第 10 章

渐近评价

“我知道，我亲爱的华生，我们对稀奇古怪的异乎寻常的东西有着共同的爱好，而对日常生活中的那些流俗和单调乏味的老一套毫无兴趣.”

——夏洛克·福尔摩斯

《红发会》

迄今，我们考虑的一直是在有限样本条件下的情形. 与之对照，我们可以考虑渐近性质，这些性质描述的是当样本量变成无穷时的一个方法. 在这一章我们将着眼于某些渐近性质，并分别考虑点估计，假设检验和区间估计. 我们将特别强调极大似然方法的渐近性.

渐近评价的力量在于，当样本量变成无穷时，计算简化了. 在有限样本情形不可能做的评价变成了常规. 这种简化还允许我们去检查其他的技术（例如自助法和M-估计），这些技术的典型特点是对其只能做渐近评价.

令样本量无限制增加（有时称其为“asymptopia”）不应仅仅被嘲笑为一个想像的练习. 反之，渐近性揭露出一个方法最基本的性质，并且给予我们一个非常强有力而应用广泛的评价工具.

10.1 点估计

10.1.1 相合性

相合性似乎是一个相当基本的性质，它要求当样本量变成无穷时估计量收敛到正确的值. 相合性非常重要，一个非相合的估计量的价值是值得怀疑的（或者至少是值得严格考查的）.

虽然人们常讲相合估计量，但相合性（也包括所有的渐近性质）关注的是一个估计量序列，而非一个单独的估计量. 如果我们按照一个分布 $f(x|\theta)$ 去观测 X_1，X_2，…，我们只要通过对于每个样本量 n 执行相同的估计过程，就可以构造出一个估计量序列 $W_n=W_n(X_1，\cdots，X_n)$. 例如，$\overline{X}_1=X_1$，$\overline{X}_2=(X_1+X_2)/2$，$\overline{X}_3=(X_1+X_2+X_3)/3$，…. 我们现在可以给相合序列下定义.

定义 10.1.1 一个估计量序列 $W_n=W_n(X_1,\cdots,X_n)$ 是参数 θ 的一个相合估计量序列（consistent sequence of estimators），如果对于每个$\epsilon>0$ 和每个 $\theta\in\Theta$，

$$\lim_{n\to\infty}P_\theta(|W_n-\theta|<\epsilon)=1. \tag{10.1.1}$$

通俗地讲，方程（10.1.1）就是说当样本量变成无穷（而且样本信息变得越来越好）时，估计量将以高概率任意接近于参数 θ，这是人们特别渴望的一个性质. 或者把事情转换一下，我们可以说一个相合估计量序列未能达到真实参数的概率很小. 方程（10.1.1）的一个等价说法是，一个相合序列 W_n 将满足：对于每个$\epsilon>0$ 和每个 $\theta\in\Theta$，

$$\lim_{n\to\infty}P_\theta(|W_n-\theta|\geqslant\epsilon)=0. \tag{10.1.2}$$

定义 10.1.1 可以和定义 5.5.1 即依概率收敛的定义，来比较. 定义 10.1.1 讲的就是一个相合估计量序列依概率收敛于被估计的参数 θ. 定义 5.5.1 是以一个概率结构来处理一个随机变量序列，而定义 10.1.1 是以用 θ 为指标的一整族概率结构做处理. 对于每个不同的 θ 值，与序列 W_n 关联的概率结构是不同的. 而且该定义说对于各个 θ 值，相应概率结构都使得序列依概率收敛到真实 θ. 这就是一个概率的定义与一个统计学定义通常的区别. 概率的定义处理一个概率结构，而统计学定义处理一整族.

例 10.1.2（$\overline{X}$的相合性） 设 X_1，X_2，…是 iid n（θ，1）的，来考虑序列

$$\overline{X}_n=\frac{1}{n}\sum_{i=1}^{n}X_i.$$

回忆有$\overline{X}_n\sim\mathrm{n}(\theta,1/n)$，所以

$$\begin{aligned}
P_\theta(|\overline{X}_n-\theta|<\epsilon)&=\int_{\theta-\epsilon}^{\theta+\epsilon}\left(\frac{n}{2\pi}\right)^{\frac{1}{2}}\mathrm{e}^{-(n/2)(\overline{x}_n-\theta)^2}\,\mathrm{d}\,\overline{x}_n && \text{（根据定义）}\\
&=\int_{-\epsilon}^{\epsilon}\left(\frac{n}{2\pi}\right)^{\frac{1}{2}}\mathrm{e}^{-(n/2)y^2}\,\mathrm{d}y && \text{（变量替换 } y=\overline{x}_n-\theta\text{）}\\
&=\int_{-\epsilon\sqrt{n}}^{\epsilon\sqrt{n}}\left(\frac{1}{2\pi}\right)^{\frac{1}{2}}\mathrm{e}^{-(1/2)t^2}\,\mathrm{d}t && \text{（变量替换 } t=y\sqrt{n}\text{）}\\
&=P(-\epsilon\sqrt{n}<Z<\epsilon\sqrt{n}) && (Z\sim\mathrm{n}(0,1))\\
&\to 1 \qquad \text{当 } n\to\infty
\end{aligned}$$

因此$\overline{X}_n$ 是θ 的一个相合估计量序列. ‖

一般情况下，为验证相合性不必做像上面那样详细的计算. 对于一个估计量 W_n，回想一下 Chebychev 不等式的陈述

$$P_\theta(|W_n-\theta|\geqslant\epsilon)\leqslant\frac{\mathrm{E}_\theta[(W_n-\theta)^2]}{\epsilon^2},$$

这样，如果对于每个 $\theta\in\Theta$ 有

$$\lim_{n\to\infty}\mathrm{E}_\theta[(W_n-\theta)^2]=0,$$

则这个估计量序列就是相合的. 此外，根据式（7.3.1）就有

$$\mathrm{E}_\theta[(W_n-\theta)^2]=\mathrm{Var}_\theta W_n+[\mathrm{Bias}_\theta W_n]^2. \tag{10.1.3}$$

把其总括在一起，我们可以给出以下的定理.

定理 10.1.3 如果 W_n 是参数 θ 的一个估计量序列，它对于每个 $\theta\in\Theta$ 都满足

(i) $\lim\limits_{n\to\infty}\mathrm{Var}_\theta W_n=0$,

(ii) $\lim\limits_{n\to\infty}\mathrm{Bias}_\theta W_n=0$,

则 W_n 是参数 θ 的一个相合估计量序列.

例 10.1.4（例 10.1.2 续） 因为

$$\mathrm{E}_\theta\,\overline{X}_n=\theta \text{ 和 } \mathrm{Var}_\theta\,\overline{X}_n=\frac{1}{n},$$

定理 10.13 的条件得到满足，从而序列 $\overline{X}_n$ 是相合的. 此外，根据定理 5.2.6，如果有一个来自均值是 θ 的任何总体的 iid 样本，只要总体具有有限的方差，则 $\overline{X}_n$ 对于 θ 是相合的. ‖

在本节开始，我们谈到一个非相合估计量序列时认为它的价值应该受到质疑. 这样评论的部分根据就是如以下定理所揭示的，存在如此之多的相合序列. 这个定理的证明留给习题 10.2.

定理 10.1.5 设 W_n 是参数 θ 的一个相合估计量序列. 设 a_1，a_2，…和 b_1，b_2，…是常数序列，满足

(i) $\lim\limits_{n\to\infty}a_n=1$,

(ii) $\lim\limits_{n\to\infty}b_n=0$.

则序列 $U_n=a_nW_n+b_n$ 是参数 θ 的一个相合估计量序列.

作为本节的结束，我们概要给出关于极大似然估计量相合性的一个更一般的结果. 这个结果表明极大似然估计量是其参数的相合估计量，而且是我们见到的求估计量的方法能保证一种最优性质的第一例.

为了使 MLE 具有相合性，其基础密度（似然函数）必须满足一定的“正则性条件”，这里我们将不深入探究，不过可以参见 10.6 节 10.6.2 了解详情.

定理 10.1.6（MLE 的相合性） 设 X_1，X_2，…是 iid $f(x|\theta)$ 的，$L(\theta|\boldsymbol{x})=\prod\limits_{i=1}^{n}f(x_i|\theta)$ 是似然函数，而 $\hat{\theta}$ 表示 θ 的 MLE. 设 $\tau(\theta)$ 是 θ 的一个连续函数. 那么在杂录 10.6.2 的关于 $f(x|\theta)$，从而也就是对 $L(\theta|\boldsymbol{x})$ 的正则性条件之下，对于每个 $\epsilon>0$ 和每个 $\theta\in\Theta$，有

$$\lim_{n\to\infty}P_\theta(|\tau(\hat{\theta})-\tau(\theta)|\geqslant\epsilon)=0.$$

这就是说，此 $\tau(\hat{\theta})$ 是 $\tau(\theta)$ 的一个相合估计量.

证明：定理的证明从证对于每个 $\theta\in\Theta$，$\frac{1}{n}\log L(\hat{\theta}|\boldsymbol{x})$ 几乎必然收敛于

$E_\theta(\log f(X|\theta))$下手. 在关于 $f(x|\theta)$的某些条件之下，这蕴涵 $\hat{\theta}$ 依概率收敛于θ，因此 $\tau(\hat{\theta})$ 依概率收敛于 $\tau(\theta)$. 细节请看 Stuart，Ord and Arnold（1999，第 18 章）. ‖

10.1.2 有效性

相合性考虑的是一个估计量的渐近精确性：它是收敛到要估计的参数吗？本节我们来看一个有关的性质，即有效性，这个性质关心的是一个估计量的渐近方差.

在计算渐近方差时，或许我们想如下进行. 给出一个基于样本量 n 的估计量 T_n，我们计算有限样本方差 $\mathrm{Var}T_n$，然后计算$\lim_{n\to\infty} k_n \mathrm{Var}T_n$，其中 k_n 是某规格化常数.（注：在很多情况中，当 $n\to\infty$时，$\mathrm{Var}T_n\to 0$，所以我们需要一个因子 k_n 以迫使它趋向一个极限.）

定义 10.1.7 对于一个估计量 T_n，如果$\lim_{n\to\infty} k_n \mathrm{Var}T_n=\tau^2<\infty$，其中$\{k_n\}$是一个常数序列，则 τ^2 叫做极限方差（limiting variance）或方差的极限.

例 10.1.8（极限方差） 关于 n 个 iid 的具有 $\mathrm{E}(X)=\mu$ 和 $\mathrm{Var}(X)=\sigma^2$ 的正态观测的平均值$\overline{X}_n$，如果我们取 $T_n=\overline{X}_n$，则$\lim_{n\to\infty} n\mathrm{Var}\,\overline{X}_n=\sigma^2$ 是 T_n 的极限方差.

但是，若我们要用 $1/\overline{X}_n$ 估计 $1/\mu$，麻烦事就发生了. 如果我们现在取 $T_n=1/\overline{X}_n$，就发现方差 $\mathrm{Var}T_n=\infty$，那么方差的极限是无穷. 然而回忆例 5.5.23，那里我们讲过 $1/\overline{X}_n$ 的近似均值和方差是

$$E\left(\frac{1}{\overline{X}_n}\right)\approx\frac{1}{\mu},$$

$$\mathrm{Var}\left(\frac{1}{\overline{X}_n}\right)\approx\left(\frac{1}{\mu}\right)^4\mathrm{Var}\,\overline{X}_n,$$

于是根据这第二次的计算，方差是 $\mathrm{Var}T_n\approx\frac{\sigma^2}{n\mu^4}<\infty$. ‖

这个例子指出了把方差的极限用作大样本量度时的问题. 当然，精确有限样本 $1/\overline{X}$的方差是∞. 但是如果 $\mu\neq 0$，则 $1/\overline{X}$取非常大值的区域会具有趋向 0 的概率. 所以例 10.1.8 中第二种近似是很现实的（同时更是有用的）. 我们采用的就是这第二种计算大样本方差的方法.

定义 10.1.9 对于一个估计量 T_n，假定有依分布收敛 $k_n(T_n-\tau(\theta))\to\mathrm{n}(0,\sigma^2)$，则参数 σ^2 叫做 T_n 的渐近方差或 T_n 的极限分布的方差.

对于计算样本均值或者其他类型平均的方差，典型情况是极限方差和渐近方差有相同值. 但是在更复杂的情况，有时极限方差将会令我们失望. 注意总是有这样的情况，渐近方差小于极限方差，这也是有趣的（见 Lehmann and Casella 1998，6.1 节）. 这里举个例子来说明.

例 10.1.10（大样本混合方差） 分层模型

$$Y_n \mid W_n = w_n \sim \mathrm{n}(0,\ w_n + (1-w_n)\sigma_n^2),$$
$$W_n \sim \mathrm{Bernoulli}(p_n),$$

可以展示出渐近方差和极限方差间的重大差别.（这个模型有时也描述为一个混合模型，在其中我们以 p_n 的概率观测 $Y_n \sim \mathrm{n}(0,\ 1)$ 而以 $1-p_n$ 的概率观测 $Y_n \sim \mathrm{n}(0,\ \sigma_n^2)$.）

首先利用定理 4.4.7，我们就有

$$\mathrm{Var}(Y_n) = p_n + (1-p_n)\sigma_n^2.$$

由此就得到，只有在 $\lim_{n\to\infty}(1-p_n)\sigma_n^2 < \infty$ 时，Y_n 的极限方差有限.

另一方面，Y_n 的渐近方差可以利用

$$P(Y_n < a) = p_n P(Z < a) + (1-p_n)P(Z < a/\sigma_n).$$

直接计算出来.

假定现在我们让 $p_n \to 1$ 及 $\sigma_n \to \infty$，并且使得 $(1-p_n)\sigma_n^2 \to \infty$. 这样就得到 $P(Y_n < a) \to P(Z < a)$，就是说 $Y_n \to \mathrm{n}(0,\ 1)$，并且我们有

$$\text{极限方差} = \lim_{n\to\infty} p_n + (1-p_n)\sigma_n^2 = \infty$$
$$\text{渐近方差} = 1$$

更多的细节参见习题 10.6. ‖

根据 Cramér-Rao 下界（定理 7.3.9）的精神，是存在一个最佳渐近方差的.

定义 10.1.11 一个估计量序列 W_n 关于一个参数 $\tau(\theta)$ 是渐近有效的，如果 $\sqrt{n}[W_n - \tau(\theta)] \xrightarrow{L} \mathrm{n}(0,\ \upsilon(\theta))$ 而且

$$\upsilon(\theta) = \frac{[\tau'(\theta)]^2}{\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X\mid\theta)\right)^2\right)},$$

就是说，W_n 的渐近方差达到了 Cramér-Rao 下界.

回忆定理 10.1.6 所讲，在一般的条件之下，MLE 是相合的. 在某些更强的正则性条件之下，关于渐近有效的同样类型定理也成立，一般我们可以把 MLE 看成是相合且渐近有效的. 关于这些正则性条件的细节也在杂录 10.6.2 中.

定理 10.1.12（MLE 的渐近有效性） 设 X_1，X_2，…是 iid $f(x\mid\theta)$，θ 的 MLE 记作 $\hat{\theta}$，设 $\tau(\theta)$ 是 θ 的一个连续函数. 那么在 10.6 节 10.6.2 的关于 $f(x\mid\theta)$，从而也就是对 $L(\theta\mid\boldsymbol{x})$ 的正则性条件之下，

$$\sqrt{n}[\tau(\hat{\theta}) - \tau(\theta)] \xrightarrow{L} \mathrm{n}(0,\ \upsilon(\theta)),$$

其中 $\upsilon(\theta)$ 是 Cramér-Rao 下界. 就是说，$\tau(\hat{\theta})$ 是 $\tau(\theta)$ 的一个相合且渐近有效的估计量.

证明：此定理证明的有趣之处在于 Taylor 级数的利用以及它发掘出 MLE 是被

定义于似然函数的导数的零点这一事实. 我们将略述 $\hat{\theta}$ 为渐近有效的证明，对 $\tau(\hat{\theta})$ 的扩充留给习题 10.7.

回忆 $l(\theta \mid \boldsymbol{x}) = \sum \log f(x_i \mid \theta)$ 是对数似然函数. 把其导数（关于 θ 的）记作 l'，l''，…. 现在真值 θ_0 的周围展开对数似然的一阶导数，

$$l'(\theta|\boldsymbol{x})=l'(\theta_0|\boldsymbol{x})+(\theta-\theta_0)l''(\theta_0|\boldsymbol{x})+\cdots, \tag{10.1.4}$$

这里，我们将忽略其高阶项（在正则性条件下这个手法是正当的）.

现在用 $\hat{\theta}$ 替换 θ，并看到等式（10.1.4）的左边是 0. 重新整理此式并且乘以 $\sqrt{n}$，就给出

$$\sqrt{n}(\hat{\theta}-\theta_0)=\sqrt{n}\frac{-l'(\theta_0 \mid \boldsymbol{x})}{l''(\theta_0 \mid \boldsymbol{x})}=\frac{-\frac{1}{\sqrt{n}}l'(\theta_0 \mid \boldsymbol{x})}{\frac{1}{n}l''(\theta_0 \mid \boldsymbol{x})}. \tag{10.1.5}$$

如果我们用 $I(\theta_0)=E\left[l'(\theta_0 \mid \boldsymbol{X})\right]^2=1/v(\theta)$ 来记关于一个观测的信息数，应用中心极限定理和弱大数定律（细节见习题 10.8）就将证明出

$$\begin{aligned} -\frac{1}{\sqrt{n}}l'(\theta_0 \mid \boldsymbol{X}) &\xrightarrow{L} \mathrm{n}(0, I(\theta_0)), \\ \frac{1}{n}l''(\theta_0 \mid \boldsymbol{X}) &\xrightarrow{p} I(\theta_0). \end{aligned} \tag{10.1.6}$$

这样，如果我们设 $W \sim \mathrm{n}[0, I(\theta_0)]$，则 $\sqrt{n}(\hat{\theta}-\theta_0)$ 依分布收敛到 $W/I(\theta_0) \sim n[0, 1/I(\theta_0)]$，定理证毕. ‖

例 10.1.13（渐近正态与相合性） 以上定理表明 MLE 具有有效性和相合性是典型情况. 我们希望注意到这个说法是有点累赘的，因为有效性只被定义在估计量是渐近正态的，而正像我们将要阐明的，渐近正态蕴涵相合性. 设

$$\sqrt{n}\frac{W_n-\mu}{\sigma} \xrightarrow{L} Z,$$

其中 $Z \sim \mathrm{n}(0, 1)$. 通过运用 Slutsky 定理（定理 5.5.17），我们断定

$$W_n-\mu=\left(\frac{\sigma}{\sqrt{n}}\right)\left(\sqrt{n}\,\frac{W_n-\mu}{\sigma}\right)\to\lim_{n\to\infty}\left(\frac{\sigma}{\sqrt{n}}\right)Z=0,$$

所以 $W_n-\mu \xrightarrow{L} 0$. 根据定理 5.5.13 我们知道，依分布收敛到一个点等价于依概率收敛，所以 W_n 是 μ 的一个相合估计量. ‖

10.1.3 计算与比较

前面几节展示的渐近公式可以提供给我们用在大样本的近似方差. 当然，我们必须要考虑正则性条件（10.6 节 10.6.2），但是这些条件是相当一般的而且几乎在通常情况下总能得到满足. 不过，有一个条件应当特别提及，就像我们在例 7.3.13

已经见过的那样，违背了它就会导致混乱. 为了使下面的近似成为正当，概率密度函数或概率质量函数的支撑集，也就是似然函数的支撑集必须与参数无关.

如果一个 MLE 是渐近有效的，定理 10.1.6 中的渐近方差就是定理 5.5.24 中（去掉 $1/n$ 项）的 Δ 方法方差. 这样，我们就可以把 Cramér-Rao 下界用作 MLE 的真实方差的一个近似. 设 X_1，X_2，…是 iid $f(x|\theta)$的，$\hat{\theta}$ 是 θ 的 MLE，而 $I_n(\theta)=\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log L(\theta\mid \boldsymbol{X})\right)^2$ 是样本的信息数. 根据 Δ 方法与 MLE 的渐近有效性，$h(\hat{\theta})$的方差可以由以下来近似

$$\text{(10.1.7)}\quad \mathrm{Var}(h(\hat{\theta})\mid\theta)\approx\frac{[h'(\theta)]^2}{I_n(\theta)}$$

$$=\frac{[h'(\theta)]^2}{\mathrm{E}_\theta\left(-\frac{\partial^2}{\partial\theta^2}\log L(\theta|\boldsymbol{X})\right)} \qquad \text{（利用引理 7.3.11 的恒等式）}$$

$$\approx\frac{[h'(\theta)]^2|_{\theta=\hat{\theta}}}{-\frac{\partial^2}{\partial\theta^2}\log L(\theta|\boldsymbol{X})|_{\theta=\hat{\theta}}} \qquad \text{（分母是 } \hat{I}_n(\hat{\theta})\text{，即观测信息数）}$$

此外，已被证明（Efron and Hinkley 1978），使用观测信息数胜于使用出现在 Cramér—Rao 下界中的期望信息数.

