则称估计量 $\hat{\theta}(X_1,\cdots,X_n)$ 有渐近正态性. ♣

定义6.6也常常记为

$$(\hat{\theta}(X_1,\cdots,X_n)-\theta)/\sigma_n(\theta) \xrightarrow{\mathcal{L}} N(0,1).$$

估计量是否具有渐近正态性是点估计优良性的一个重要标志, 因为一般而言估计量 $\hat{\theta}(X_1,\cdots,X_n)$ 的分布很难得到, 而我们在作统计推断 (区间估计和假设检验) 时要以估计量的分布为基础. 渐近正态性则提供了估计量的一个近似分布, 利用它我们才能完成相关的统计推断.

在一般条件下, 矩估计和最大似然估计都会有渐近正态性, 例如总体均值 μ 的点估计 $\overline{X}_n$, 由林德伯格-莱维中心极限定理, 我们有

$$\sqrt{n}(\overline{X}_n-\mu)/\sigma \xrightarrow{\mathcal{L}} N(0,1).$$

对于参数函数 $g(\theta)$ 估计量的相合性和渐近正态性也可以作类似的考虑. 关于这些问题的讨论可以参见**?**.

6.6 扩展阅读：德军坦克问题

在二战时期, 首先出现在意大利的德军五号坦克（豹式坦克）因其速度快, 装备有 75 mm/L70 长管战车炮, 对盟军造成了很大的损失. 美军的谢尔曼坦克对阵德军的三号和四号坦克表现不俗, 但不敌德军五号坦克, 需要多辆谢尔曼坦克才能击败一辆德军五号坦克. 就在诺曼底登陆 (D 日) 不久前, 有消息称德国有大量的五号坦克. 为了确定这种说法的真实性, 盟军需要估计德军生产该五号坦克的数量. 具体有两种不同的做法：第一种是使用传统的情报窃取, 第二种是使用统计学的方法来进行估计. 盟军最初通过间谍、解码和逼供等传统手段收集信息, 得出的结论是从 1940 年 6 月到 1942 年 9 月期间每月生产 1400 辆坦克. 但这个数字显然与事实不符, 因为在长达 8 个月的斯大林格勒战役中轴心国使用了 1200 辆坦克, 也就是说 1400 的数字高的离谱. 最终盟军找到了关键线索：序列号. 盟军发现每辆被俘坦克上都有一个独一无二的序列号. 通过仔细观察, 他们发现从这些序列号可以推算出坦克的生产量. 这些序列号主要是变速箱的号码. 假设坦克是按从 1 到 N 顺序生产的, 那么如何使用统计方法估算坦克总数呢?

统计估计方法

在这个问题中, 我们观测到 k 辆坦克. 直观上可以使用观测到的最大序列号 m 来估计 N. 这是因为在 N 辆坦克中随机抽查 k 辆, 记观测到最大序列号为 $X_{(k)}$, 则 $P(X_{(k)}=m)$ 的为

$$P(X_{(k)}=m)=\binom{m-1}{k-1}\Big/\binom{N}{k}$$

给定总数 N 和样本量 k, 则

$$EX_{(k)}=\mu=\sum_{m=k}^{N} m\binom{m-1}{k-1}\Big/\binom{N}{k}=\frac{k(N+1)}{k+1}.$$

因此, 可以得到

$$N = \mu(1 + k^{-1}) - 1.$$

于是由矩估计方法可知 N 的矩估计量为

$$\hat{N} = X_{(k)}(1 + k^{-1}) - 1 = m(1 + k^{-1}) - 1.$$

可以证明, $\hat{N}$ 是 N 的最小方差无偏估计. 其方差为

$$Var(\hat{N}) = (1 + k^{-1})^2 Var(X_{(k)}) = \frac{1}{k}\frac{(N-k)(N+1)}{k+2} \approx \frac{N^2}{k^2} \text{当} k \ll N.$$

从而标准差约为 N/k.

模拟研究

统计方法得到的估计量有多好呢? 我们通过蒙特卡洛随机模拟来说明. 考虑 N 从 100 到 10^6, 我们随机观测到 $k = 20$ 个数, 分别计算 $\hat{N}$ 和估计的偏差, 样本最大值 m 等量. 最后通过图示来展示结果. 下面对不同的 N, 从中随机抽取 $k = 20$ 个数, 记录抽取的数并分别计算 $\hat{N}$ 和估计的偏差, 样本最大值 m 等量. 重复此过程 50 次.

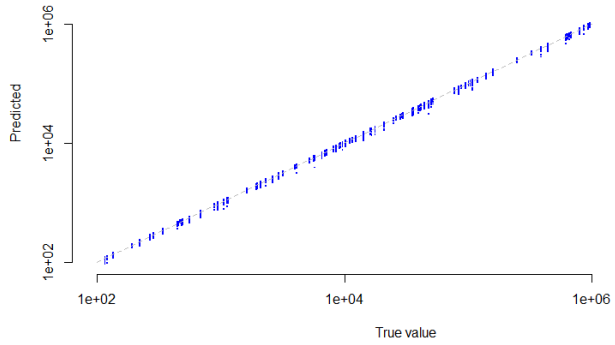


图 6.2: 预测值和真实值图



实验 15：德国坦克问题

可以看出, 预测值和真实值非常好的匹配在对角线附近. 那么残差是如何波动的呢? 很明显, 残差随着 N 的增大而扩大. 注意到方差约为 N/k, 在上面的模拟研究中, 我们固定了 k, 而增加 N, 因此随着 N 增大, 自然方差越来越大, 相应地残差的波动也越来越大.

从这里例子中, 我们可以看出统计方法非常有效的解决了这个棘手的问题. 使用统计方法来确定德军坦克的生产速度, 得出的结论是：1940 年夏天到 1942 年秋天之间, 德军每月生产 246 辆坦克. 而根据战后德军的内部数据, 德军实际产量为 245 辆, 只比估算值少一辆！那么本例中, 如果使用最大似然估计方法, 会怎么样呢?