进第一次作业里面，强相合性、渐进正态性，在**第一次作业解答及补充.pdf** 里面已经写过证明了，此处不再赘述.
至于均方相合性，注意到 S^2 是无偏的，所以要证 MSE 收敛，只需要算方差就行了。
样本方差的方差计算是一个经典的繁琐计算，具体过程可参考一篇知乎帖子：https://zhuanlan.zhihu.com/p/268376613.
$E(S-\sigma)^2 = \mathrm{Var}\,(S^2) = \frac{\nu_4}{n} - \frac{n-3}{n(n-1)}\sigma^4 = \frac{\sigma_4}{n}(3 - \frac{n-3}{n-1}) \to 0$，当$n \to \infty$.
值得注意是，这里用到了正态的四阶中心矩是方差的平方的 **3 倍** 这一条性质，假如不是 3 倍而是 1 倍，那么就不一定有均方收敛性了。

解法二：

由于学过矩估计的强相合性与渐进正态性，而 $S^2 = \frac{n}{n-1}(\hat{\alpha_2} - \hat{\alpha_1}^2)$ 是样本一阶矩与二阶矩的函数，并且也是总体方差的无偏估计，所以由第二节 PPT 第 6、7 页的性质，则可以得知 S^2 的强相合性与渐近正态性成立.
至于渐近方差的计算，用 $G\Sigma G^T$ 来算即可.
$\Sigma = (\alpha_{i+j} - \alpha_i\alpha_j)_{2\times 2}$
注意到，当我们用 $(X_i - a)$ 代替 X_i 时，S^2 的形式完全不变，故不妨设 $a = 0$，因此上面的 α_i 可以看成中心矩.
而梯度 $G = (0, 1)$，当 $n \to \infty$ 时.
$G\Sigma G^T = \mu_4 - \mu_2^2 = 2\sigma^4$.

评注：

总体来说，这两种解法的本质是相同的，只是用课堂上讲过的定理可以简化一些证明与计算.
批改反馈：有些同学没有利用这里的平移技巧，所以算出来的方差很大一堆没有化简，实际上是可以用中心矩化简的。