注意方差估计的步骤是一个两步的过程，这个事实或多或少被式（10.1.7）掩盖了. 为估计 $\mathrm{Var}_\theta h(\hat{\theta})$，首先我们近似 $\mathrm{Var}_\theta h(\hat{\theta})$，然后再估计这个近似结果，而这通常是用 $\hat{\theta}$ 替换 θ. 作为结果的估计，可以记作 $\widehat{\mathrm{Var}}_\theta h(\hat{\theta})$ 或 $\widehat{\mathrm{Var}}_\theta h(\hat{\theta})$.

从定理 10.1.6 得出 $-\frac{1}{n}\frac{\partial^2}{\partial\theta^2}\log L(\theta\mid\boldsymbol{X})|_{\theta=\hat{\theta}}$ 是 $I(\theta)$ 的一个相合估计量，所以就得到 $\widehat{\mathrm{Var}}_\theta h(\hat{\theta})$ 是 $\mathrm{Var}_\theta h(\hat{\theta})$ 的一个相合估计量.

例 10.1.14（近似二项方差） 在例 7.2.7 中我们看到 $\hat{p}=\sum X_i/n$ 是 p 的 MLE，其中我们有来自一个 Bernoulli（p）总体的随机样本 X_1，…，X_n. 我们通过直接计算还知道

$$\mathrm{Var}_p\hat{p}=\frac{p(1-p)}{n},$$

并且 $\mathrm{Var}_p h(\hat{p})$的一个合理估计是

$$\text{(10.1.8)}\quad \widehat{\mathrm{Var}}_p\,\hat{p}=\frac{\hat{p}(1-\hat{p})}{n}.$$

如果我们把式（10.1.7）的近似式应用于 $h(p)=p$，就得到 $\mathrm{Var}_p\,\hat{p}$ 的一个估计

$$\widehat{\mathrm{Var}}_p\,\hat{p}\approx\frac{1}{-\frac{\partial^2}{\partial p^2}\log L(p|\boldsymbol{x})|_{p=\hat{p}}}.$$

因

$$\log L(p|\boldsymbol{x})=n\hat{p}\log(p)+n(1-\hat{p})\log(1-p),$$

于是就有

$$\frac{\partial^2}{\partial p^2}\log L(p|\boldsymbol{x})=-\frac{n\hat{p}}{p^2}-\frac{n(1-\hat{p})}{(1-p)^2}.$$

在 $p=\hat{p}$ 计算此二阶导数就有

$$\frac{\partial^2}{\partial p^2}\log L(p|\boldsymbol{x})\Big|_{p=\hat{p}}=-\frac{n\hat{p}}{\hat{p}^2}-\frac{n(1-\hat{p})}{(1-\hat{p})^2}=-\frac{n}{\hat{p}(1-\hat{p})}.$$

这就给出了一个方差的近似，它等于式（10.1.8）．我们现在运用定理 10.1.6 就可以断言 $\hat{p}$ 的渐近有效性，而且特别地就是

$$\sqrt{n}(\hat{p}-p)\xrightarrow{L}\mathrm{n}[0,p(1-p)].$$

如果我们再使用定理 5.5.17（Slutsky 定理），我们就可以断定

$$\sqrt{n}\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})}}\xrightarrow{L}\mathrm{n}[0,1].$$

估计 $\hat{p}$ 的方差事实上并不那样困难，而且也不必引入全部的这些近似步骤．但如果我们估计一个稍微复杂的函数，这些计算可能需要一些技巧．回忆在例 5.5.22 中我们曾用 Δ 方法近似胜率 $p/(1-p)$ 的一种估计量 $\hat{p}/(1-\hat{p})$ 的方差．现在我们看到，事实上这个估计量是胜率的 MLE，而且我们可以通过下面的方法估计它的方差：

$$\begin{aligned}\widehat{\mathrm{Var}}\left(\frac{\hat{p}}{1-\hat{p}}\right)&=\frac{\left[\frac{\partial}{\partial p}\left(\frac{p}{1-p}\right)\right]^2\Big|_{p=\hat{p}}}{-\frac{\partial^2}{\partial p^2}\log L(p|\boldsymbol{x})\Big|_{p=\hat{p}}}\\&=\frac{\left[\frac{(1-p)+p}{(1-p)^2}\right]^2\Big|_{p=\hat{p}}}{\frac{n}{p(1-p)}\Big|_{p=\hat{p}}}\\&=\frac{\hat{p}}{n(1-\hat{p})^3}.\end{aligned}$$

此外，我们还认识到这个估计量是渐近有效的． ‖

MLE 方差近似在很多情况下是成功的，但也并非一贯这么好．特别地，当函数 $h(\hat{\theta})$ 非单调的时候我们必须要小心．在这种情况下，导数 h' 将会有一个符号的改变，而这可能导致一个被低估的渐近方差．要认识到因为近似方法是基于 Cramér-Rao 下界的，所以就有可能低估．而非单调函数可使这个问题更坏．

例 10.1.15（例 10.1.14 续） 设现在我们要估计 Bernoulli 分布的方差 $p(1-p)$．这个方差的 MLE 由 $\hat{p}(1-\hat{p})$ 给出，这个估计量方差的估计可以通过应用式

(10.1.7) 的近似式获得. 我们有

$$\begin{aligned}\widehat{\mathrm{Var}}(\hat{p}(1-\hat{p})) &= \frac{\left[\frac{\partial}{\partial p}(p(1-p))\right]^2\Big|_{p=\hat{p}}}{-\frac{\partial^2}{\partial p^2}\log L(p|\boldsymbol{x})\Big|_{p=\hat{p}}} \\ &= \frac{(1-2p)^2\big|_{p=\hat{p}}}{\frac{n}{p(1-p)}\Big|_{p=\hat{p}}} \\ &= \frac{\hat{p}(1-\hat{p})(1-2\hat{p})^2}{n},\end{aligned}$$

其中如果$\hat{p}=\frac{1}{2}$就得到 0，这显然是一个对$\hat{p}(1-\hat{p})$之方差的低估. 函数 $p(1-p)$为非单调这个事实就是造成这个问题的一个原因.

使用定理 10.1.6，我们可以断定只要 $p\neq 1/2$，我们的估计量就是渐近有效的. 如果 $p=1/2$，我们就需要使用如定理 5.5.26 所给出的二阶近似（见习题 10.10）. ‖

渐近有效的性质给予我们一个在求渐近方差时希望达到的基准点（见 10.6 节 10.6.1）. 通过渐近相对效率（asymptotic relative efficiency）的概念，我们还能将渐近方差当作比较估计量的一个工具.

定义 10.1.16 如果两个估计量 W_n 和 V_n 满足

$$\sqrt{n}[W_n-\tau(\theta)] \xrightarrow{L} n[0,\sigma_W^2],$$

$$\sqrt{n}[V_n-\tau(\theta)] \xrightarrow{L} n[0,\sigma_V^2],$$

V_n 关于 W_n 的渐近相对效率（ARE）是

$$\mathrm{ARE}(V_n, W_n) = \frac{\sigma_W^2}{\sigma_V^2}$$

例 10.1.17（Poisson 估计量的 ARE） 设 X_1，X_2，…是 iid Poisson（λ）的，而我们对估计 0 概率感兴趣. 例如，在一个给定时间段内进入一家银行的顾客数有时用一个泊松随机变量来建模，而 0 概率就是在该时间段将没有一个顾客进入此银行的概率. 如果 $X\sim$Poisson（λ），则 $P(X=0)=\mathrm{e}^{-\lambda}$，而一个自然的（然而有些朴素的）估计量$\hat{\tau}$是

$$\hat{\tau}=\frac{1}{n}\sum_{i=1}^{n} Y_i.$$

其中 $Y_i=I(X_i=0)$. 这些 Y_i 服从分布 Bernoulli（$\mathrm{e}^{-\lambda}$），于是由此就可得到

$$\mathrm{E}(\hat{\tau})=\mathrm{e}^{-\lambda}\text{和 }\mathrm{Var}(\hat{\tau})=\frac{\mathrm{e}^{-\lambda}(1-\mathrm{e}^{-\lambda})}{n}.$$

另外一种方法，$\mathrm{e}^{-\lambda}$的 MLE 是 $\mathrm{e}^{-\hat{\lambda}}$，其中$\hat{\lambda}=\sum_i X_i/n$ 是λ 的 MLE. 使用 Δ 方

法近似，我们就有

$$\mathrm{E}(\mathrm{e}^{-\hat{\lambda}})\approx \mathrm{e}^{-\lambda} \text{和 } \mathrm{Var}(\mathrm{e}^{-\hat{\lambda}})\approx \frac{\lambda \mathrm{e}^{-2\lambda}}{n}.$$

因为

$$\sqrt{n}[\hat{\tau}-\mathrm{e}^{-\lambda}] \xrightarrow{L} n[0,\mathrm{e}^{-\lambda}(1-\mathrm{e}^{-\lambda})]$$

$$\sqrt{n}[\mathrm{e}^{-\hat{\lambda}}-\mathrm{e}^{-\lambda}] \xrightarrow{L} n[0,\lambda \mathrm{e}^{-2\lambda}],$$

所以$\hat{\tau}$关于 MLE $\mathrm{e}^{-\hat{\lambda}}$的渐近相对效率（ARE）是

$$\mathrm{ARE}(\hat{\tau},\mathrm{e}^{-\hat{\lambda}})=\frac{\lambda \mathrm{e}^{-2\lambda}}{\mathrm{e}^{-\lambda}(1-\mathrm{e}^{-\lambda})}=\frac{\lambda}{\mathrm{e}^{\lambda}-1}.$$

对这个函数的考察表明它是严格减的，在 $\lambda=0$ 达到最大值 1（这是$\hat{\tau}$所能达到的最好状态），而且随 $\lambda\to\infty$快速地逐渐变小（在 $\lambda=4$ 小于 10%）趋向渐近线 0.（参见习题 10.9） ‖

因为在典型的情况下 MLE 是渐近有效的，所以不可能期望另外的估计量有比它更小的渐近方差. 然而其他估计量也许具有其他的令人满意的性质（易于计算，对基础假定稳健）使其合乎愿望. 在这种情形下，如果我们采用别的估计量，那么 MLE 的有效性对于校准我们将会放弃什么就变得很重要.

我们来看最后一个例，其中要得到最佳方差计算可没那么容易. 下一节将讨论稳健性.

例 10.1.18（估计一个伽玛分布均值） 眼见为实，伽玛分布均值的估计不是一件容易的任务. 回忆伽玛分布的概率密度函数 $f(x|\alpha,\beta)$由下式给出

$$f(x|\alpha,\beta)=\frac{1}{\Gamma(\alpha)\beta^{\alpha}}x^{\alpha-1}\mathrm{e}^{-x/\beta}.$$

这个分布的均值是 $\alpha\beta$，为计算极大似然估计量我们必须处理 Γ 函数的导数（称为双 Γ 函数 digamma function），这种计算很困难. 与之对比，矩法给予我们一个易于计算的估计.

具体来说，设我们有来自上面伽玛密度的一随机样本 X_1，…，X_n，此密度中参数已经改设，把均值作为显参数，记为 $\mu=\alpha\beta$，从而密度成为

$$f(x\mid\mu,\beta)=\frac{1}{\Gamma(\mu/\beta)\beta^{\mu/\beta}}x^{\mu/\beta-1}\mathrm{e}^{-x/\beta},$$

而且 μ 的矩估计量是$\overline{X}$，具有方差 $\beta\mu/n$.

为计算 MLE，我们利用对数似然

$$l(\mu,\beta|\boldsymbol{x})=\sum_{i=1}^{n}\log f(x_i|\mu,\beta).$$

为了计算简便，假定 β 已知，于是我们求解$\frac{\mathrm{d}}{\mathrm{d}\mu}l(\mu,\beta|\boldsymbol{x})=0$ 以得到 MLE $\hat{\mu}$. 不存在

显式解，所以我们用数值方法.

根据定理 10.1.6 我们知道$\hat{\mu}$是渐近有效的. 让我们有兴趣的问题是使用易于计算的矩估计量我们会有多少损失. 为了做比较，我们计算渐近相对效率

$$\mathrm{ARE}(\hat{\mu},\overline{X})=[\beta\mu]\mathrm{E}\left[-\frac{\mathrm{d}^2}{\mathrm{d}\mu^2}l(\mu,\beta|\boldsymbol{X})\right],$$

并且选择了几个β值把它们的图显示在图 10.1.1 中. 当然，我们知道此 ARE 必定大于 1，但是从图我们看到对于较大的β值，进行较复杂计算和使用 MLE 是值得的. (参见习题 10.11 做的扩充，计算细节参见例 A.0.7.) ‖

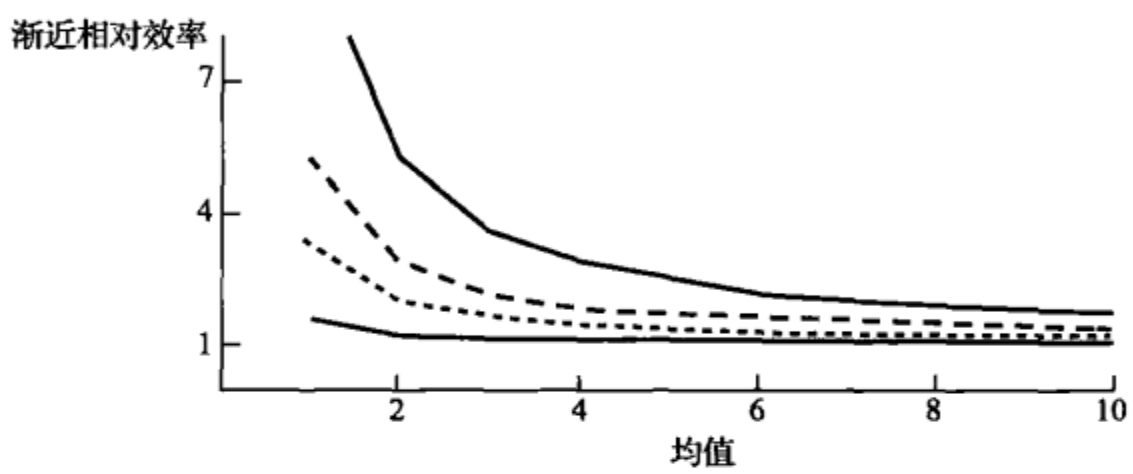

图 10.1.1 Γ均值矩法估计量对 MLE 估计量的渐近相对效率. 四条曲线相应于尺度参数的四个值（1，3，5，10），其中较高的曲线相应于较大的尺度参数的值

10.1.4 自助法标准误差

我们在例 1.2.20 中曾首次看到自助法（bootstrap）. 自助法提供了计算标准误差的一个替代方法.（它还能够提供更多，参见 10.6 节 10.6.3）

自助法建立在一个简单然而强有力的思想上（它可能涉及相当多的数学）[⊖]. 在统计学上，我们通过提取样本获悉总体特征. 因为样本代表总体，相似的样本特征就应该给予我们关于总体特征的信息. 自助法通过提取再抽样样本（resample）帮助我们了解样本特征（即我们从原始样本中再提取样本），并且利用这些信息去推断总体. 自助法是由 Efron 在 1970 年代后期发展起来的，其最初的想法出现在 Efron（1970a，b），并且有 Efron（1982）的专著. 另外参看 Efron（1998）有关近期更多的想法和发展.

让我们首先看一个简单的，实际上并不需要自助法的例子.

例 10.1.19（一个方差的自助法） 在例 1.2.20 中我们计算过选自

2，4，9，12

的所有可能的四个数的平均，那里我们是有放回的取数. 这是最简单的自助法形

⊖ 参见 Lehmann（1999，6.5 节）.

式，有时叫做非参数自助法. 图 1.2.2 在一个直方图中显示了这些值.

我们所建立的是样本均值可能值的再抽样. 我们看到有$\binom{4+4-1}{4}=35$个不同的可能值，但是这些值不是等概率的（这样，就不能视为随机样本）. 而$4^4=256$个（非不同的）再抽样样本都是等可能的，所以它们可以被视为随机样本. 对于第i个再抽样样本，我们设$\overline{x}_i^*$是该再抽样样本的平均数. 则我们可以用下式估计样本均值$\overline{X}$的方差

$$\mathrm{Var}^*(\overline{X})=\frac{1}{n^n-1}\sum_{i=1}^{n^n}(\overline{x}_i^*-\overline{\overline{x}}^*)^2, \tag{10.1.9}$$

其中$\overline{\overline{x}}^*=\frac{1}{n^n-1}\sum_{i=1}^{n^n}\overline{x}_i^*$为再抽样样本的平均数. （这里一律用*表示自助的，或者说再抽样的值）

在我们这个例中，自助法均值和方差为$\overline{\overline{x}}^*=6.75$和$\mathrm{Var}^*(\overline{X})=3.94$. 结果表明，就所关注的均值和方差而言，自助法估计与通常估计几乎是相同的（见习题 10.13）. ‖

现在我们已经看到怎样计算一个自助法标准误差，但在上述问题中并不真正需要这样做. 然而自助法的真正优越性，就像 Δ 方法，公式（10.1.9）几乎可以应用于任何的估计量. 这样，对任何的估计量$\hat{\theta}(\boldsymbol{x})=\hat{\theta}$，我们可以写成

$$\mathrm{Var}^*(\hat{\theta})=\frac{1}{n^n-1}\sum_{i=1}^{n^n}(\hat{\theta}_i^*-\overline{\hat{\theta}}^*)^2, \tag{10.1.10}$$

其中$\hat{\theta}_i^*$是从第i个再抽样样本计算出的估计量，而$\overline{\hat{\theta}}^*=\frac{1}{n^n-1}\sum_{i=1}^{n^n}\hat{\theta}_i^*$是再抽样估计值的平均数.

例 10.1.20（自助法　二项方差） 在例 10.1.15 中，我们使用 Δ 方法估计$\hat{p}(1-\hat{p})$的方差. 基于一组容量为n的样本，对于这个方差的估计我们可以把它替换为

$$\mathrm{Var}^*(\hat{p}(1-\hat{p}))=\frac{1}{n^n-1}\sum_{i=1}^{n^n}(\hat{p}(1-\hat{p})_i^*-\overline{\hat{p}(1-\hat{p})}^*)^2. \quad \|$$

但是现在出现一个问题. 对于我们的例 10.1.19，$n=4$，在自助法求和中有 256 项. 在更典型的样本量情况，这个项数增加得如此之大以至无法计算. （作者确信，当$n>15$，列举出全部再抽样样本几乎就是不可能的.）但是现在记住我们是统计学家，我们抽取再抽样样本的一组样本.

这样，对于样本$\boldsymbol{x}=(x_1, x_2, \cdots, x_n)$和一个估计量$\hat{\theta}(x_1, x_2, \cdots, x_n)=\hat{\theta}$，取$B$个再抽样样本［或叫自助样本（bootstrap samples）］并计算

$$\text{Var}_B^*(\hat{\theta}) = \frac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}_i^* - \bar{\hat{\theta}}^*)^2. \tag{10.1.11}$$

例 10.1.21（例 10.1.20 结论） 对于一组容量为 $n=24$ 的样本，我们计算 $\hat{p}(1-\hat{p})$ 的 Δ 方法方差估计以及对 $B=1000$ 的自助法方差估计. 对于 $\hat{p}\neq 1/2$，我们用例 10.1.5 的一阶 Δ 方法方差，而对于 $\hat{p}=1/2$，我们用定理 5.5.26 的二阶方差估计（见习题 10.16）. 在表 10.1.1 中我们看到，在所有情况下自助法方差估计值都更接近方差真值，而 Δ 方法方差估计是低估的. （这是不应奇怪的，根据式 (10.1.7)，表明 Δ 方法方差估计是基于一个下界的）

表 10.1.1 $\hat{p}(1-\hat{p})$ 的自助法和 Δ 方法方差. 当 $\hat{p}=1/2$ 的时候使用二阶 Δ 方法（参见定理 5.5.26）. 方差的真值是假定 $\hat{p}=p$ 用数值计算得出.

	$\hat{p}=1/4$	$\hat{p}=1/2$	$\hat{p}=2/3$
自助法	0.00508	0.00555	0.00561
Δ 方法	0.00195	0.00022	0.00102
真值	0.00484	0.00531	0.00519

Δ 方法是一个“一阶”近似，它基于 Taylor 级数展开的第一项. 当该项等于 0 的时候（如 $\hat{p}=1/2$ 时一样），我们必须使用二阶 Δ 方法. 相反，自助法经常能够有“二阶”准确度，获得比展开式第一项更准确的结果（见 10.6 节 10.6.3）. 因而此处当 $\hat{p}=1/2$ 时自助法自动起到校正作用.（注意 $24^{24}\approx 1.33\times 10^{13}$，这是一个巨大的数，所以穷举自助样本是不现实的） ‖

至此我们已经谈论过的自助法的类型叫做非参数自助法（nonparametric bootstrap），它对于总体的概率密度函数或概率质量函数没有函数形式的假定. 与之对照，我们也可以有参数自助法（parametric bootstrap）.

设我们有一组来自一个概率密度函数是 $f(x|\theta)$ 的样本 X_1，X_2，…，X_n，其中 θ 可以是一个参数向量. 于是我们可以用 MLE $\hat{\theta}$ 来估计 θ，然后抽取样本

$$X_1^*, X_2^*, \cdots, X_n^* \sim f(x|\hat{\theta}).$$

如果我们抽取了 B 组这样的样本，就可以利用式（10.1.11）估计 $\hat{\theta}$ 的方差. 注意，这些样本不是数据的再抽样样本，而实际上是抽自 $f(x\mid\hat{\theta})$ 的随机样本，这个分布有时叫做插件分布（plug-in distribution）.

例 10.1.22（参数自助法） 设我们有一组样本

$$-1.81,\ 0.63,\ 2.22,\ 2.41,\ 2.95,\ 4.16,\ 4.24,\ 4.53,\ 5.09$$

其 $\bar{x}=2.71$ 和 $s^2=4.82$. 如果我们假定基础分布是正态的，则参数自助法应抽取样本

$$X_1^*,\ X_2^*,\ \cdots,\ X_n^* \sim n(2.71,\ 4.82).$$

基于 $B=1000$ 组样本，我们算出 $\mathrm{Var}_B^*(S^2)=4.33$. 基于正态理论，$S^2$ 的方差是 $2(\sigma^2)^2/8$，我们可以用 MLE 估计它是 $2(4.82)^2/8=5.81$. 这些数据值实际上是从方差为 4 的正态分布生成出来的，所以 $\mathrm{Var}S^2=4.00$. 参数自助法在这里给出一个较好的估计值.（在例 5.6.6 中我们估计 S^2 的分布所使用的就是现在我们知道的参数自助法.） ‖

既然有了一个通用的计算标准误差的方法，我们如何知道它是一个好的方法呢？在例 10.1.21 中它似乎做得比 Δ 方法更好，而我们知道后者具有某些好性质. 特别地，我们知道 Δ 方法是基于极大似然估计，通常它将会产生相合估计量. 我们能够讲自助法也是这样吗？虽然我们不能对这个问题给出普遍回答，但是我们说，在很多情形下自助法的确提供给我们一个合理的相合估计量.

稍微确切一点地讲，我们把自助法估计量的计算分成两个不同部分.

a. 确立当 $B\to\infty$，(10.1.11) 收敛到 (10.1.10)，即

$$\mathrm{Var}_B^*(\hat{\theta})\xrightarrow{B\to\infty}\mathrm{Var}^*(\hat{\theta}).$$

b. 确立 (10.1.10) 的相合性，其中使用的是全部自助样本，即

$$\mathrm{Var}^*(\hat{\theta})\xrightarrow{n\to\infty}\mathrm{Var}(\hat{\theta}).$$

(a) 部分能够利用大数定律得到确立（习题 10.15）. 还要注意 (a) 部分全都发生在样本里.（Lehmann 1999，6.5 节把 $\mathrm{Var}_B^*(\hat{\theta})$ 叫做一个逼近量（approximator）而不是估计量.）

(b) 部分的确立需要稍许精细的处理，而这就是相合性确立之所在. 典型地，在 iid 抽样将获得相合性，但是在更一般情况它未必发生.（Lehmann 1999，6.5 节给出一个例子.）关于相合性的更详细的讨论（必然是在一个更高水平上的），参见 Shao and Tu（1995，3.3.2 节）或 Shao（1999，5.5.3 节）.

10.2 稳健性

到现在为止，我们在基础模型正确的假定下已经评价了估计量的性能. 在这个假定下，我们已经推导出在某种意义下的最佳估计量. 然而，如果基础模型不正确，则就不能保证我们的估计量最佳.

我们无法预防所有可能的情形，此外，如果我们的模型是经过仔细考虑做出的，则也不必这样预防. 但是我们关心对于假定的模型有小的或中等偏离的情形. 这样，这就引导我们去考虑稳健估计量（robust estimator）. 这种估计量放弃在假定模型上的最佳性，而换以当假定模型不是真实模型时的合理表现. 这样，我们就有一个平衡得失问题，最佳性或稳健性哪个标准更重要，这或许最好根据具体情况来决定.

术语“稳健性”可以有多种解释，但是也许 Huber（1981，1.2 节）概括得最好，他指出[⊖]：

任何统计方法应该具有如下特性：

（1）在假定模型下它应当具有一个合理的好（最佳或接近最佳）效率.

（2）应该在这样的意义下是稳健的：对于假定模型的微小偏离应该仅引起性能的轻微损伤

（3）对模型大一些的偏离也不应导致灾难性后果.

我们首先看一些易于理解这些条款的简单例子；然后再继续看更一般的估计量和稳健性的度量.

10.2.1 均值和中位数

样本均值是一个稳健的统计量吗？它也许严格地依赖于我们怎样表征稳健性的度量.

例 10.2.1（样本均值的稳健性） 设 X_1，X_2，…，X_n 是 iid $\mathrm{n}(\mu,\sigma^2)$的. 我们知道$\overline{X}$具有方差 $\mathrm{Var}(\overline{X})=\sigma^2/n$，而该方差是 Cramér-Rao 下界，这样在假定模型下它达到最好的方差因此$\overline{X}$满足（1）.

为研究（2），即研究对模型微小的偏离之下$\overline{X}$的性能，我们首先需要决定这意味着什么. 一种通常的解释就是使用一个 δ-污染模型（δ-contamination model），就是说，对于一个小 δ，假定我们的观测

$$X_i\sim\begin{cases}\mathrm{n}(\mu,\sigma^2) & \text{以 } 1-\delta \text{ 的概率}\\ f(x) & \text{以 } \delta \text{ 的概率}\end{cases},$$

其中 $f(x)$是某个其他分布.

设我们把 $f(x)$取为任何的一个均值是 θ 而方差是 τ^2 的密度. 则

$$\mathrm{Var}(\overline{X})=(1-\delta)\frac{\sigma^2}{n}+\delta\frac{\tau^2}{n}+\frac{\delta(1-\delta)(\theta-\mu)^2}{n}.$$

这看上去真是不错，因为如果 $\theta\approx\mu$ 且$\sigma\approx\tau$，$\overline{X}$将接近最佳. 然而，我们可以对模型做更大一点的扰动，并把事情搞得相当糟. 考虑如果 f（x）是一个 Cauchy 概率密度函数将会发生什么. 这时立即就得出 $\mathrm{Var}(\overline{X})=\infty$.（细节参见习题 10.18 而习题 10.19 讨论的是另一个情况.）‖

现在转向（3）. 我们问，如果出现一个常见程度的异常观测值会发生什么. 我们想像一个样本值的特别集合并考虑最大观测值增大的影响. 例如，设 $X_{(n)}=x$，其中 $x\to\infty$. 这样的一个观测值的影响可以认为是“灾难性的”. 虽然$\overline{X}$的分布性质没有受到影响，但是其观测值将是“无意义的.”这说明了崩溃值的概念，这个概

⊖ 特性的前二条.

念归于 Hample (1974).

定义 10.2.2 设 $X_{(1)}<\cdots<X_{(n)}$ 是容量为 n 的顺序样本，而设 T_n 是一个基于这个样本的统计量. T_n 具有崩溃值（breakdown value）b，$0\leqslant b\leqslant 1$，如果对于每一个 $\epsilon>0$，都有

$$\lim_{X_{(\{(1-b)n\})}\to\infty} T_n<\infty \text{和} \lim_{X_{(\{(1-(b+\epsilon))n\})}\to\infty} T_n=\infty.$$

(关于百分位数记号回忆定义 5.4.2.)

容易看出 $\overline{X}$ 的崩溃值是 0；就是说，如果这个样本中任何比例的样本值趋向无穷，则 $\overline{X}$ 的值也趋向无穷. 与此鲜明对照的是，样本中位数在样本值的这种变化下是不变的. 这种对于极端观测值的不敏感性有的时候被认为是样本中位数的一个优点，它的崩溃值为 50%. (关于崩溃值的更多内容参见习题 10.20.)

由于中位数在稳健性方面对于均值有改善，我们就可以问，转而使用一个更加稳健的估计量（当然我们必须这样做!）是否会失去什么. 例如在例 10.2.1 的简单正态模型中，如果模型正确，则样本均值是最优的无偏估计量. 因此就可以推出，对这个正态模型（以及它的附近），样本均值是一个良好的估计量. 但关键问题是对这个正态模型样本均值到底比中位数好多少? 如果我们能够回答这个问题，在做出使用哪个估计量以及侧重考虑哪个准则（最佳性或者稳健性）的选择时，就有了更丰富的信息. 为了在某种普遍意义下回答这个问题，我们号召用渐近相对效率准则.

为了计算中位数对均值的 ARE，首先必须建立中位数的渐近正态性并计算其渐近分布的方差.

例 10.2.3（中位数的渐近正态性） 为了求中位数的极限分布，我们采用一种类似于在定理 5.4.3 和 5.4.4 证明当中的论证，即基于二项分布的论证.

设 X_1，X_2，…，X_n 是来自具有概率密度函数为 f 和 cdf 为 F（设是可微的）的总体的样本，并且 $P(X_i\leqslant\mu)=1/2$，所以 μ 是总体中位数. 设 M_n 是样本中位数，并考虑对于某 a 计算

$$\lim_{n\to\infty} P(\sqrt{n}(M_n-\mu)\leqslant a)$$

如果我们通过

$$Y_i=\begin{cases}1 & \text{如果 } X_i\leqslant\mu+a/\sqrt{n}\\ 0 & \text{其他}\end{cases}$$

定义随机变量 Y_i，由此就可推出 Y_i 是一个成功概率为 $p_n=F(\mu+a/\sqrt{n})$ 的 Bernoulli 随机变量. 为避免复杂，我们将假定 n 是奇数，这样事件 $\{M_n\leqslant\mu+a/\sqrt{n}\}$ 就等价于事件 $\left\{\sum_i Y_i\geqslant(n+1)/2\right\}$.

经过一些代数计算就得到

$$P(\sqrt{n}(M_n-\mu)\leqslant a)=P\left(\frac{\sum_i Y_i-np_n}{\sqrt{np_n(1-p_n)}}\geqslant\frac{(n+1)/2-np_n}{\sqrt{np_n(1-p_n)}}\right).$$

现在 $p_n\to p=F(\mu)=1/2$，于是我们能应用中心极限定理证明 $\dfrac{\sum_i Y_i-np_n}{\sqrt{np_n(1-p_n)}}$ 依分布收敛到一个标准正态随机变量 Z. 简单的极限计算又证明有

$$\frac{(n+1)/2-np_n}{\sqrt{np_n(1-p_n)}}\to -2aF'(\mu)=-2af(\mu).$$

把这些都放在一起就得到

$$P(\sqrt{n}(M_n-\mu)\leqslant a)\to P(Z\geqslant -2af(\mu)),$$

而且因此 $\sqrt{n}(M_n-\mu)$ 的渐近分布是均值 0 为而方差为 $1/[2f(\mu)]^2$ 的正态分布. （关于细节，见习题 10.22，而严谨推导以及对于这个结果发展的更一般结果，参见 Shao，1999，5.3 节.） ‖

例 10.2.4（中位数对均值的渐近相对效率） 由于对于均值和中位数的渐近方差有简单的表达式，所以 ARE 是易于计算的. 下面的表给出相应三种对称分布的渐近相对效率. 我们发现，如所料当分布的尾部越重则得到的 ARE 越大. 这就是说，在重尾分布情况，用中位数性能会改善. 更多的比较见习题 10.23.

中位数/均值的渐近相对效率		
正态分布	罗吉斯蒂克分布	双指数分布
0.64	0.82	2

‖

10.2.2 M-估计量

我们所使用的很多统计量是最小化一个特别的准则的结果. 例如，如果 X_1，X_2，…，X_n 是来自 $f(x|\theta)$ 的，那么，可能的估计量有：样本均值，它是使 $\sum(x_i-a)^2$ 最小的量；样本中位数，它是使 $\sum|x_i-a|$ 最小的量；再就是 MLE，它是使 $\prod_{i=1}^n f(x_i|\theta)$ 最大（或者使负的对数似然最小）的量. 作为获得一个稳健估计量的系统方法，我们应当试图写下一个准则函数，它的最小值导致一个具有令人满意的稳健性质的估计量.

在试图定义一个稳健准则时，Huber（1964）曾考虑一种均值和中位数间的折中方案. 均值的准则是一个平方，它使之具有敏感性，但是在“尾部”平方对大的观测值给出太多的权重. 与之相反，中位数的绝对值准则不偏重大的或者小的观测值. 折中方案就是最小化准则函数

$$\sum_{i=1}^n \rho(x_i-a) \tag{10.2.1}$$

其中函数 ρ 是由

$$\rho(x)=\begin{cases}\frac{1}{2}x^2 & \text{若}|x|\leqslant k\\ k|x|-\frac{1}{2}k^2 & \text{若}|x|>k\end{cases}. \tag{10.2.2}$$

函数 $\rho(x)$ 的性态对于 $|x|\leqslant k$ 像 x^2 而对于 $|x|>k$ 像 $|x|$. 此外，因为 $\frac{1}{2}k^2=k|k|-\frac{1}{2}k^2$，所以这个函数连续（见习题 10.28）. 事实上 ρ 是可微的. 常数 k 可被称为一个调节参数，它控制着混合，对于较小的 k 值，则产生一个更“像中位数”的估计量.

表 10.2.1 Huber 估计量

k	0	1	2	3	4	5	6	8	10
估计值	−2.1	0.03	−0.04	0.29	0.41	0.52	0.87	0.97	1.33

例 10.2.5（Huber 估计量） 定义使（10.2.1）和（10.2.2）达最小的估计量叫做 Huber 估计量. 为了解这个估计量怎样工作以及 k 的选择如何重要，考虑以下含有 8 个标准正态偏离值与 3 个“离群值”的数据集合：

$\boldsymbol{x}=-1.28,\ -0.96,\ -0.46,\ -0.44,\ -0.26,\ -0.21,\ -0.063,\ 0.39,\ 3,\ 6,\ 9$

对于这些数据，均值是 1.33 而中位数是 −0.21. 当 k 变化时，我们得到一列 Huber 估计的值，列在表 10.2.1 中. 我们看到，当 k 增大时 Huber 估计值在中位数与均值间变动，因而我们解释为随着 k 的增大，则对离群值的稳健性下降. ‖

最小化（10.2.2）的估计量是 Huber 所研究的估计量的一个特例. 对于一般的函数 ρ，我们把使 $\sum_i \rho(x_i-\theta)$ 达最小的估计量叫做一个 M-估计量（M-estimator），这个名字使我们联想起它们是极大似然类型的估计量. 注意到如果把 ρ 选成负的对数似然 $-l(\theta|x)$，则 M-估计量就是通常的 MLE. 但是更灵活地选择这个欲最小化的函数，可以推演出具有各种不同性质的估计量.

由于最小化一个函数的典型做法是通过解出其导数的零点（指我们能够求导数的时候）而进行，定义 $\psi=\rho'$，我们看到，M−估计就是

$$\sum_{i=1}^{n}\psi(x_i-\theta)=0 \tag{10.2.3}$$

的解. 把估计量刻画为一个方程的根对于获取估计量的性质是特别有用的，这是由于那些在极大似然估计量中使用过的论证方法能够扩展. 特别地，看 10.1.2 节，尤其是定理 10.1.12 的证明. 我们假定函数 $\rho(x)$ 是对称的，而它的导数 $\psi(x)$ 是单调增的（这保证（10.2.3）的根是唯一的最小点）. 于是，就像定理 10.1.12 的证明，我们写出 ψ 的 Taylor 展开式为

$$\sum_{i=1}^{n}\psi(x_i-\theta)=\sum_{i=1}^{n}\psi(x_i-\theta_0)+(\theta-\theta_0)\sum_{i=1}^{n}\psi'(x_i-\theta_0)+\cdots,$$

其中 θ_0 是真值，而且我们忽略高阶项. 设 $\hat{\theta}_M$ 是方程（10.2.3）的解并且用它替换 θ 就得到

$$0=\sum_{i=1}^{n}\psi(x_i-\theta_0)+(\hat{\theta}_M-\theta_0)\sum_{i=1}^{n}\psi'(x_i-\theta_0)+\cdots,$$

其中左侧为 0 是因为 $\hat{\theta}_M$ 是方程（10.2.3）的解. 现在，再次类似于定理 10.1.12 的证明，我们重排这些项，然后除以 $\sqrt{n}$，并且忽略余项就得到

$$\sqrt{n}(\hat{\theta}_M-\theta_0)=\frac{\frac{-1}{\sqrt{n}}\sum_{i=1}^{n}\psi(x_i-\theta_0)}{\frac{1}{n}\sum_{i=1}^{n}\psi'(x_i-\theta_0)}.$$

现在我们假定 θ_0 满足 $E_{\theta_0}\psi(X-\theta_0)=0$（这通常被当作 θ_0 的定义）. 于是就可得到

$$\frac{-1}{\sqrt{n}}\sum_{i=1}^{n}\psi(x_i-\theta_0)=\sqrt{n}\left[\frac{-1}{n}\sum_{i=1}^{n}\psi(x_i-\theta_0)\right]\xrightarrow{L}\mathrm{n}(0,E_{\theta_0}\psi(X-\theta_0)^2), \tag{10.2.4}$$

而且根据大数定律得到

$$\frac{1}{n}\sum_{i=1}^{n}\psi'(x_i-\theta_0)\xrightarrow{p}\mathrm{E}_{\theta_0}\psi'(X-\theta_0). \tag{10.2.5}$$

把这些放在一起，我们就得到

$$\sqrt{n}\ (\hat{\theta}_M-\theta_0)\to\mathrm{n}\left(0,\ \frac{\mathrm{E}_{\theta_0}\psi\ (X-\theta_0)^2}{[\mathrm{E}_{\theta_0}\psi'\ (X-\theta_0)]^2}\right). \tag{10.2.6}$$

例 10.2.6（Huber 估计量的极限分布） 设 X_1，X_2，…，X_n 是 iid 的，来自概率密度函数 $f(x-\theta)$，其中 f 关于 0 对称，则对于由式（10.2.2）给出的 ρ，我们有

$$\psi(x)=\begin{cases}x & \text{若}|x|\leqslant k\\ k & \text{若}x>k\\ -k & \text{若}x<-k\end{cases} \tag{10.2.7}$$

而且因此有

$$\begin{aligned}\mathrm{E}_\theta\psi(X-\theta)&=\int_{\theta-k}^{\theta+k}(x-\theta)f(x-\theta)\mathrm{d}x-\\ &\quad k\int_{-\infty}^{\theta-k}f(x-\theta)\mathrm{d}x+k\int_{\theta+k}^{\infty}f(x-\theta)\mathrm{d}x\\ &=\int_{-k}^{k}yf(y)\mathrm{d}y-k\int_{-\infty}^{-k}f(y)\mathrm{d}y+k\int_{k}^{\infty}f(y)\mathrm{d}y=0,\end{aligned} \tag{10.2.8}$$

其中我们做了替换 $y=x-\theta$. 几个积分相加得 0 是由于 f 的对称. 因此，Huber 估计量具有正确的均值（见习题 10.25）.

为了计算方差，我们需要 ψ' 的期望值. 虽然 ψ 是不可微的，但是越过不可微点 $(x=\pm k)$，ψ' 将为 0. 因此我们只需要处理对于 $|x|\leqslant k$ 的期望，从而我们有

$$\mathrm{E}_\theta\psi'(X-\theta)=\int_{\theta-k}^{\theta+k}f(x-\theta)\mathrm{d}x=P_0(|X|\leqslant k),$$

$$\begin{aligned}\mathrm{E}_\theta\psi(X-\theta)^2&=\int_{\theta-k}^{\theta+k}(x-\theta)^2f(x-\theta)\mathrm{d}x+k^2\int_{\theta+k}^{\infty}f(x-\theta)\mathrm{d}x+k^2\int_{-\infty}^{\theta-k}f(x-\theta)\mathrm{d}x\\&=\int_{-k}^{k}x^2f(x)\mathrm{d}x+2k^2\int_{k}^{\infty}f(x)\mathrm{d}x.\end{aligned}$$

这样，我们可以得出结论，Huber 统计量是渐近正态的，具有均值 θ 和渐近方差

$$\frac{\int_{-k}^{k}x^2f(x)\mathrm{d}x+2k^2P_0(|X|>k)}{[P_0(|X|\leqslant k)]^2}.$$

‖

如我们在例 10.2.4 中所做的那样，现在针对几种不同分布考察 Huber 统计量的 ARE.

例 10.2.7（Huber 估计量的 ARE） 因为 Huber 估计量在某种意义上是均值与中位数的折中，我们将考察它相对于这两个估计量的相对效率.

Huber 估计量的渐近相对效率，$k=1.5$

	正态分布	罗吉斯蒂克分布	双指数分布
与均值比较	0.96	1.08	1.37
与中位数比较	1.51	1.31	0.68

对于正态及罗吉斯蒂克分布，Huber 估计量的表现类似于均值而比中位数有改进. 对于双指数分布，Huber 估计量比均值有改进但是不如中位数好. 回忆均值是关于正态的 MLE，而中位数是关于双指数分布的 MLE（所以 ARE<1 是料想之中的）. 对于这些分布，Huber 估计量具有类似于 MLE 的性能，但是在其他情况似乎也保持合理性. ‖

我们看到 M—估计量是稳健性和效率的一个折衷. 现在我们更仔细地分析为了得到稳健性，在效率方面我们可能放弃了什么.

式（10.2.6）中的渐近方差的分母含有 $\mathrm{E}_{\theta_0}\psi'(X-\theta_0)$ 项，我们可以把它写成

$$\mathrm{E}_\theta\psi'(X-\theta)=\int\psi'(x-\theta)f(x-\theta)\mathrm{d}x=-\int\left[\frac{\partial}{\partial\theta}\psi(x-\theta)\right]f(x-\theta)\mathrm{d}x.$$

现在我们运用乘积微分法则就得到

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}\theta}\int\psi(x-\theta)f(x-\theta)\mathrm{d}x\\&=\int\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\psi(x-\theta)\right]f(x-\theta)\mathrm{d}x+\int\psi(x-\theta)\left[\frac{\mathrm{d}}{\mathrm{d}\theta}f(x-\theta)\right]\mathrm{d}x.\end{aligned}$$

因为 $\mathrm{E}_\theta\psi(X-\theta)=0$，所以上式的左侧是 0，于是我们有

$$-\int\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\psi(x-\theta)\right]f(x-\theta)\mathrm{d}x=\int\psi(x-\theta)\left[\frac{\mathrm{d}}{\mathrm{d}\theta}f(x-\theta)\right]\mathrm{d}x$$

$$=\int\psi(x-\theta)\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(x-\theta)\right]f(x-\theta)\mathrm{d}x,$$

这里我们用到$\frac{\mathrm{d}}{\mathrm{d}y}g(y)/g(y)=\frac{\mathrm{d}}{\mathrm{d}y}\log g(y)$这个事实. 最后一个表达式可以写成$\mathrm{E}_\theta[\psi(X-\theta)l'(\theta|X)]$，其中$l(\theta\mid x)$是对数似然函数，这样就产生出恒等式

$$\mathrm{E}_\theta\psi'(X-\theta)=-\mathrm{E}_\theta\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\psi(x-\theta)\right]=\mathrm{E}_\theta[\psi(X-\theta)l'(\theta|X)]$$

(这里，当我们取$\psi=l'$，就得出（我们希望的）熟悉的等式$-\mathrm{E}_\theta[l''(\theta\mid X)]=\mathrm{E}_\theta l'(\theta\mid X)^2$；见引理7.3.11).

现在比较一个M－估计量和MLE的渐近方差就是一件简单的事情了，回忆MLE $\hat\theta$的渐近方差是$1/\mathrm{E}_\theta l'(\theta|X)^2$，所以借助于Cauchy-Schwarz不等式我们有

$$\mathrm{ARE}(\hat\theta_M,\hat\theta)=\frac{[\mathrm{E}_\theta\psi(X-\theta_0)l'(\theta|X)]^2}{\mathrm{E}_\theta\psi(X-\theta)^2\mathrm{E}_\theta l'(\theta|X)^2}\leqslant 1 \tag{10.2.9}$$

因此，一个M-估计量的效率总比MLE低，只有当ψ和l'成比例时它的效率才能与MLE相匹敌（见习题10.29).

这一节我们没有试图对所有的稳健估计量分类，而是限于一些例子. 有许多很好的详细论述稳健性的书籍；有兴趣的读者可以试阅Staudte and Sheather（1990）或者Hettmansperger and McKean（1998).

10.3 假设检验

与在10.1节一样，本节描述几种在复杂的问题中获得某些检验的方法. 在这些问题当中，不存在或不知道有像以前几节中所定义的那些最佳（例如UMP无偏）检验. 在这种情况下，任何合理的检验的推导都可能有用. 下面我们将用两小节讨论似然比检验的大样本性质和其他近似的大样本检验.

10.3.1 LRT的渐近分布

对于复杂模型最有用的方法之一就是构造检验的似然比方法，因为它给出检验统计量的一个显式的定义

$$\lambda(\boldsymbol{x})=\frac{\sup_{\Theta_0}L(\theta|\boldsymbol{x})}{\sup_{\Theta}L(\theta|\boldsymbol{x})},$$

而且给出了拒绝区域的一个显式的形式，即$\{\boldsymbol{x}:\lambda(\boldsymbol{x})\leqslant c\}$. 在观测到数据$\boldsymbol{X}=\boldsymbol{x}$之后，似然函数$L(\theta\mid\boldsymbol{x})$就是变量$\theta$的一个完全被定义了的函数. 即使$L(\theta\mid\boldsymbol{x})$在集合$\Theta_0$和集合$\Theta$上的两个上确界不能被解析地得到，它们一般也可以用数值方法计算. 因此，即使$\lambda(\boldsymbol{x})$没有方便的定义式可用，对于观测的数据点，检验统计

量 $\lambda(\boldsymbol{x})$ 值仍可以得到.

为了定义一个水平 α 检验，必须选择常数 c 以使得

$$\sup_{\Theta_0} P_\theta(\lambda(\boldsymbol{X}) \leqslant c) \leqslant \alpha. \tag{10.3.1}$$

如果我们不能得到 $\lambda(\boldsymbol{x})$ 的一个简单公式，似乎就没有希望得出 $\lambda(\boldsymbol{X})$ 的样本分布也就不知如何挑选 c 以使方程（10.3.1）成立. 然而，如果我们借助于渐近分布，我们就能够得到一个近似答案.

类似于定理 10.1.12，我们有以下结果.

定理 10.3.1（LRT 的渐近分布简单 H_0） 关于检验 $H_0: \theta=\theta_0$ 对 $H_1: \theta\neq\theta_0$，设 $X_1, \cdots, X_n$ 是 iid $f(x|\theta)$，$\hat{\theta}$是 θ 的 MLE，并且 $f(x|\theta)$满足在杂录 10.6.2 中的正则性条件. 则在 H_0 之下，当 $n \to \infty$，

$$-2\log \lambda(\boldsymbol{X}) \xrightarrow{L} \chi_1^2,$$

其中 χ_1^2 是一个具有自由度 1 的 χ^2 分布随机变量.

证明： 首先在$\hat{\theta}$的邻域展开 $\log L(\theta|\boldsymbol{x})=l(\theta|\boldsymbol{x})$为 Taylor 级数，有

$$l(\theta|\boldsymbol{x})=l(\hat{\theta}|\boldsymbol{x})+l'(\hat{\theta}|\boldsymbol{x})(\theta-\hat{\theta})+l''(\hat{\theta}|\boldsymbol{x})\frac{(\theta-\hat{\theta})^2}{2}+\cdots.$$

现在把 $l(\theta_0|\boldsymbol{x})$的展开式代入$-2\log\lambda(\boldsymbol{x})=-2l(\theta_0|\boldsymbol{x})+2l(\hat{\theta}|\boldsymbol{x})$中，得到

$$-2\log\lambda(\boldsymbol{x})\approx\frac{(\theta-\hat{\theta})^2}{-l''(\hat{\theta}|\boldsymbol{x})},$$

这里我们用到 $l'(\hat{\theta}\mid\boldsymbol{x})=0$ 这个事实. 因为分母就是观测信息数$\hat{I}_n(\hat{\theta})$ 并且$\hat{I}_n(\hat{\theta})\to I(\theta_0)$，于是根据定理 10.1.12 和 Slutsky 定理（定理 5.5.17）就推断出 $-2\log\lambda(\boldsymbol{X})\to\chi_1^2$. ‖

例 10.3.2（Poisson LRT） 考虑基于 iid Poisson（λ）的样本 $X_1, \cdots, X_n$ 的检验 $H_0: \lambda=\lambda_0$ 对 $H_1: \lambda\neq\lambda_0$，我们有

$$-2\log\lambda(\boldsymbol{x}) = -2\log\left(\frac{\mathrm{e}^{-n\lambda_0}\lambda_0^{\sum x_i}}{\mathrm{e}^{-n\hat{\lambda}}\hat{\lambda}^{\sum x_i}}\right) = 2n[(\lambda_0-\hat{\lambda})-\hat{\lambda}\log(\lambda_0/\hat{\lambda})],$$

其中$\hat{\lambda}=\sum x_i/n$ 是λ 的 MLE. 运用定理 10.3.1，如果 $-2\log\lambda(\boldsymbol{x})>\chi^2_{1,\alpha}$ 我们就将在水平 α 上拒绝 H_0.

为了对这个渐近的准确度有些认识，这里给出这个检验的一个小型模拟. 设 $\lambda_0=5$ 和 $n=25$，图 10.3.1. 显示的是把 $-2\log\lambda(\boldsymbol{x})$ 的 10000 个值做成的直方图与 χ_1^2 的概率密度函数图放在一起. 从图看起来，符合是较好的. 此外，在下面的表中给出了模拟的（确切的）和 χ_1^2（近似的）分界点的比较值，它表明两者的分位点非常近似.

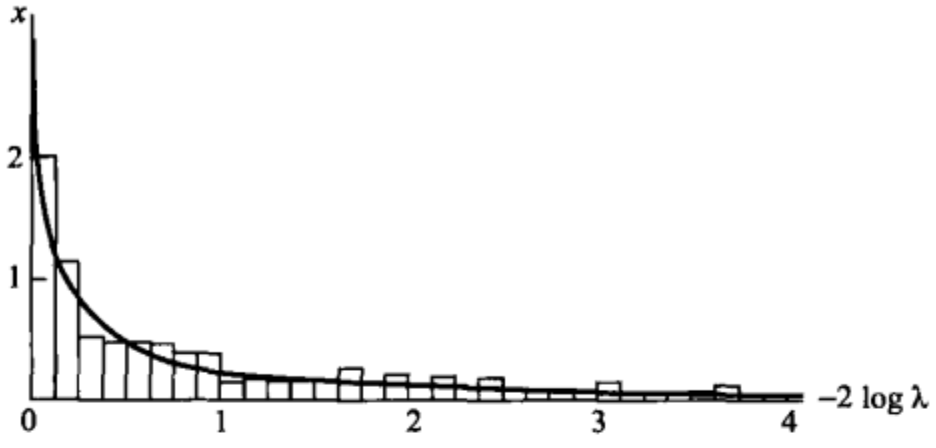

图 10.3.1 $-2\log\lambda(\boldsymbol{x})$ 的 10000 个值做成的直方图与 χ_3^2 的概率密度函数图，$\lambda_0=5$ 和 $n=25$

Poisson LRT 统计量的模拟的（确切的）和近似的百分位数

百分位数	0.80	0.90	0.95	0.99
模拟的	1.630	2.726	3.744	6.304
χ^2	1.642	2.706	3.841	6.635

定理 10.3.1 可以扩展到其中的原假设涉及的是一个参数向量的情况. 我们不加证明地叙述下面的推广，它允许我们确信式（10.3.1）是正确的，至少对于大样本正确. 这个课题的完整讨论可以在 Stuart，Ord and Arnold（1999，第 22 章）中找到.

定理 10.3.3 设 X_1，…，X_n 是来自一个概率密度函数或概率质量函数 $f(x\mid\theta)$的随机样本. 在杂录 10.6.2 中的正则性条件之下，如果 $\theta\in\Theta_0$，则统计量 $-2\log\lambda(\boldsymbol{X})$ 的分布在样本容量 $n\to\infty$时收敛到一个 χ^2 分布. 这个极限分布的自由度是由 $\theta\in\Theta$ 指明的自由参数个数与由 $\theta\in\Theta_0$ 指明的自由参数个数之差.

对于 $\lambda(\boldsymbol{X})$ 过小的值拒绝 $H_0:\theta\in\Theta_0$ 等价于对于 $-2\log\lambda(\boldsymbol{X})$ 过大的值作出拒绝. 因此

$$H_0\text{被拒绝，当且仅当}\quad -2\log\lambda(\boldsymbol{X})\geqslant\mathcal{X}^2_{\nu,\alpha}$$

其中 ν 是定理 10.3.3 中指出的自由度. 如果 $\theta\in\Theta_0$ 且样本量很大，犯第一类错误的概率将近似为 α. 这样，对于大的样本量，方程（10.3.1）将近似地得到满足，从而也定义了一个渐近的真实水平 α 检验. 注意定理实际上仅仅蕴涵

$$\lim_{n\to\infty}P_\theta(\text{拒绝 }H_0)=\alpha,\text{对于每个 }\theta\in\Theta_0,$$

而不是 $\sup_{\theta\in\Theta_0}P_\theta$（拒绝 H_0）收敛到 α. 渐近的真实水平 α 检验情况通常是这样.

检验统计量自由度的计算通常是直接的. 最经常的是，Θ 可以表示为 q 维欧氏空间的一个子集合，它包含 $\mathbf{R}^q$ 中的一个开子集，而 Θ_0 可以表示为 p 维欧氏空间的一个子集合，它包含 $\mathbf{R}^p$ 中的一个开子集，其中 $p<q$. 则 $q-p=\nu$ 就是这个检验统计量的自由度.

例 10.3.4（多项分布 LRT） 设 $\theta=(p_1, p_2, p_3, p_4, p_5)$，其中这些 p_j 非负并且和是 1. 设 $X_1, \cdots, X_n$ 是 iid 的离散随机变量，而且 $P_\theta(X_i=j)=p_j$，$j=1, 2, 3, 4, 5$. 这样，X_i 的概率质量函数是 $f(j|\theta)=p_j$，而且似然函数是

$$L(\theta \mid \boldsymbol{x}) = \prod_{i=1}^{n} f(x_i \mid \theta) = p_1^{y_1} p_2^{y_2} p_3^{y_3} p_4^{y_4} p_5^{y_5},$$

其中 $y_j=x_1, \cdots, x_n$ 中等于 j 的个数. 考虑检验

$$H_0: p_1=p_2=p_3, \text{并且 } p_4=p_5 \text{ 对 } H_1: H_0 \text{ 不真.}$$

完全的参数空间 Θ 实际是一个四维集合，这是因为 $p_5=1-p_1-p_2-p_3-p_4$，所以只有四个自由参数. 这个参数集合定义为

$$\sum_{j=1}^{4} p_j \leqslant 1, \text{并且 } p_j \geqslant 0, j=1, \cdots, 4,$$

它是 R^4 的一个子集合，包含 R^4 中的一个开子集. 因此 $q=4$. H_0 所指出的集合里只有一个自由参数，这是因为一旦 p_1 被固定，$0 \leqslant p_1 \leqslant \frac{1}{3}$，$p_2=p_3$ 就必须等于 p_1，且 $p_4=p_5$ 必须等于 $\frac{1-3p_1}{2}$. 因此 $p=1$，而自由度 $\nu=4-1=3$.

为了计算 $\lambda(\boldsymbol{x})$，必须确定在 Θ_0 和 Θ 之下 θ 的 MLE. 通过使

$$\frac{\partial}{\partial p_j} \log L(\theta \mid \boldsymbol{x}) = 0, \text{对于每个 } j=1, \cdots, 4,$$

并且利用 $p_5=1-p_1-p_2-p_3-p_4$ 和 $y_5=n-y_1-y_2-y_3-y_4$，我们能够验证 Θ 之下 p_j 的 MLE 是 $\hat{p}_j=y_j/n$. 在 H_0 之下，似然函数则化简为

$$L(\theta \mid \boldsymbol{x}) = p_1^{y_1+y_2+y_3} \left(\frac{1-3p_1}{2}\right)^{y_4+y_5}.$$

再次利用使其导数等于 0 的通常方法就证明出在 H_0 之下 p_1 的 MLE 是 $\hat{p}_{10}=(y_1+y_2+y_3)/(3n)$. 则有 $\hat{p}_{10}=\hat{p}_{20}=\hat{p}_{30}$ 和 $\hat{p}_{40}=\hat{p}_{50}=(1-3\hat{p}_{10})/2$. 把这些值和 $\hat{p}_j$ 值代入 $L(\theta \mid \boldsymbol{x})$ 并且把相同的指数项结合就得到

$$\lambda(\boldsymbol{x})=\left(\frac{y_1+y_2+y_3}{3y_1}\right)^{y_1}\left(\frac{y_1+y_2+y_3}{3y_2}\right)^{y_2}\left(\frac{y_1+y_2+y_3}{3y_3}\right)^{y_3}\left(\frac{y_4+y_5}{2y_4}\right)^{y_4}\left(\frac{y_4+y_5}{2y_5}\right)^{y_5}.$$

这样，检验统计量是

$$-2\log\lambda(\boldsymbol{x}) = 2\sum_{i=1}^{5} y_i \log\left(\frac{y_i}{m_i}\right), \tag{10.3.2}$$

其中 $m_1=m_2=m_3=(y_1+y_2+y_3)/3$ 和 $m_4=m_5=(y_4+y_5)/2$. 渐近真实水平 α 检验当 $-2\log\lambda(\boldsymbol{x}) \geqslant \chi^2_{3,\alpha}$ 时拒绝 H_0. 有一大类经常利用似然比检验渐近理论解决的检验问题，本例是其中的一个. ‖

10.3.2 其他大样本检验

另外一个构造大样本检验统计量的通用方法是建立在一个具有渐近正态分布的

估计量之上的. 设我们要检验关于一个实值参数 θ 的假设，而 $W_n=W(X_1,\cdots,X_n)$ 是一个基于样本容量 n 的通过某种方法得到的 θ 的点估计量. 例如，W_n 可能是 θ 的 MLE. 于是一个基于正态近似的近似检验可以通过下面途径证明它的合理性. 如果把 W_n 的方差记作 σ_n^2，而且如果我们能够用某种形式的中心极限定理证明当 $n\to\infty$ 时，$(W_n-\theta)/\sigma_n$ 依分布收敛到一个标准正态随机变量，则 $(W_n-\theta)/\sigma_n$ 就可以比作一个 n（0，1）分布. 我们因此就有了一个近似检验的基础.

当然，在前段论述中有很多细节要验证，但是这些想法的确应用于很多情形. 例如，如果 W_n 是一个 MLE，则可以使用定理 10.1.12 来证实上面论述的正确性. 要注意，W_n 的分布，也许还有 σ_n 的值依赖于 θ 的值. 因此，关于收敛更正式的说法是，对于每个固定的 $\theta\in\Theta$，如果我们对 W_n 使用其相应的分布而对 σ_n 使用其相应的值，则 $(W_n-\theta)/\sigma_n$ 收敛到一个标准正态分布. 如果对于每个 n，σ_n 是一个可计算的常数（它可能依赖于 θ 但不依赖于其他未知参数），那么就可以推导出一个基于 $(W_n-\theta)/\sigma_n$ 的检验.

在某些情况中，σ_n 还依赖于未知的参数. 在这种情况，我们寻找 σ_n 的一个的估计 S_n，满足 σ_n/S_n 依概率收敛到 1. 然后运用 Slutsky 定理（如例 5.5.18）我们就可以推出 $(W_n-\theta)/S_n$ 同样依分布收敛到一个标准正态分布. 这样就可以建立一个大样本检验.

设我们想检验双侧假设 $H_0:\theta=\theta_0$ 对 $H_1:\theta\neq\theta_0$. 一个近似检验就可以建立在统计量 $Z_n=(W_n-\theta_0)/S_n$ 之上，并且当且仅当 $Z_n<-z_{\alpha/2}$ 或 $Z_n>z_{\alpha/2}$ 时拒绝 H_0. 如果 H_0 为真，则 $\theta=\theta_0$ 并且 Z_n 依分布收敛到 $Z\sim$n（0，1）. 这样，犯第一类错误的概率

$$P_{\theta_0}(Z_n<-z_{\alpha/2}\text{ 或 }Z_n>z_{\alpha/2})\to P_{\theta_0}(Z<-z_{\alpha/2}\text{ 或 }Z>z_{\alpha/2}),$$

从而这是一个渐近的真实水平 α 检验.

现在考虑另一个 $\theta\neq\theta_0$ 的参数值. 我们可以写成

$$Z_n=\frac{W_n-\theta_0}{S_n}=\frac{W_n-\theta}{S_n}+\frac{\theta-\theta_0}{S_n}. \tag{10.3.3}$$

不管 θ 的值是什么，都有 $(W_n-\theta)/S_n\to$n（0，1）. 典型情况下还有当 $n\to\infty$ 时，$\sigma_n\to0$.（回忆，$\sigma_n=\mathrm{Var}W_n$，而典型情况下当 $n\to\infty$ 时估计量变得愈渐精确.）这样，S_n 将依概率收敛到 0，$(\theta-\theta_0)/S_n$ 将依概率收敛到 $+\infty$ 或 $-\infty$，依赖于 $(\theta-\theta_0)$ 为正或负. 所以 Z_n 将依概率收敛到 $+\infty$ 或 $-\infty$，而且

$$P_\theta(\text{拒绝 }H_0)=P_\theta(Z<-z_{\alpha/2}\text{ 或 }Z>z_{\alpha/2})\to1,\text{当 }n\to\infty.$$

这样，就可以构造出一个具有渐近真实水平 α 和渐近功效 1 的检验.

我们想检验单侧假设 $H_0:\theta\leqslant\theta_0$ 对 $H_1:\theta>\theta_0$，可以构造一个类似的检验. 这时将再次利用检验统计量 $Z_n=(W_n-\theta_0)/S_n$，而且当且仅当 $Z_n>z_\alpha$ 时拒绝 H_0. 运用与前面类似的推理，我们可以断言这个检验的功效函数根据 $\theta<\theta_0$，$\theta=\theta_0$ 或

$\theta>\theta_0$收敛到 0，α 或 1. 因此，这个检验也具有合理的渐近功效性质.

一般而言，一个 Wald 检验（Wald test）是一个基于形式为

$$Z_n=\frac{W_n-\theta_0}{S_n}$$

的统计量的检验，其中 θ_0 是参数 θ 的一个假设值，W_n 是 θ 的一个估计量，而 S_n 是 W_n 的标准误差，即 W_n 标准差的一个估计. 如果 W_n 是 θ 的 MLE，那么就如 10.1.3 节所讨论，$1/\sqrt{I_n(W_n)}$ 是 W_n 的一个合理的标准误差. 也经常用 $1/\sqrt{\hat{I}_n(W_n)}$替换它，其中

$$\hat{I}_n(W_n)=-\frac{\partial^2}{\partial\theta^2}\log L(\theta|\boldsymbol{X})\Big|_{\theta=W_n}$$

是观测信息数［参见（10.1.7)］.

例 10.3.5（大样本二项检验） 设 X_1，…，X_n 是来自总体 Bernoulli（p）的随机样本. 考虑检验 H_0：$p\leqslant p_0$ 对 H_1：$p>p_0$，其中 $0<p_0<1$ 是一个指定的值. p 的基于样本容量 n 的 MLE 是$\hat{p}_n=\sum_{i=1}^{n}X_i/n$. 因为$\hat{p}_n$ 正好是样本均值，所以中心极限定理适用，而且说明对于任何的 p，$0<p<1$，$(\hat{p}_n-p)/\sigma_n$ 收敛到一个标准正态随机变量，其中 $\sigma_n=\sqrt{p(1-p)/n}$是一个依赖于未知参数 p 的值. σ_n 的一个合理的估计是 $S_n=\sqrt{\hat{p}_n(1-\hat{p}_n)/n}$，而且可以证明（见习题 5.32）$\sigma_n/S_n$ 依概率收敛到 1. 这样，对于任何的 p，$0<p<1$，

$$\frac{\hat{p}_n-p}{\sqrt{\dfrac{\hat{p}_n(1-\hat{p}_n)}{n}}}\to \mathrm{n}(0,1).$$

Wald 检验统计量 Z_n 是在上式中把 p 换成 p_0，而大样本 Wald 检验当 $Z_n>z_\alpha$ 时拒绝 H_0. 作为 σ_n 的一种替换的估计，容易验证 $1/I_n(\hat{p}_n)=\hat{p}_n(1-\hat{p}_n)/n$. 所以，如果我们使用信息数去推导$\hat{p}_n$ 的标准误差，则得到相同的统计量 Z_n.

如果对于双侧检验 H_0：$p=p_0$ 对 H_1：$p\neq p_0$ 感兴趣，其中 $0<p_0<1$ 是一个指定的值，可以再次应用上面的策略. 然而在这种情况下，有另外一个近似检验. 根据中心极限定理，对于任何的 p，$0<p<1$，

$$\frac{\hat{p}_n-p}{\sqrt{p(1-p)/n}}\to \mathrm{n}(0,1).$$

因此，如果原假设为真，则统计量

$$Z'_n=\frac{\hat{p}_n-p_0}{\sqrt{p_0(1-p_0)/n}}\sim \mathrm{n}(0,1)\quad\text{（近似地）}\tag{10.3.4}$$

这个近似的水平 α 检验在 $|Z'_n|>z_{\alpha/2}$时拒绝 H_0.

在两个检验都适用的情况，例如，当检验假设 H_0：$p=p_0$ 时，不清楚选择哪

个. 它们的功效函数（指实际的，而不是近似的）互相交叉，所以每一个检验都是在一部分的参数空间上功效更强. （Ghosh 1979 对这个问题给出一些启示. Robbins 1977 及 Eberhardt and Fligner 1977 讨论了关于两样本二项分布问题的论证. 习题 10.31 给出了关于这个问题的两个不同的检验.）

当然，任何对于功效函数的比较都被如下的事实所混扰，即这些检验是近似的而不必保持水平 α. 利用连续性校正（见例 3.3.2）有助于这个问题. 在很多情况中，使用连续性校正的近似方法是保守的，就是说，它们保持其名义 α 水平（见例 10.4.6）. ‖

式（10.3.4）是另一个有用的大样本检验，即记分检验（score test）的一个特例，记分统计量（score statistics）的定义是

$$S(\theta)=\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)=\frac{\partial}{\partial\theta}\log L(\theta|\boldsymbol{X}).$$

根据式（7.3.8）我们知道，对于所有的 θ，$\mathrm{E}_\theta S(\theta)=0$. 特别地，如果我们在检验 $H_0:\theta=\theta_0$ 并且 H_0 为真，则 $S(\theta_0)$ 的均值是 0. 进一步，根据式（7.3.10），

$$\mathrm{Var}_\theta S(\theta)=\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log L(\theta\mid\boldsymbol{X})\right)^2\right)=-\mathrm{E}_\theta\left(\frac{\partial^2}{\partial\theta^2}\log L(\theta\mid\boldsymbol{X})\right)=I_n(\theta);$$

这里的信息数是记分统计量的方差. 记分检验的检验统计量是

$$Z_S=S(\theta_0)/\sqrt{I_n(\theta_0)}.$$

如果 H_0 为真，Z_S 具有 0 均值和 1 方差. 根据定理 10.1.12 就可推出如果 H_0 为真，则 Z_S 收敛到一个标准正态随机变量. 这样，近似的水平 α 记分检验当 $|Z_S|>z_{\alpha/2}$ 时拒绝 H_0. 如果 H_0 是复合假设，$\hat{\theta}_0$ 是假定 H_0 真时 θ 的估计，则把 Z_S 中的 θ_0 替换成 $\hat{\theta}_0$. 如果 $\hat{\theta}_0$ 是限制的 MLE，限制极大化可利用拉格朗日乘数法实现. 因此，这个记分检验有时称为拉格朗日乘数检验（Lagrange multiplier test）.

例 10.3.6（二项记分检验） 再来考虑例 10.3.5 的 Bernoulli 模型，并且考虑检验 $H_0:p=p_0$ 对 $H_1:p\neq p_0$. 直接的计算得出

$$S(p)=\frac{\hat{p}_n-p}{p(1-p)/n}\text{和}\ I_n(p)=\frac{n}{p(1-p)}.$$

因此，记分统计量是

$$Z_S=\frac{S(p_0)}{\sqrt{I_n(p_0)}}=\frac{\hat{p}_n-p_0}{\sqrt{p_0(1-p_0)/n}},$$

它和式（10.3.4）相同. ‖

即将考虑的最后一类近似检验是稳健检验（见杂录 10.6.6）. 在 1.2 节，我们曾看到如果 $X_1,\cdots,X_n$ 是 iid 地来自一个位置族而 $\hat{\theta}_M$ 是一个 M-估计量，则

$$\sqrt{n}(\hat{\theta}_M-\theta_0)\to\mathrm{n}(0,\mathrm{Var}_{\theta_0}(\hat{\theta}_M)),\tag{10.3.5}$$

其中 $\mathrm{Var}_{\theta_0}(\hat{\theta}_M)=\dfrac{\mathrm{E}_{\theta_0}\psi(X-\theta_0)^2}{[\mathrm{E}_{\theta_0}\psi'(X-\theta_0)]^2}$ 是其渐近方差. 这样我们就可以构造一个“广义”的记分统计量，

$$Z_{GS}=\sqrt{n}\frac{\hat{\theta}_M-\theta_0}{\sqrt{\mathrm{Var}_{\theta_0}(\hat{\theta}_M)}},$$

或者一个广义的 Wald 统计量，

$$Z_{GW}=\sqrt{n}\frac{\hat{\theta}_M-\theta_0}{\sqrt{\widehat{\mathrm{Var}}_{\theta_0}(\hat{\theta}_M)}},$$

其中 $\widehat{\mathrm{Var}}_{\theta_0}(\hat{\theta}_M)$ 可以是任意的相合估计量. 例如，我们可以使用标准误差的一个自助估计，或者简单地把一个估计量代入到式（10.2.6）中并且用

$$\widehat{\mathrm{Var}}_1(\hat{\theta}_M)=\frac{\frac{1}{n}\sum_{i=1}^{n}[\psi(x_i-\hat{\theta}_M)]^2}{\left[\frac{1}{n}\sum_{i=1}^{n}\psi'(x_i-\hat{\theta}_M)\right]^2}. \tag{10.3.6}$$

方差估计的选择可能会很重要；有关指导请参看 Boos（1992）或 Carroll，Ruppert and Stefanski（1995，附录 A.3).

例 10.3.7（基于 Huber 估计量的检验） 设 X_1，…，X_n 是 iid 来自一个概率密度函数 $f(x-\theta)$，其中 f 关于 0 对称，则对于 Huber M-估计量使用式（10.2.2）中的 ρ 函数和式（10.2.7）中的 ψ 函数，我们得到一个渐近方差

$$\frac{\int_{-k}^{k}x^2f(x)\mathrm{d}x+k^2P_0(|X|>k)}{[P_0(|X|\leqslant k)]^2}. \tag{10.3.7}$$

因此，基于 M—估计的渐近正态性，我们可以（例如）在水平 α 检验假设 $H_0:\theta=\theta_0$ 对 $H_1:\theta\neq\theta_0$，如果 $|Z_{GS}|>z_{\alpha/2}$ 就拒绝 H_0. 为更实用些，我们将考虑一种使用标准误差的估计的近似检验. 我们将使用统计量 Z_{GW}，但是将把我们的方差估计建立在方程（10.3.7）上，就是

$$\widehat{\mathrm{Var}}_2(\hat{\theta}_M)= \tag{10.3.8}$$

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\theta}_M)^2I(|x_i-\hat{\theta}_M|<k)+k^2\left(\frac{1}{n}\sum_{i=1}^{n}I(|x_i-\hat{\theta}_M|>k)\right)}{\left(1-\frac{1}{n}\sum_{i=1}^{n}I(|x_i-\hat{\theta}_M|<k)\right)^2}$$

此外，我们增添一种“朴素”的检验 Z_N，它使用一个简单的方差估计

$$\widehat{\mathrm{Var}}_3(\hat{\theta}_M)=\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\theta}_M)^2. \tag{10.3.9}$$

这些检验怎么样？解析的评价是困难的，但是表 10.3.1 的小型模拟表明 $z_{\alpha/2}$ 分位点通常都太小（忽略方差估计中的不同），真实水平通常大于名义真实水平. 但

对有一类分布却是一致的，双指数分布就是最好的情况. （最后的这一点并不完全令人惊讶，因为 Huber 估计量对于指数拖尾分布具有一种最优性；参见 Huber 1981，第 4 章.）‖

表 10.3.1. 基于 Z_{GW} 和 Z_N 的名义水平 $\alpha=0.1$ 检验的在指定的参数值的功效

样本容量 $n=15$（模拟次数 10000）

	基础 pdf							
	正态		t_5		罗吉斯蒂克		双指数	
	Z_{GW}	Z_N	Z_{GW}	Z_N	Z_{GW}	Z_n	Z_{GW}	Z_N
θ_0	0.16	0.16	0.14	0.13	0.15	0.15	0.11	0.09
$\theta_0+0.25\sigma$	0.27	0.29	0.29	0.27	0.27	0.27	0.31	0.26
$\theta_0+0.5\sigma$	0.58	0.60	0.65	0.63	0.59	0.60	0.70	0.64
$\theta_0+0.75\sigma$	0.85	0.87	0.89	0.89	0.85	0.87	0.92	0.90
$\theta_0+1\sigma$	0.96	0.97	0.98	0.97	0.96	0.97	0.98	0.98
$\theta_0+2\sigma$	1	1	1	1	1	1	1	1

10.4 区间估计

像我们在前面两节已做的那样，现在来探索几种近似和渐近置信集合形式. 就像以往，我们的目的是用例子说明一些将被用于更加复杂情况的方法，将得到某种解答的方法. 这里得到的解答几乎一定不是最好的，但是一定不是最坏的. 在很多情况下，它们却是我们所能做到的最好的.

仍像过去，我们从基于 MLE 的近似开始.

10.4.1 近似极大似然区间

根据 10.1.2 节的讨论，并且运用定理 10.1.12，我们就有了求 MLE 渐近分布的一般方法，从而就有了构造一个置信区间的一般方法.

如果 X_1，…，X_n 是 iid $f(x|\theta)$ 的而 $\hat{\theta}$ 是 θ 的 MLE，则根据式(10.1.7)，$\hat{\theta}$ 的一个函数 $h(\hat{\theta})$ 的方差可以由

$$\widehat{\mathrm{Var}}(h(\hat{\theta})\,|\,\theta)\approx\frac{[h'(\theta)]^2\,|_{\theta=\hat{\theta}}}{-\dfrac{\partial^2}{\partial\theta^2}\log L(\theta|\boldsymbol{x})\,|_{\theta=\hat{\theta}}}$$

近似. 现在，对于一个固定的但是任意的 θ 值，我们对

$$\frac{h(\hat{\theta})-h(\theta)}{\sqrt{\widehat{\mathrm{Var}}(h(\hat{\theta})\,|\,\theta)}}$$

的渐近分布感兴趣. 从定理 10.1.12 和 Slutsky 定理（定理 5.5.17）(见习题 10.33) 就可推出

$$\frac{h(\hat{\theta})-h(\theta)}{\sqrt{\widehat{\mathrm{Var}}(h(\hat{\theta})\,|\,\theta)}}\to \mathrm{n}(0,1),$$

于是给出近似的置信区间

$$h(\hat{\theta})-z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}(h(\hat{\theta})\,|\,\theta)}\leqslant h(\theta)\leqslant h(\hat{\theta})+z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}(h(\hat{\theta})\,|\,\theta)}.$$

例 10.4.1（例 10.1.14 续） 我们有来自一个 Bernoulli（p）总体的随机样本 X_1，…，X_n. 我们曾看到为估计胜率 $p/(1-p)$ 可以用其 MLE $\hat{p}/(1-\hat{p})$，还看到此估计有近似方差

$$\widehat{\mathrm{Var}}\left(\frac{\hat{p}}{1-\hat{p}}\right)\approx\frac{\hat{p}}{\mathrm{n}(1-\hat{p})^3}.$$

我们因此就可以构造近似的置信区间

$$\frac{\hat{p}}{1-\hat{p}}-z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}\left(\frac{\hat{p}}{1-\hat{p}}\right)}\leqslant\frac{p}{1-p}\leqslant\frac{\hat{p}}{1-\hat{p}}+z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}\left(\frac{\hat{p}}{1-\hat{p}}\right)}.$$ ‖

基于记分统计量（见 10.3.2 节），可以构造似然逼近的一种限制形式. 这种方法，在其适用时，可以给出更好的区间. 随机量

$$Q(\boldsymbol{X}|\theta)=\frac{\dfrac{\partial}{\partial\theta}\log L(\theta|\boldsymbol{X})}{\sqrt{-\mathrm{E}_\theta\left(\dfrac{\partial^2}{\partial\theta^2}\log L(\theta|\boldsymbol{X})\right)}} \tag{10.4.1}$$

当 $n\to\infty$ 时有渐近分布 n（0，1）. 因此，集合

$$\{\theta:|Q(\boldsymbol{x}|\theta)|\leqslant z_{\alpha/2}\} \tag{10.4.2}$$

是一个近似的 $1-\alpha$ 置信集合. 注意，运用 7.3.2 节的结果，我们有

$$\mathrm{E}_\theta(Q(\boldsymbol{X}|\theta))=\frac{\mathrm{Var}_\theta\left(\dfrac{\partial}{\partial\theta}\log L(\theta|\boldsymbol{X})\right)}{\sqrt{-\mathrm{E}_\theta\left(\dfrac{\partial^2}{\partial\theta^2}\log L(\theta|\boldsymbol{X})\right)}}=0$$

和

$$\mathrm{Var}_\theta(Q(\boldsymbol{X}|\theta))=\frac{\mathrm{Var}_\theta\left(\dfrac{\partial}{\partial\theta}\log L(\theta|\boldsymbol{X})\right)}{-\mathrm{E}_\theta\left(\dfrac{\partial^2}{\partial\theta^2}\log L(\theta|\boldsymbol{X})\right)}=1, \tag{10.4.3}$$

因此这个近似与一个 n(0，1) 随机变量的前两阶矩准确匹配. Wilks（1938）证明了这些区间具有一种渐近最优性质；它们是渐近地在某一个区间类中最短的.

当然，这些区间不够一般，对一个函数 $h(\theta)$ 就未必适用. 此时，我们必须能把式（10.4.2）表示成 $h(\theta)$ 的一个函数才行.

例 10.4.2（二项记分区间） 仍旧利用一个二项分布的例，如果 $Y=\sum_{i=1}^{n}X_i$，

其中每个 X_i 是一个独立的 Bernoulli（p）随机变量，我们有

$$\begin{aligned} Q(\boldsymbol{X}\mid p) &= \frac{\frac{\partial}{\partial p}\log L(p\mid \boldsymbol{X})}{\sqrt{-\mathrm{E}_p\left(\frac{\partial^2}{\partial p^2}\log L(p\mid \boldsymbol{X})\right)}} \\ &= \frac{\frac{y}{p}-\frac{n-y}{1-p}}{\sqrt{\frac{n}{p(1-p)}}} \\ &= \frac{\hat{p}-p}{\sqrt{p(1-p)/n}}, \end{aligned}$$

其中 $\hat{p}=y/n$. 由式（10.4.2），就给出了一个近似的 $1-\alpha$ 置信区间

$$\left\{p:\left|\frac{\hat{p}-p}{\sqrt{p(1-p)/n}}\right|\leqslant z_{\alpha/2}\right\}. \tag{10.4.4}$$

这就是由反转记分统计量（见例 10.3.6）得到的区间. 为了计算这个区间我们需要解一个关于 p 的二次方程；关于细节参见例 10.4.6. ‖

在 10.3 节我们曾推出另外一个基于 $-2\log\lambda(\boldsymbol{X})$ 具有渐近 χ^2 分布这个事实的似然检验. 这就表明如果 $X_1,\cdots,X_n$ 是 iid $f(x|\theta)$的并且$\hat{\theta}$是θ 的 MLE，则集合

$$\left\{\theta:-2\log\left(\frac{L(\theta|\boldsymbol{x})}{L(\hat{\theta}|\boldsymbol{x})}\right)\leqslant\chi^2_{1,\alpha}\right\} \tag{10.4.5}$$

是一个近似的 $1-\alpha$ 置信区间. 情况就是如此，并且给了我们另一个近似的似然区间.

当然，式（10.4.5）恰好就是我们最初通过反转 LRT 统计量推出的最高似然区域（9.2.7）. 然而，现在我们有了一条自动附加上近似置信水平的途径.

例 10.4.3（二项 LRT 区间） 设 $Y=\sum_{i=1}^{n}X_i$，其中每个 X_i 是一个独立的 Bernoulli（p）随机变量，我们有近似的 $1-\alpha$ 置信集合

$$\left\{p:-2\log\left(\frac{p^y(1-p)^{n-y}}{\hat{p}^y(1-\hat{p})^{n-y}}\right)\leqslant\chi^2_{1,\alpha}\right\}.$$

将在例 10.4.7 中对这个置信集合连同基于记分和 Wald 检验的区间进行比较. ‖

10.4.2 其他大样本区间

大多数近似置信区间是基于求近似的（或渐近的）枢轴或者反转近似的水平 α 检验统计量. 如果有任何的统计量 W 与 V 和一个参数θ 使得当 $n\to\infty$，有

$$\frac{W-\theta}{V}\to \mathrm{n}(0,1),$$

则我们就可以通过

$$W-z_{\alpha/2}V\leqslant\theta\leqslant W+z_{\alpha/2}V$$

构成一个关于 θ 的近似的置信区间，它本质上是一个 Wald-型区间. 直接应用中心极限定理连同 Slutsky 定理，我们通常将给出一个近似的置信区间. （注意，前一节的近似极大似然区间都反映了这个策略.）

例 10.4.4（近似区间） 如果 X_1，…，X_n 是 iid 的具有均值 μ 与方差 σ^2，那么根据中心极限定理，

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\to \mathrm{n}(0,1).$$

进一步，根据 Slutsky 定理，如果 $S^2\xrightarrow{p}\sigma^2$，则

$$\frac{\overline{X}-\mu}{S/\sqrt{n}}\to n\ (0,\ 1),$$

这就给出近似的 $1-\alpha$ 置信区间

(10.4.6) $$\overline{x}-z_{\alpha/2}s/\sqrt{n}\leqslant\mu\leqslant\overline{x}+z_{\alpha/2}s/\sqrt{n}.$$

为了看它近似得有多好，我们这里给出关于各种概率密度函数的近似区间精确覆盖概率的一个小型模拟计算结果. 注意到因为上述区间是枢轴的，所以覆盖概率不依赖于参数值，它是常数因此就是置信系数. 我们从表 10.4.1 看到，甚至对于 $n=15$ 这么小的样本容量，这个枢轴置信区间仍然很合理，不过清楚看出它没达到名义置信系数. 无疑这是由于乐观地使用了分位点 $z_{\alpha/2}$，而没有考虑到 S 的变异. 当样本容量增大，近似将得到改善. ‖

表 10.4.1 枢轴区间（10.4.6）的置信系数，样本容量 $n=15$，模拟次数 10000

	基础 pdf			
名义水平	正态	t_5	罗吉斯蒂克	双指数
$1-\alpha=0.90$	0.879	0.864	0.880	0.876
$1-\alpha=0.95$	0.931	0.924	0.931	0.933

在上例中，我们没有指定样本的分布形式就能够得到一个近似的置信区间. 当我们确实指定了其形式时，我们就应当做得更好.

例 10.4.5（近似的 Poisson 区间） 如果 X_1，…，X_n 是 iid Poisson（p）的，则我们知道有

$$\frac{\overline{X}-\lambda}{S/\sqrt{n}}\to n(0,1).$$

但是，这即使在我们不从 Poisson 总体抽取样本的情况下也是对的. 利用 Poisson 假定，我们就知道 $\mathrm{Var}(X)=\lambda=E(\overline{X})$ 以及 $\overline{X}$ 是 λ 的一个好的估计量（见例 7.3.12）. 因此，根据 Poisson 假定，我们也可以由

$$\frac{\overline{X}-\lambda}{\sqrt{\overline{X}/n}}\to \mathrm{n}(0,1)$$

这个事实得到一个近似的置信区间，它就是从反转 Wald 检验所得到的区间. 我们可以从另外一条路径利用 Poisson 假定. 因为 $\mathrm{Var}(X)=\lambda$，由此就可以得出

$$\frac{\overline{X}-\lambda}{\sqrt{\lambda/n}}\to n(0,1).$$

由此导致相应于记分检验的区间，它也就是式（10.4.2）的似然区间并且根据 Wilks（1938）的结论，它是最好的（见习题 10.40）. ‖

一般而言，一个合理的经验法则是，在近似中要尽可能少用估计多用参数. 其道理很简单，参数被固定且不把任何附加的变动引入近似当中，而每个统计量都代入更多的变动.

例 10.4.6（二项记分区间续） 对于一个抽自 Bernoulli（p）总体的随机样本 $X_1,\cdots,X_n$，我们在例 10.3.5 中曾看到，当 $n\to\infty$，

$$\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}}\text{和}\frac{\hat{p}-p}{\sqrt{p(1-p)/n}}$$

都依分布收敛到一个标准正态的随机变量，其中 $\hat{p}=\sum x_i/n$. 在例 10.3.5 中我们看到，可以基于两个近似中的任何一个来建立检验，前者是 Wald 检验而后者是记分检验. 我们还知道可以利用任何一个近似去构造关于 p 的置信区间. 然而，记分检验近似（具有更少的统计量和更多的参数）将给出例 10.4.2 的区间（10.4.4），它是渐近最优的，就是说

$$\left\{p:\left|\frac{\hat{p}-p}{\sqrt{p(1-p)/n}}\right|\leqslant z_{\alpha/2}\right\}$$

是更好的近似区间.

这个区间并不一目了然，但是我们可以明确解出它的值集合. 如果把两边平方并且重排项，我们要求出 p 的集合以满足

$$\left\{p:(\hat{p}-p)^2\leqslant z_{\alpha/2}^2\frac{p(1-p)}{n}\right\}.$$

这是一个关于 p 的二次不等式，经过进一步整理可以把它表示成更常见的形式

$$\left\{p:\left(1+\frac{z_{\alpha/2}^2}{n}\right)p^2-\left(2\hat{p}+\frac{z_{\alpha/2}^2}{n}\right)p+\hat{p}^2\leqslant 0\right\}.$$

由于在这个二次式中 p^2 的系数为正，二次曲线开口向上，因此，如果 p 处于这个二次式的两个根之间则满足此不等式. 这两个根是

$$\frac{2\hat{p}+z_{\alpha/2}^2/n\pm\sqrt{(2\hat{p}+z_{\alpha/2}^2/n)^2-4\hat{p}^2(1+z_{\alpha/2}^2/n)}}{2(1+z_{\alpha/2}^2/n)},\tag{10.4.7}$$

这两个根确定了关于 p 的置信区间的端点. 虽然关于这些根的表达式有些令人不快，但事实上这个区间是一个关于 p 的很好的区间. 这个区间可以被进一步改进，不过要通过利用连续性校正（见例 3.3.2）. 为此，我们将要分别求解两个二次方程

（见习题 10.45），

$$\left|\frac{\hat{p}+\frac{1}{2n}-p}{\sqrt{p\ (1-p)\ /n}}\right|\leqslant z_{\alpha/2}\qquad\text{（其大的根＝区间上端点）}$$

$$\left|\frac{\hat{p}-\frac{1}{2n}-p}{\sqrt{p\ (1-p)\ /n}}\right|\leqslant z_{\alpha/2}\qquad\text{（其小的根＝区间下端点）}$$

对于端点有显然的修改. 如果 $\sum x_i=0$，则区间下端点取 0，而如果 $\sum x_i=n$，则区间上端点取 1. Blyth（1986）有一些好的近似. ‖

我们现在已经看到三种关于 Bernoulli 比例的区间：基于 Wald 和记分统计量的区间以及例 10.4.3 的 LRT 区间. 典型地，Wald 区间是最不被喜欢的，然而对三者都进行比较是有意义的.

例 10.4.7（比较二项区间） 设 $Y=\sum_{i=1}^{n} X_i$，X_1，…，X_n 是 iid 的，来自一个 Bernoulli（p）总体，Wald 区间是

$$\hat{p}-z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\leqslant p\leqslant\hat{p}+z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\tag{10.4.8}$$

例 10.4.6 描述了记分区间（经连续性校正的），而近似的 LRT 区间由例 10.4.3 给出. 为比较它们，我们来看一个例.

对于 $n=12$，图 10.4.1 显示了这三种方法的实现区间. LRT 方法产生最短的区间，而记分方法产生的区间最长. 对于这个图，我们在 Wald 区间已做了两个修正. 首先，在 $y=0$ 未修正的区间是（0，0），于是我们把上端点改为 $1-(\alpha/2)^n$，在 $y=n$ 对下端点做类似修正. 另外，有些情况其 Wald 区间的端点落到［0，1］之外，这些区间已经被截断.

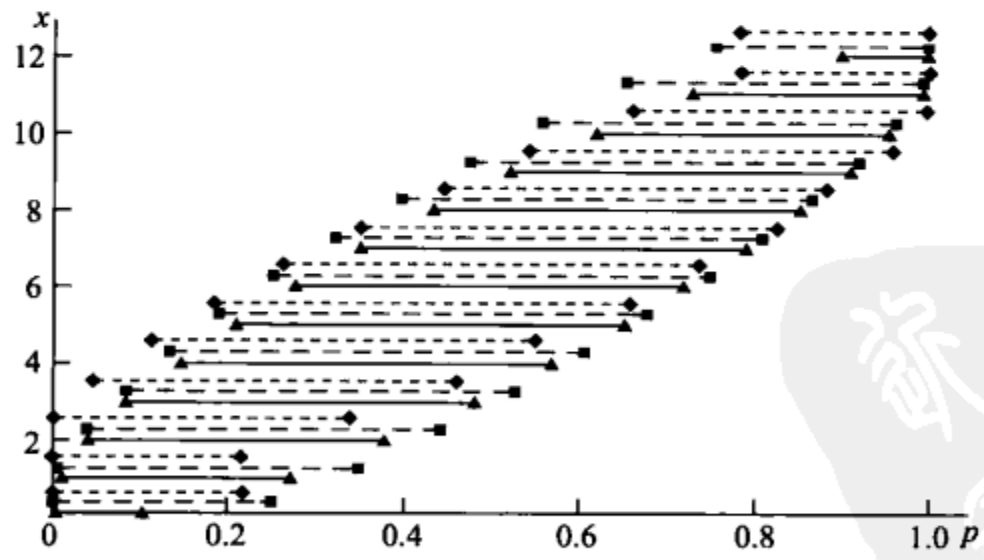

图 10.4.1 来自不同方法的关于 Bernoulli 比例 p 的区间

LRT 方法（实线），记分方法（长破折号），修正 Wald 的方法（短破折号）

在图 10.4.2 中，记分区间较长的长度反映在它较高的覆盖概率上. 确实，记

分区间是唯一的（在这三个中）保持覆盖概率在 0.9 之上的一个，从而是唯一的具有置信系数 0.9 的区间. LRT 和 Wald 区间显得似乎太短，它们的覆盖概率远低于 0.9，从而不能接受它们. 当然，通过增加 n 可以改进它们的表现.

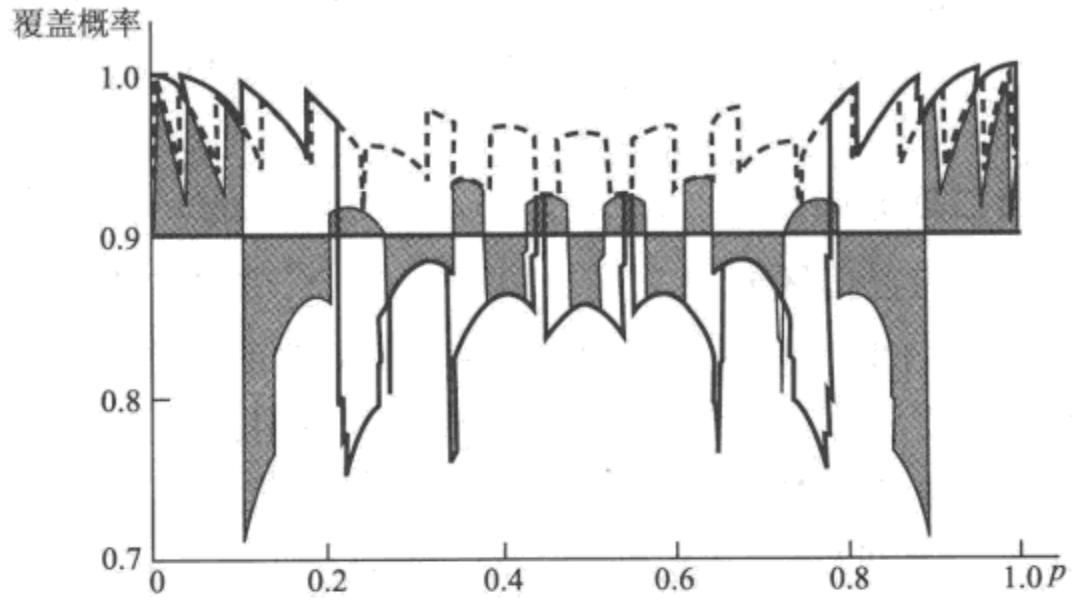

图 10.4.2 关于 Bernoulli 比例 p 的名义 0.9 置信方法的覆盖概率
LRT 方法（细实线，灰色阴影），记分方法（破折线），修正 Wald 的方法（粗实线）

因此看来，连续性校正的记分区间，它尽管较长，但当 n 小的时候是合适的选择（但是在习题 10.44 有另一选择）. LRT 和 Wald 方法产生的区间对于小的 n 却太短了，另外 Wald 区间还要遭受端点的弊病. ‖

就像我们在 10.3.2 节所做，我们简要查看一下基于稳健估计量的区间.

例 10.4.8（基于 Huber 估计量的区间） 类似例 10.3.7 的方法，我们可以基于 Huber 的 M—估计量构造渐近置信区间. 如果 X_1，…，X_n 是 iid 的，来自一个概率密度函数 $f(x-\theta)$，其中 f 关于 0 对称，则我们有关于 θ 的近似区间

$$\hat{\theta}_M \pm z_{a/2}\sqrt{\frac{\mathrm{Var}(\hat{\theta}_M)}{n}},$$

其中 Var（$\hat{\theta}_M$）由式（10.3.7）给出. 现在我们把 Var（$\hat{\theta}_M$）替换成估计式（10.3.8）和式（10.3.9）以得到 Wald 型区间. 为了评价这些区间，我们制作一个类似表 10.4.1 的表. 有趣的是，除了双指数分布之外，表 10.4.2 中的区间与表 10.4.1 中基于常用均值和方差的区间比起来相当糟糕. 对此，除了再次归咎于分位点 $z_{a/2}$ 的过分乐观外，我们没有好的解释. ‖

表 10.4.2 基于 Huber M—估计量的、名义系数 $1-\alpha=0.9$ 的区间的置信系数
样本容量 $n=15$，模拟次数 10000

名义水平	所用概率密度函数			
	正态	t_5	罗吉斯蒂克	双指数
方差估计（10.3.8）	0.844	0.856	0.855	0.889
方差估计（10.3.9）	0.837	0.867	0.855	0.910

到目前为止，所有提到的近似都以 $n\to\infty$ 为基础. 然而，在某些其他情况下，我们也可以利用近似区间. 在例 9.2.17 中，我们曾经需要考虑当参数趋于无穷时的近似. 另一种情况是，在例 2.3.13 中我们看到对于某种参数的结构，Poisson 分布可以用来近似二项分布. 这就是说，如果认为是这样一种参数结构的话，可以基于 Poisson 分布来构造近似的二项区间. 用类似的道理我们来说明下列不太常见的情况.

例 10.4.9（负二项区间） 设 X_1，…，X_n 是 iid 的，服从 NB（r，p）. 我们假定 r 已知并对 p 的置信区间感兴趣. 由于 $Y=\sum X_i\sim NB(nr,p)$，我们可以用多种方式求出置信区间. 应用二项－F 分布之间关系的一个变形，我们可以构造一个精确置信区间（见习题 9.22），或者我们也可以使用正态近似（见习题 10.41）. 有另外一种近似，不是基于大 n，而是基于小 p.

在习题 2.38 中已经证明，当 $p\to 0$ 时，依分布

$$2pY\to\chi^2_{2r}.$$

所以，对于小的 p 值，$2pY$ 是一个枢轴量！利用这个事实，我们可以构造一个对小的 p 值有效的枢轴 $1-\alpha$ 置信区间：

$$\left\{p:\frac{x^2_{2r,1-\alpha/2}}{2y}\leqslant p\leqslant\frac{x^2_{2r,\alpha/2}}{2y}\right\}$$

细节放在习题 10.47 中.

10.5 习题

10.1 X_1，…，X_n 是抽自概率密度函数为

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

的总体的随机样本. 求 θ 的一个相合估计量并证明其相合性.

10.2 证明定理 10.1.5.

10.3 X_1，…，X_n 是抽自总体 n（θ，θ）的随机样本，其中 $\theta>0$.

（a）证明 θ 的 MLE，即$\hat\theta$是二次方程 $\theta^2+\theta-W=0$ 的一个根，其中 $W=(1/n)\sum_{i=1}^{n}X_i^2$，并确定哪个根是 MLE.

（b）用 10.1.3 节中的技术求$\hat\theta$的近似方差.

10.4 习题 7.19 中模型的一个变化是令随机变量 Y_1，…，Y_n 满足

$$Y_i=\beta X_i+\epsilon_i,\quad i=1,\cdots,n,$$

其中 X_1，…，X_n 为独立 n(μ，τ^2) 随机变量，ϵ_1，…，ϵ_n 为 iid n(0，σ^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) $\sum X_iY_i / \sum X_i^2$.

(b) $\sum Y_i / \sum X_i$.

(c) $\sum (Y_i/X_i) /n$.

10.5 就例 10.1.8 中的情形，对 $T_n=\sqrt{n}/\overline{X}_n$ 证明：

(a) Var $(T_n)=\infty$.

(b) 如果 $\mu\neq 0$，并且从样本空间中删去 $(-\delta, \delta)$，则 Var $(T_n)<\infty$.

(c) 如果 $\mu\neq 0$，区间 $(-\delta, \delta)$ 中的概率当 $n\to\infty$时趋近于 0.

10.6 就例 10.1.10 中的情形证明

(a) $EY_n=0$ 且 $\mathrm{Var}(Y_n)=p_n+(1-p_n)\sigma_n^2$.

(b) $P(Y_n<a)\to P(Z<a)$，并且因此有 $Y_n\to$n(0,1)(回忆 $p_n\to 1$，$\sigma_n\to\infty$以及 $(1-p_n)\sigma_n^2\to\infty$).

10.7 在定理 10.1.12 的证明中，已经证明了 MLE $\hat{\theta}$是 θ 的一个渐近有效估计量. 证明，若 $\tau(\theta)$ 是 θ 的连续和可导的函数，则 $\tau(\hat{\theta})$ 是 $\tau(\theta)$ 的相合估计量和渐近有效估计量.

10.8 建立式 (10.1.6) 中的两个收敛结果，完成定理 10.1.6 的证明.

(a) 证明

$$\frac{1}{\sqrt{n}} l'(\theta_0 \mid \boldsymbol{X}) = \sqrt{n}\left[\frac{1}{n}\sum_i W_i\right],$$

其中 $W_i=\dfrac{\dfrac{\mathrm{d}}{\mathrm{d}\theta} f(X_i|\theta)}{f(X_i|\theta)}$ 有均值 0 和方差 $I(\theta_0)$. 再利用中心极限定理证明收敛于n(0，$I(\theta_0)$).

(b) 证明

$$-\frac{1}{n} l''(\theta_0 \mid \boldsymbol{X}) = \frac{1}{n}\sum_i W_i^2 - \frac{1}{n}\sum_i \frac{\dfrac{\mathrm{d}^2}{\mathrm{d}\theta^2} f(X_i|\theta)}{f(X_i|\theta)}$$

并且第一部分的均值是 $I(\theta_0)$，而第二部分的均值是 0. 应用 WLLN.

10.9 假定 X_1，…，X_n 是 iid Poisson (λ) 的. 求下列量的最佳无偏估计量：

(a) $e^{-\lambda}$，这是 $X=0$ 的概率.

(b) $\lambda e^{-\lambda}$，这是 $X=1$ 的概率.

(c) 对 (a) 和 (b) 中的这些最佳无偏估计量，计算它们相对于 MLE 的渐近相对效率. 你喜欢哪个估计？为什么？

(d) 对于可能致癌的化合物，可以通过测量暴露于这种化合物下的微生物的突

变率来进行基本检测. 试验人员把这种化合物放在 15 个皮氏培养皿中，记录到下列数目的突变群体：

$$10, 7, 8, 13, 8, 9, 5, 7, 6, 8, 3, 6, 6, 3, 5.$$

估计 $e^{-\lambda}$，即没有突变群体出现的概率，以及 $\lambda e^{-\lambda}$，也就是只有一个突变群体出现的概率. 计算最佳无偏估计和 MLE.

10.10 继续例 10.1.15 中的计算，那里考察了 $p(1-p)$ 的估计的性质.

(a) 证明，如果 $p \neq \frac{1}{2}$，MLE $\hat{p}(1-\hat{p})$ 是渐近有效的.

(b) 如果 $p=\frac{1}{2}$，用定理 5.5.26 求 $\hat{p}(1-\hat{p})$ 的极限分布.

(c) 求 Var $(\hat{p}(1-\hat{p}))$ 的精确表达式. 近似失败的原因清楚了吗?

10.11 本题将考虑例 10.1.18 中计算的细节，并作出扩展.

(a) 重新画图 10.1.1，对已知的 β 计算 ARE. (可以遵照例 A.0.7 进行计算，或者自行编程.)

(b) 验证，不论 β 已知与否，ARE $(\overline{X}, \hat{\mu})$ 都一样.

(c) 对于已知 μ 时 β 的估计，说明矩方法估计和 MLE 一样. (用 (α, β) 参数化可能更容易.)

(d) 对于 μ 未知时 β 的估计，说明矩方法估计和 MLE 不一样. 用渐近相对效率比较这些估计，并产生像图 10.1.1 那样的图，其中不同的曲线对应不同的 μ 值.

10.12 验证杂录 10.6.1 中的超有效估计量 d_n 是渐近正态的，该正态分布的方差当 $\theta \neq 0$ 时为 $v(\theta)=1$，而当 $\theta=0$ 时为 $v(\theta)=a^2$. (关于超有效估计量的更多讨论，参见 Lehmann and Casella 1998，6.2 节.)

10.13 见例 10.1.19.

(a) 验证，样本 2，4，9，12 的自助法均值和方差分别是 6.75 和 3.94.

(b) 验证原始样本的均值是 6.75.

(c) 验证，当用 n 代替 $n-1$ 去除时，均值的自助法方差以及均值之方差的通常估计是一致的.

(d) 说明如何用 $\binom{4+4-1}{4}=35$ 个不同的可能重抽样本计算自助法均值和标准误差.

(e) 对一般样本 $X_1, \cdots, X_n$ 建立 (b) 和 (c).

10.14 在下列的每一个情形中，我们都将看到参数和非参数自助法. 比较这些估计值，讨论这些方法的长处和短处.

(a) 见例 10.1.22，用非参数自助法估计 S^2 的方差.

(b) 在例 5.6.6 中，我们对于从 Poisson 样本得到的 S^2 的分布进行了参数自助

抽样. 用非参数自助法作为替代画出这个分布的直方图.

(c) 在例 10.1.18 中我们考虑了估计伽玛均值的问题. 假定我们从分布 Gamma (α, β) 抽到一个随机样本

$$0.28, 0.98, 1.36, 1.38, 2.4, 7.42.$$

用极大似然估计和自助法估计分布的均值和方差.

10.15 (a) 证明, 当 $B\to\infty$ 时, 式 (10.1.11) 中的 $\mathrm{Var}_B^*(\hat{\theta})$ 收敛到式 (10.1.10) 中的 $\mathrm{Var}^*(\hat{\theta})$.

(b) 对于固定的 B_i 和 $i=1, 2, \cdots$, 计算自助方差 $\mathrm{Var}_{B_i}^*(\hat{\theta})$. 用大数定律证明当 $m\to\infty$ 时 $(1/m)\sum_{i=1}^{m}\mathrm{Var}_{B_i}^*(\hat{\theta})\to\mathrm{Var}^*(\hat{\theta})$.

10.16 对于例 10.1.21 中的情形, 如果我们观察到 $\hat{p}=\frac{1}{2}$, 就可以从定理 5.5.26 得到方差的估计. 说明这个方差估计是 $2[\mathrm{Var}(\hat{p})]^2$.

(a) 如果 $\hat{p}=11/24$, 验证这个方差估计值为 0.00007.

(b) 用模拟的方法计算当 $n=24$ 和 $p=11/24$ 时 $\hat{p}(1-\hat{p})$ 的"准确方差". 验证这个值是 0.00529.

(c) 你认为为什么在这种情况下 Δ 方法如此糟糕? 二阶 Δ 方法会好一些吗? 自助法估计怎么样?

10.17 Efron (1982) 分析了法学院入学数据, 目的在于考察 LAST (Law School Admission Test, 法学院入学考试) 分数与一年级 GPA (grade point average, 年级平均成绩) 之间的相关性. 对于 15 个法学院, 我们有数据对 (平均 LAST, 平均 GPA):

(576, 3.39) (635, 3.30) (558, 2.81) (578, 3.03) (666, 3.44)
(580, 3.07) (555, 3.00) (661, 3.43) (651, 3.36) (605, 3.13)
(653, 3.12) (575, 2.74) (545, 2.76) (572, 2.88) (594, 2.96)

(a) 计算 LAST 分数和 GPA 之间的相关系数.

(b) 用非参数自助法估计相关系数的标准差. 用 $B=1000$ 个重抽样本, 并画出这些样本的直方图.

(c) 用参数自助法估计相关系数的标准差. 假定 (LAST, GPA) 有二元正态分布, 估计其中的 5 个参数. 然后从这个二元正态分布产生容量为 15 的 1000 个样本.

(d) 如果 (X, Y) 是二元正态的, 相关系数为 ρ, r 为样本相关系数, 则用 Δ 方法可以证明

$$\sqrt{n}(r-\rho)\to\mathrm{n}(0,(1-\rho^2)^2).$$

用这个事实估计 r 的标准差. 它与自助法估计相比如何? 给出 r 的近似概率密

度函数.

(e) Fisher z 变换是相关系数的一个**方差稳定化**变换（见习题 11.4）. 如果 (X, Y) 是二元正态的，相关系数是 ρ，r 为样本相关系数，则

$$\frac{1}{2}\left[\log\left(\frac{1+r}{1-r}\right)-\log\left(\frac{1+\rho}{1-\rho}\right)\right]$$

是近似正态的. 用这个事实给出 r 的近似概率密度函数.

(要建立 (d) 中的正态结果，需要一些乏味的矩阵计算，见 Lehmann and Casella 1998，例 6.5. (e) 中的 z 变换比 (d) 中的 Δ 方法收敛于正态的速率更快. Diaconis and Holmes 1994 就这个问题穷尽了自助抽样，枚举出所有 77,558,760 个相关系数.)

10.18 对于例 10.2.1 的情形，即如果 $X_1, \cdots, X_n$ 是 iid 的，$X_i \sim \mathrm{n}(\mu, \sigma^2)$ 的概率为 $1-\delta$，而 $X_i \sim f(x)$ 的概率为 δ，其中 $f(x)$ 是均值为 θ、方差为 τ^2 的任意密度，证明

$$\mathrm{Var}(\overline{X})=(1-\delta)\frac{\sigma^2}{n}+\delta\frac{\tau^2}{n}+\frac{\delta(1-\delta)(\theta-\mu)^2}{n}.$$

另外，证明对于具有 Cauchy 概率密度函数的污染，总导致无限方差. (提示：把这个混合模型写成多层模型. 令 $Y=0$ 的概率为 $1-\delta$，$Y=1$ 的概率为 δ，则 $\mathrm{Var}(X_i)=\mathrm{E}[\mathrm{Var}(X_i|Y)]+\mathrm{Var}(\mathrm{E}[X_i|Y])$.)

10.19 违反所用假设的另一种方式是抽样中存在相关，这可以严重影响到样本均值的性质. 假设我们在例 10.2.1 讨论的情况中引入相关，即我们观测到 $X_1, \cdots, X_n$，其中 $X_i \sim \mathrm{n}(\theta, \sigma^2)$，但这些 X_i 不再独立.

(a) 对等相关的情形，即对 $i \neq j$，$\mathrm{Corr}(X_i, X_j)=\rho$，证明

$$\mathrm{Var}(\overline{X})=\frac{\sigma^2}{n}+\frac{n-1}{n}\rho\,\sigma^2,$$

从而当 $n \to \infty$ 时 $\mathrm{Var}(\overline{X}) \not\to 0$.

(b) 如果这些 X_i 是沿时间（或距离）观测到的，有时假设相关随时间（或距离）而下降，一个特殊的模型是假设 $\mathrm{Corr}(X_i, X_j)=\rho^{|i-j|}$. 证明在这种情形下

$$\mathrm{Var}(\overline{X})=\frac{\sigma^2}{n}+\frac{2\sigma^2}{n^2}\frac{\rho}{1-\rho}\left(n-\frac{1-\rho^n}{1-\rho}\right),$$

从而当 $n \to \infty$ 时 $\mathrm{Var}(\overline{X}) \to 0$. (关于相关的其他效应见杂录 5.8.2.)

(c) (b) 中的相关结构出现在**自回归 AR (1) 模型**中，在这个模型中，我们假定 $X_{i+1}=\rho X_i+\delta_i$，$\delta_i$ iid n(0, 1). 如果 $|\rho|<1$ 并定义 $\sigma^2=1/(1-\rho^2)$，证明 $\mathrm{Corr}(X_1, X_i)=\rho^{i-1}$.

10.20 参见定义 10.2.2 中关于崩溃值的定义.

(a) 如果 $T_n=\overline{X}_n$，即样本均值，证明 $b=0$.

(b) 如果 $T_n=M_n$，即样本中位数，证明 $b=0.5$.

在敏感性上界于均值和中位数之间的一个估计量是 α－**截尾均值**，$0<\alpha<\frac{1}{2}$，定义如下：α－截尾均值 $\overline{X}_n^\alpha$ 是去掉 αn 个最小的观测值和 αn 个最大的观测值，然后取其余观测值的算术平均得到的.

(c) 如果 $T_n=\overline{X}_n^\alpha$，即样本的 α－截尾均值，$0<\alpha<\frac{1}{2}$，证明 $0<b<\frac{1}{2}$.

10.21 均值和中位数的崩溃现象在尺度参数的相应估计上也有表现. 对于样本 X_1，…，X_n，

(a) 证明样本方差 $S^2=\sum(X_i-\overline{X})^2/(n-1)$ 的崩溃值是 0.

(b) 一个稳健估计量是**中位绝对偏差**（**median absolute deviation**），或 MAD，即 $|X_1-M|$，…，$|X_n-M|$ 的中位数，其中 M 为样本中位数. 证明这个估计量的崩溃值是 50%.

10.22 本题考虑例 10.2.3 中的一些细节.

(a) 验证当 n 是奇数时，

$$P(\sqrt{n}(M_n-\mu)\leqslant a)=P\left(\frac{\sum_i Y_i-np_n}{\sqrt{np_n(1-p_n)}}\geqslant\frac{(n+1)/2-np_n}{\sqrt{np_n(1-p_n)}}\right).$$

(b) 验证 $p_n\to p=F(\mu)=1/2$ 以及

$$\frac{(n+1)/2-np_n}{\sqrt{np_n(1-p_n)}}\to-2aF'(\mu)=-2af(\mu).$$

(提示：证明 $\frac{(n+1)/2-np_n}{\sqrt{n}}$ 是一个导数的极限形式.)

(c) 解释如何从陈述

$$P(\sqrt{n}(M_n-\mu)\leqslant a)\to P(Z\geqslant-2af(\mu))$$

得到结论“$\sqrt{n}(M_n-\mu)$ 渐近于均值为 0、方差为 $1/[2f(\mu)]^2$ 的正态分布”.

(注意，仅当 p_n 不依赖于 n 时才能直接应用 CLT. 当 p_n 依赖于 n 时，需要做更多的工作才能严格地得到极限正态分布的结论. 在这些工作后，所得结果会如所期望.)

10.23 在本题中，我们将进一步探讨中位数相对于均值的 ARE，即 ARE $(M_n,\ \overline{X})$.

(a) 验证例 10.2.4 中给出的三个 ARE.

(b) 证明 ARE $(M_n,\ \overline{X})$ 不受尺度变化的影响，即无论所研究的概率密度函数是 $f(x)$ 还是 $(1/\sigma)f(x/\sigma)$，该渐近相对效率不变.

(c) 当所研究的分布是自由度为 ν 的学生 t 分布时，计算 ARE $(M_n,\ \overline{X})$，其中 $\nu=3,\ 5,\ 10,\ 25,\ 50,\ \infty$. 关于 ARE 和分布的尾部能得到什么结论?

(d) 当所研究的分布为

$$X\sim\begin{cases}\mathrm{n}(0,\ 1) & \text{以概率 } 1-\delta\\ \mathrm{n}(0,\ \sigma^2) & \text{以概率 } \delta\end{cases}$$

时计算 ARE $(M_n,\ \overline{X})$. 对于 δ 和 σ 的一个范围，计算 ARE. 关于均值和中位数的相对表现你有什么结论?

10.24 假定 θ_0 满足 $\mathrm{E}_{\theta_0}\psi(X-\theta_0)=0$，证明式（10.2.4）和式（10.2.5）蕴涵式（10.2.6）.

10.25 如果 f（x）是关于 0 对称的概率密度函数，而 ρ 是一个对称函数，证明 $\int\psi(x-\theta)f(x-\theta)\,\mathrm{d}x=0$，其中 $\psi=\rho'$ 是一个奇函数. 由此可以得到，如果 X_1，…，X_n 是 iid 的，来自于概率密度函数 $f(x-\theta)$，且 $\hat{\theta}_M$ 是 $\sum_i\rho(x_i-\theta)$ 的最小点，则 $\hat{\theta}_M$ 是渐近正态的，且正态分布的均值为 θ 的真值.

10.26 这里我们考虑例 10.2.6 中结论的一些细节.

(a) 验证 $\mathrm{E}_\theta\psi'(X-\theta)$ 和 $\mathrm{E}_\theta[\psi(X-\theta)]^2$ 的表达式，从而验证 $\hat{\theta}_M$ 的方差公式.

(b) 当计算 ψ' 的期望值时，我们注意到 ψ 是不可导的，但我们可以使用可导的部分. 另一个方法是，认识到 ψ 的期望值是可导的，以及式（10.2.5）中的结论，我们可以得到

$$\frac{1}{n}\sum_{i=1}^{n}\psi'(x_i-\theta_0)\to\frac{\mathrm{d}}{\mathrm{d}\theta}\mathrm{E}_{\theta_0}\psi(X-\theta)\Big|_{\theta=\theta_0}.$$

证明这与方程（10.2.5）的结论相同.

10.27 考虑例 10.6.2 中的情形.

(a) 证明 $IF(\overline{X},x)=x-\mu$.

(b) 如果 $P(X\leqslant m)=1/2$ 或 $m=F^{-1}(1/2)$，则对于中位数，我们有 $T(F)=m$. 如果 $X\sim F_\delta$，证明

$$P(X\leqslant a)=\begin{cases}(1-\delta)F(a) & \text{如果 } x>a\\ (1-\delta)F(a)+\delta & \text{其他}\end{cases}$$

从而

$$T(F_\delta)=\begin{cases}F^{-1}\left(\dfrac{1}{2(1-\delta)}\right) & \text{如果 } x>F^{-1}\left(\dfrac{1}{2(1-\delta)}\right)\\ F^{-1}\left(\dfrac{1/2-\delta}{1-\delta}\right) & \text{其他}\end{cases}$$

(c) 证明

$$\frac{1}{\delta}\left[F^{-1}\left(\frac{1}{2\ (1-\delta)}\right)-F^{-1}\left(\frac{1}{2}\right)\right]\to\frac{1}{2f\ (m)},$$

并完成计算 IF（M，x）的论述.

（提示：记 $a_\delta=F^{-1}\left(\frac{1}{2\ (1-\delta)}\right)$，再说明极限为 $a'_\delta\mid_\delta=0$，而这个量可以用隐函数微分法和事实 $(1-\delta)^{-1}=2F\ (a_\delta)$ 来计算.）

10.28 证明，如果 ρ 由式（10.2.2）定义，则 ρ 和 ρ' 都是连续的.

10.29 由式（10.2.9）我们知道，一个 M—估计量绝不可能比极大似然估计量更有效. 但是我们知道什么时候它同样有效.

(a) 证明如果我们选择 $\psi(x-\theta)=cl'(\theta|x)$，其中 l 是对数似然，而 c 是常数，则式（10.2.9）是等式.

(b) 对于下列分布，验证相应的 ψ 函数给出渐近有效的 M—估计量.

(i) 正态：$f(x)=\mathrm{e}^{-x^2/2}/(\sqrt{2\pi})$，$\psi(x)=x$

(ii) 罗吉斯蒂克：$f(x)=\mathrm{e}^{-x}/(1+\mathrm{e}^{-x})^2$，$\psi(x)=\tanh(x)$，其中 tanh（$x$）是**双曲正切**

(iii) Cauchy：$f(x)=[\pi(1+x^2)]^{-1}$，$\psi(x)=2x/(1+x^2)$

(iv) 最小信息分布：

$$f(x)=\begin{cases}C\mathrm{e}^{-x^2/2} & |x|\leqslant c\\ C\mathrm{e}^{-c|x|+c^2/2} & |x|>c\end{cases}$$

$\psi(x)=\max\{-c,\min(c,x)\}$，$C$ 和 c 都是常数.

（更多的细节见 Huber 1981，3.5 节.）

10.30 对于 M—估计量，ψ 函数与崩溃值之间有着联系. 此中细节相当复杂（Huber 1981，3.2 节），但可以总结如下：若 ψ 是有界函数，则对应的 M—估计量的崩溃值由

$$b^*=\frac{\eta}{1+\eta},\text{其中 }\eta=\min\left\{-\frac{\psi(-\infty)}{\psi(\infty)},-\frac{\psi(\infty)}{\psi(-\infty)}\right\}$$

给出.

(a) 计算习题 10.29 中有效 M—估计量的崩溃值. 哪个估计量既是有效的又是稳健的?

(b) 计算下列 M—估计量的崩溃值：

(i) 由式（10.2.1）给出的 Huber 估计量.

(ii) Tukey 双加权：$\psi(x)=x(c^2-x^2)$对$|x|\leqslant c=0$，对其他 x；c 是常数.

(iii) Andrew 的正弦波：$\psi(x)=c\sin(x/c)$对$|x|\leqslant c\pi$，$=0$ 对其他 x.

(c) 当研究的分布是（i）正态和（ii）双边指数时，计算（b）中的估计量相对于 MLE 的 ARE.

10.31 从多个总体收集到的二项数据常常用**列联表**表现出来. 在两个总体的情形，列联表形如

	总体 1	总体 2	总和
成功	S_1	S_2	$S=S_1+S_2$
失败	F_1	F_2	$F=F_1+F_2$
总和	n_1	n_2	$n=n_1+n_2$

其中总体 1 是 binomial（n_1，p_1），有 S_1 个成功，F_1 个失败；总体 2 是 binomial（n_2，p_2），有 S_2 个成功，F_2 个失败. 经常感兴趣的一个假设是

$$H_0：p_1=p_2 \text{ 对 } H_1：p_1\neq p_2.$$

（a）说明可以基于统计量

$$T=\frac{(\hat{p}_1-\hat{p}_2)^2}{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)(\hat{p}(1-\hat{p}))}$$

进行检验，其中$\hat{p}_1=S_1/n_1$，$\hat{p}_2=S_2/n_2$，$\hat{p}=(S_1+S_2)/(n_1+n_2)$. 另外，证明当 n_1，$n_2\to\infty$时，T 的分布趋于 χ_1^2.（这是所谓**独立性 χ^2 检验**的一个特殊情形.）

（b）测量与 H_0 相违背的另一个方法是计算**期望频数表**. 这个表的构造方法是，在给定边缘总和的条件下，根据 $H_0：p_1=p_2$ 填充表格，即

期望频数

	1	2	总和
成功	$\frac{n_1S}{n_1+n_2}$	$\frac{n_2S}{n_1+n_2}$	$S=S_1+S_2$
失败	$\frac{n_1F}{n_1+n_2}$	$\frac{n_2F}{n_1+n_2}$	$F=F_1+F_2$
总和	n_1	n_2	$n=n_1+n_2$

用这个期望频数表中的所有格子，计算统计量 T^*：

$$T^*=\sum\frac{(\text{观测频数}-\text{期望频数})^2}{\text{期望频数}}$$

$$=\frac{\left(S_1-\frac{n_1S}{n_1+n_2}\right)^2}{\frac{n_1S}{n_1+n_2}}+\cdots+\frac{\left(F_2-\frac{n_2F}{n_1+n_2}\right)^2}{\frac{n_2F}{n_1+n_2}}$$

用代数运算证明 $T^*=T$，因此 T^* 是渐近 χ^2 的.

（c）检验 p_1 和 p_2 相等的另一个统计量是

$$T^{**}=\frac{\hat{p}_1-\hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}},$$

证明，在 H_0 下，T^{**} 是渐近 n（0，1）的，因此，其平方渐近于 χ_1^2. 进一步，证

明 $(T^{**})^2 \neq T^*$.

(d) 在什么情况下一个统计量比另一个更好?

(e) 19 世纪后期 Joseph Lister 进行了一个著名的医学试验. 当时手术的死亡率很高, 而 Lister 猜测使用抗感染剂石炭酸可能有助于降低死亡率. 在几年期间, Lister 做了 75 例截肢手术, 有的用了石炭酸, 有的没有用. 数据如下:

		用了石炭酸? 用	没用
患者存活?	存活	34	19
	没有	6	16

用这些数据检验石炭酸的使用是否与患者死亡有关.

10.32 (a) 设 $(X_1, \cdots, X_n) \sim$multinomial $(m, p_1, \cdots, p_n)$. 考虑检验 $H_0: p_1 = p_2$ 对 $H_1: p_1 \neq p_2$. 一个常用的检验是所谓 **McNemar 检验**, 当

$$\frac{(X_1 - X_2)^2}{X_1 + X_2} > \chi^2_{1,\alpha}$$

时拒绝 H_0. 证明这个检验统计量有形式

$$\sum_{1}^{n} \frac{(\text{观测频数} - \text{期望频数})^2}{\text{期望频数}},$$

其中 X_i 为观测到的格子频数, 期望格子频数为在假设 $p_1 = p_2$ 下 mp_i 的 MLE.

(b) McNemar 检验经常用在下列类型的问题中. 问调查对象是否同意某个说法, 然后让他们读到一些关于这个说法的信息, 再问他们是否同意. 把每种情况的响应数量总结在下面的 2×2 表中:

		前 同意	不同意
后	同意	X_3	X_2
	不同意	X_1	X_4

假设 $H_0: p_1 = p_2$ 是说从同意变到不同意的人在所有人中的比例与从不同意变到同意的人在所有人中的比例相同. 可能检验的另一个假设是原来同意的人中改变态度的人的比例与原来不同意的人中改变态度的人的比例相同. 用条件概率表述这个假设, 并说明它不同于上述的 H_0. (这个假设可以用习题 10.31 中的 χ^2 检验来进行.)

10.33 完成定理 10.3.1. 用定理 10.1.12 和 Slutsky 定理 (定理 5.5.17) 证明 $(\theta - \hat{\theta})/\sqrt{-l''(\hat{\theta}|\boldsymbol{x})} \to \mathrm{n}(0,1)$, 从而 $-2\log\lambda(\boldsymbol{X}) \to \chi^2_1$.

10.34 为检验 $H_0: p = p_0$ 对 $H_1: p \neq p_0$, 假定我们观测到 $X_1, \cdots, X_n$ iid Bernoulli (p).

(a) 推导出 $-2\log\lambda(\boldsymbol{x})$ 的表达式, 这里 $\lambda(\boldsymbol{x})$ 为 LRT 统计量.

(b) 像例 10.3.2 那样，模拟 $-2\log\lambda(\boldsymbol{x})$ 的分布，并把结果与 χ^2 近似进行比较.

10.35 设 X_1，…，X_n 为来自总体 n(μ，σ^2) 的随机样本.

(a) 如果 μ 未知而 σ 已知，证明 $Z=\sqrt{n}\,(\overline{X}-\mu_0)/\sigma$ 是检验 $H_0:\mu=\mu_0$ 的 Wald 统计量.

(b) 如果 σ 未知而 μ 已知，求检验 $H_0:\sigma=\sigma_0$ 的一个 Wald 统计量.

10.36 设 X_1，…，X_n 为来自总体 gamma (α，β) 的随机样本. 假定 α 已知而 β 未知. 考虑检验 $H_0:\beta=\beta_0$.

(a) β 的 MLE 是什么?

(b) 推导检验 H_0 的一个 Wald 统计量，在统计量的分子和分母中都用 MLE.

(c) 重复 (b)，但在标准误处用样本标准差.

10.37 设 X_1，…，X_n 为来自总体 n (μ，σ^2) 的随机样本.

(a) 如果 μ 未知而 σ 已知，证明 $Z=\sqrt{n}(\overline{X}-\mu_0)/\sigma$ 是检验 $H_0:\mu=\mu_0$ 的记分统计量.

(b) 如果 σ 未知而 μ 已知，求检验 $H_0:\sigma=\sigma_0$ 的一个记分统计量.

10.38 设 X_1，…，X_n 为来自总体 gamma (α，β) 的随机样本. 假定 α 已知而 β 未知，考虑检验 $H_0:\beta=\beta_0$. 求检验 H_0 的记分统计量.

10.39 扩充例 10.3.7 中所做的比较.

(a) 基于 Huber M—估计量的另一个检验使用基于式 (10.3.6) 的方差估计. 考察这个检验统计量的表现，评论它作为式 (10.3.8) 或式 (10.3.9) 的替代的长短之处.

(b) 基于 Huber M—估计量的另一个检验使用自助法方差估计. 考察这个检验统计量的表现.

(c) $\hat{\theta}_M$ 的一个稳健竞争者是中位数. 考察基于中位数的位置参数检验的表现.

10.40 在例 10.4.5 中我们看到，由 Poisson 假设和中心极限定理，得到以下事实并由此导出一个近似区间：

$$\frac{\overline{X}-\lambda}{\sqrt{\lambda/n}}\to \mathrm{n}(0,1).$$

证明这个逼近按照 Wilks (1938) 的说法是最优的，即证明

$$\frac{\overline{X}-\lambda}{\sqrt{\lambda/n}}=\frac{\dfrac{\partial}{\partial\lambda}\log L(\lambda\,|\,\boldsymbol{X})}{\sqrt{-E_\lambda\left(\dfrac{\partial^2}{\partial\lambda^2}\log L(\lambda\,|\,\boldsymbol{X})\right)}}.$$

10.41 设 X_1，…，X_n 是 iid 的，服从负二项分布 NB (r，p). 我们要构造负二项分布参数的近似置信区间.

(a) 计算 Wilks 近似 (10.4.1)，并说明如何用这个表达式形成置信区间.

(b) 求负二项分布的均值的近似 $1-\alpha$ 置信区间，并说明如何对求出的区间做连续性校正.

(c) 用负二项分布作为习题 9.23 中蚜虫数据的模型，用 (b) 中的结果构造近似 90%的置信区间. 把这个区间与习题 9.23 中基于 Poisson 分布的置信区间进行比较.

10.42 证明，对于任何固定的水平 α，式 (10.4.5) 等价于式 (9.2.7) 中的最高似然区域，它们产生同样的置信集合.

10.43 在例 10.4.7 中，对 Wald 区间进行了两项修改.

(a) 在 $y=0$ 时，上区间端点变为 $1-(\alpha/2)^{1/n}$，在 $y=n$ 时，下区间端点变为 $(\alpha/2)^{1/n}$. 说明选择这些端点的合理性. (提示：见 9.2.3 节.)

(b) 第 2 个修改是把区间截断，使之在 $[0, 1]$ 之内. 说明这个变化，连同 (a) 中的另一个变化一起，改进了原来的 Wald 区间.

10.44 Agresti and Coull (1998) "强烈推荐"对于二项参数使用记分区间，但他们也关心像式 (10.4.7) 那样的区间在统计的基础教材里是否有点令人生畏. 为得到一个简单、合理的二项区间，他们建议对 Wald 区间进行如下修改：增加 2 个成功和 2 个失败，然后用原来的 Wald 公式 (10.4.8)，即用 $\tilde{p}=(y+2)/(n+4)$ 代替 $\hat{p}=y/n$. 用长度以及覆盖概率，比较这个区间与二项记分区间. 你同意"它是记分区间的一个合理替换"这个说法吗?

(Samuels and Lu 1992 曾建议基于样本容量对 Wald 区间做另一种修改. Agresti and Caffo 2000 把这个修改扩充到两样本问题.)

10.45 用例 10.4.6 中给出的连续性校正求解近似二项置信区间的端点. 证明这个区间比没有连续性校正的要宽，并且连续性校正置信区间有一致更高的覆盖概率. (事实上，未校正的区间的覆盖概率不保持 $1-\alpha$，对于某些参数值，覆盖概率低于这个水平. 校正后的区间对于所有的参数保持大于 $1-\alpha$ 的覆盖概率.)

10.46 扩展例 10.4.8 中的比较.

(a) 产生类似于表 10.4.2 的表，考察基于中位数的位置参数的置信区间的稳健性. (基于均值的置信区间的结果在表 10.4.1 中给出.)

(b) 另一个基于 Huber M-估计量的置信区间使用自助法计算的方差. 考察这个区间的稳健性.

10.47 设 $X_1, \cdots, X_n$ 是 iid 的，服从负二项分布 NB (r, p).

(a) 完成例 10.4.9 的细节，即对于小的 p 值，区间

$$\left\{p: \frac{\chi^2_{2nr,1-\alpha/2}}{2\sum x} \leqslant p \leqslant \frac{\chi^2_{2nr,\alpha/2}}{2\sum x}\right\}$$

是一个近似 $1-\alpha$ 置信区间.

(b) 说明要得到最短长度的 $1-\alpha$ 置信区间，如何选择端点.

10.48 对于 Fieller 置信集合的情形（见杂录 9.5.3），即设随机样本 $(X_1, Y_1), \cdots, (X_n, Y_n)$ 来自于参数为 $(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ 的二元正态分布，求 $\theta = \mu_Y/\mu_X$ 的近似置信区间. 用例 5.5.27 中的近似矩计算，并应用中心极限定理.

10.6 杂录

10.6.1 超有效性

虽然定理 7.3.9 中的 Cramér－Rao 下界是名副其实的方差下界，但定义 10.1.11 和定理 10.1.6 中渐近方差的下界却可能被突破. 打破定义 10.1.11 中的界的一个例子由 Hodges（见 LeCam 1953）给出.

如果 $X_1, \cdots, X_n$ 是 iid $n(\theta, 1)$ 的，θ 的无偏估计量的 Cramér－Rao 下界是 $v(\theta) = 1/n$. 估计量

$$d_n = \begin{cases} \overline{X} & \text{如果 } |\overline{X}| \geqslant 1/n^{1/4} \\ a\overline{X} & \text{如果 } |\overline{X}| < 1/n^{1/4} \end{cases}$$

满足：依分布有

$$\sqrt{n}(d_n - \theta) \to n[0, v(\theta)],$$

其中 $v(\theta) = 1$ 当 $\theta \neq 0$，$v(\theta) = a^2$ 当 $\theta = 0$. 如果 $a < 1$，则不等式 (7.3.5) 在 $\theta = 0$ 不成立.

像 d_n 这样的估计量称为**超有效估计量**；虽然在某些一般的情况下可以构造出这种估计量，然而在现实中，在更大的程度上这种估计量是一种理论上的怪物. 这是因为使得方差低于下界的 θ 的值的集合是一个 Lebesgue（勒贝格）0 测度集合. 然而，超有效估计量的存在性提醒我们，在建立估计量的性质时，总要小心检查所做的假设（也提醒我们在一般情况下要小心!）.

10.6.2 适当的正则性条件

“在适当的正则条件下”这句话有乱用之嫌，因为只要给予充分的假设我们大概可以证明任何想要的结果. 然而，“正则条件”常是一组技术性很强的、相当乏味的、也是在大多数合理的问题中经常能够得到满足的条件. 但它们又是必需的，因此我们应该与之打交道. 为了完整起见，我们给出一组正则条件，用这组条件足以严格建立定理 10.1.6 和定理 10.1.12. 这组条件不是最普遍的条件，但对于许多应用而言是足够普遍的（一个值得提到的例外是 MLE 位于参数空间的边界的情形）. 需要事先说明的是，下面的内容不是给弱者的，略过这些内容对于理解定理而言无伤大局.

这些条件主要与密度的可微性及微分与积分的可交换性有关（像定理 7.3.9 中的条件一样）. 更多的细节和一般性结果见 Stuart, Ord, and Arnold (1999, 第 18 章), Ferguson (1996, 第 4 部分), 或 Lehmann and Casella (1998, 6.3 节).

下列假定足以保证定理 10.1.6, 即 MLE 的相合性.

(A1) 我们观测到 $X_1, \cdots, X_n$, $X_i \sim f(x|\theta)$ 是 iid 的.

(A2) 参数是**可识别的**, 即如果 $\theta \neq \theta'$, 则 $f(x|\theta) \neq f(x|\theta')$.

(A3) 各个密度 $f(x|\theta)$ 有共同支撑集, 并且 $f(x|\theta)$ 关于 θ 可导.

(A4) 参数空间 Ω 包含一个开集 ω, 以真参数值 θ_0 为该开集的一个内点.

下列两个假定, 连同 (A1) ～ (A4) 一起, 保证了定理 10.1.12, 即 MLE 的渐近正态性和渐近效率.

(A5) 对于每个 $x \in \mathcal{X}$, 密度 $f(x|\theta)$ 关于 θ 是三阶可导的, 其三阶导数是 θ 的连续函数, 并且 $\int f(x|\theta)\mathrm{d}x$ 可以在积分号下微分三次.

(A6) 对任何 $\theta_0 \in \Omega$, 存在一个正数 c 和一个函数 $M(x)$ (二者都可以依赖于 θ_0) 使得

$$\left|\frac{\partial^3}{\partial\theta^3}\log f(x|\theta)\right| \leqslant M(x), \text{ 对于所有 } x \in \mathcal{X},\ \theta_0 - c < \theta < \theta_0 + c,$$

以及 $\mathrm{E}_{\theta_0}|M(X)| < \infty$.

10.6.3 再谈自助法

理论

自助法背后的理论相当复杂, 基于 **Edgeworth 展开**. Edgeworth 展开是分布函数围绕正态分布的展开 (与 Taylor 级数同理). 例如, 对于 $X_1, \cdots, X_n$ iid, 有密度 f, 均值 μ 和方差 σ^2, $\dfrac{\sqrt{n}(\overline{X}-\mu)}{\sigma}$ 的累积分布函数的一个 Edgeworth 展开是 (Hall 1992, 方程 2.17)

$$P\left(\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} \leqslant \omega\right) = \Phi(\omega) + \phi(\omega)\left[\frac{-1}{6\sqrt{n}}\kappa(\omega^2-1) + R_n\right]$$

其中 nR_n 有界, Φ 和 ϕ 分别是标准正态分布的分布和密度函数, $\kappa = \mathrm{E}(X_1-\mu)^3$ 是偏度. 展开式中的第一项是通常的正态逼近, 由于增加了更多的项, 展开式变得更加精确了.

令人惊奇的是, 在某些情况下自助法自动精确到展开式中的第二项 (因此达到"二阶"精确). 这个结果并不总是成立, 但在对一个枢轴量使用自助法时确实成立. 有关自助法的 Edgeworth 展开, Hall (1992) 中有透彻的讨论, 也可见 Shao and Tu (1995).

实践

我们只把自助法用在了计算标准误上，但它还有许多其他的应用，其中最受欢迎的是置信区间的构造. 对于不同的情形，自助法也发展出许多不同的变化. 特别，用自助法对付相关数据不失为妙手. Efron and Tibshirani (1993) 介绍了自助法的许多应用和其他内容.

局限性

虽然自助法或许是近年来统计方法的最重要的进展，但它也不是没有局限性和非议的. 除了独立同分布抽样和枢轴量的情形外，自助法未必自动有用，但仍然可能是非常有用的. 关于这些问题的有趣的讨论，见 LePage and Billard (1992) 或 Young (1994).

10.6.4 影响函数

关于灾难性后果的一个度量是**影响函数**，它考虑分布的性质，同时也度量一个异常观测值产生的效果. 影响函数可以解释为导数，由此又可以得到一些有趣的结果.

一个统计量的影响函数实际上可以用与它对应的总体量来计算. 例如，样本均值的影响函数用总体均值来计算，因为它测量了总体扰动的影响. 类似地，样本中位数的影响函数用总体中位数来计算. 为了用适当的方式表达这个概念，要把一个估计量视为对累积分布函数 F 或经验累积分布函数 F_n 进行运算的函数. 这种以其他函数作为自变量的函数叫做**泛函**.

注意，对于一个样本 X_1，…，X_n 来说，关于样本的知识等价于关于经验累积分布函数 F_n 的知识，因为 F_n 在每个 X_i 处有大小为 $1/n$ 的跳跃. 因而，一个统计量 $T=T(X_1, \cdots, X_n)$ 可以等价地写为 $T(F_n)$. 如此，我们就可以把相应的总体量记为 $T(F)$.

定义 10.6.1 对于来自累积分布函数为 F 的总体的样本 X_1，…，X_n，统计量 $T=T(F_n)$ 在点 x 处的**影响函数**为

$$IF(T,x)=\lim_{\delta\to 0}\frac{1}{\delta}[T(F_\delta)-T(F)],$$

其中 $X\sim F_\delta$ 如果

$$X\sim\begin{cases}F & \text{依概率 } 1-\delta\\ x & \text{依概率 } \delta\end{cases}$$

也就是说，F_δ 是 F 和点 x 的混合.

例 10.6.2 (均值和中位数的影响函数) 假设我们有一个总体，它有连续的累积分布函数 F 和概率密度函数 f. 以 μ 记其总体均值，$\overline{X}$ 记样本均值，$T(\cdot)$ 是计算总体均值的泛函. 则 $T(F_n)=\overline{X}$，$T(F)=\mu$，并且

$$T(F_\delta)=(1-\delta)\mu+\delta x,$$

所以 $IF(\overline{X},x)=x-\mu$，当 x 增大时，它对于$\overline{X}$的影响也增大.

对于中位数，我们有（见习题 10.27）

$$IF(M,x)=\begin{cases}\dfrac{1}{2f(m)} & \text{如果 } x>m\\ -\dfrac{1}{2f(m)} & \text{其他.}\end{cases}$$

因此，与均值不同，中位数的影响函数是有界的. ‖

为什么有界影响函数是重要的呢？为回答这个问题，我们来看 M—估计量的影响函数，均值和中位数都是 M-估计的特殊情况.

令$\hat{\theta}_M$ 为由方程 $\sum_i\psi(x_i-\theta)=0$ 解得的 M—估计量，其中 $X_1,\cdots,X_n$ 是 iid 的，有累积分布函数 F. 在 10.2.2 节我们看到$\hat{\theta}_M$ 是满足 $\mathrm{E}_{\theta_0}\psi(X-\theta_0)=0$ 的 θ_0 的相合估计量. $\hat{\theta}_M$ 的影响函数是

$$IF(\hat{\theta}_M,x)=\frac{\psi(x-\theta_0)}{-\int\psi'(t-\theta_0)f(t)\mathrm{d}t}=\frac{\psi(x-\theta_0)}{-\mathrm{E}_{\theta_0}(\psi'(X-\theta_0))}.$$

现在如果我们回想式（10.2.6），就会看到影响函数平方的期望给出了$\hat{\theta}_M$ 的渐近方差，即依概率

$$\sqrt{n}(\hat{\theta}_\mathrm{M}-\theta_0)\to\mathrm{n}(0,\mathrm{E}_{\theta_0}[IF(\hat{\theta}_M,X)]^2).$$

因此，影响函数与渐近方差有直接的联系.

10.6.5 自助法区间

在 10.1.4 节中我们看到，自助法是获得任何一个统计量的标准误的简单而又具有一般性的方法. 在计算这些标准误时，我们实际上构造了一个统计量的分布，即**自助分布**. 这就自然引出了一个问题：有用自助法构造置信区间的简单而又一般的方法吗？自助法的确可以用来构造很好的置信区间，但是，在计算置信区间时，计算标准误时享受到的简单性没有了.

使用自助分布的百分位数，或者对 t 统计量（枢轴量）应用自助法，看上去都有作为一般方法应用的潜在可能. 然而，Efron and Tibshirani（1993，13.4 节）指出“一般来说这两种区间都不好”. Hall（1992，第 3 章）倾向于使用 t 统计量方法，并指出对一个枢轴量应用自助法，一般来说是比较好的.

百分位数和百分位数 t 区间只是自助置信区间家族中的沧海一粟，很多方法都有出色的表现. 然而，我们无法用简单的办法做一个总结；不同的问题需要不同的方法.

10.6.6 稳健区间

虽然我们在 10.2 节中就点估计量的稳健性讨论了一些细节，但除例 10.3.7 和例 10.4.8 以外，关于稳健检验和置信区间却没有深入细节问题. 这并不是由于不重要而是由于需要更多的篇幅.

当我们考察点估计量的稳健性质时，主要关心当假定有偏离（包括小偏离和大偏离）时估计量的表现. 对于检验和区间，我们也关心同样的问题，并期望稳健的点估计量能够导致稳健的检验和区间. 特别，我们还要求当对于假定的偏离在一定范围内时，稳健检验能够保持功效，而稳健区间能够保持覆盖概率. 这个要求能够达到，因为一个检验的功效函数与用来构造这个检验的点估计量的影响函数相联系（见 Staudte and Sheather 1990，5.3.3 节）. 当然这也立即意味着相应的区间估计的覆盖性质也与影响函数相联系.

Boos (1992) 通过估计方程和记分检验，对于稳健检验做了一个很好的介绍. Staudte and Sheather (1990)，Hettmansperger and McKean (1998) 以及当代的经典著作 Huber (1981) 也都是很好的参考书.