概率论与数理统计

庄玮玮

weizh@ustc.edu.cn

安徽 合肥

2020 年 4 月

第 五 章　　参数估计

§5.1 总体和参数

现代生活是建立在数据之上的，没有数据，一切很难想象. 统计学是利用数据解释自然规律的科学，内容包括如何收集和分析数据.

基于统计学的数据处理方法称为统计方法. 在科学研究、工农业生产、新产品开发、产品质量的提高乃至政治、教育、社会科学等各个领域，使用统计方法和不使用统计方法获得的结果是大不相同的. 只要统计方法使用得当，就能够收到事半功倍的效果.这也是统计学能随着科学技术和国民经济的发展而快速发展的重要原因.

统计学没有自己的基于试验的专门研究对象，但是可以为物理学家、化学家、医生、社会学家、心理学家等提供一套研究问题的有效方法. 这套方法可以帮助各个领域的研究工作者更快地获得成功.

日常生活中我们总是自觉或不自觉地和总体与样本打交道. 买桔子时, 先要尝尝这批桔子甜不甜. 这时称这批桔子是一个总体, 单个的桔子是个体.

在仅关心桔子的甜度时, 我们可以称单个桔子的甜度是个体, 称所有桔子的甜度为总体. 这样就可以把桔子的甜不甜数量化.

要了解一批桔子的甜度情况, 你只需品尝一两个, 然后通过这一两个桔子的甜度判断这批桔子的甜度. 这就是用个体推断总体.

为把上面的实际情况总结出来, 需要引入一些术语.

A. 总体、个体和总体均值

在统计学中，我们把所要调查对象的全体叫做总体(population)，把总体中的每个成员叫做个体 (individual).

总体中的个体总可以用数量表示. 为了叙述的简单和明确，我们把个体看成数量，把总体看成数量的集体. 我们要调查的是总体的性质.

总体中个体的数目有时是确定的，有时较难确定，但是往往并不影响总体的确定，也不影响问题的解决. 在判断一批桔子甜不甜时，你没有必要知道一共有多少个桔子.

总体平均是总体的平均值，也称为总体均值(mean)．在统计学中，常用μ 表示总体均值．当总体含有N个个体，第i个个体是y_i时，总体均值

$$\mu = \frac{y_1 + y_2 + \cdots + y_N}{N}.$$

当$y_1, y_2, \cdots, y_N$是总体的全部个体，μ是总体均值时，称

$$\sigma^2 = \frac{(y_1 - \mu)^2 + (y_2 - \mu)^2 + \cdots + (y_N - \mu)^2}{N}$$

为总体方差 或方差(variance).

总体方差描述了总体中的个体向总体均值μ的集中程度．方差越小，个体向μ集中得越好．总体方差σ^2也描述了总体中个体的分散程度或波动幅度，方差越小，个体就越整齐．

总体平均是总体的平均值，也称为总体均值(mean). 在统计学中，常用μ 表示总体均值. 当总体含有N个个体，第i个个体是y_i时，总体均值

$$\mu = \frac{y_1 + y_2 + \cdots + y_N}{N}.$$

当$y_1, y_2, \cdots, y_N$是总体的全部个体，μ是总体均值时，称

$$\sigma^2 = \frac{(y_1 - \mu)^2 + (y_2 - \mu)^2 + \cdots + (y_N - \mu)^2}{N}$$

为总体方差 或方差(variance).

总体方差描述了总体中的个体向总体均值μ的集中程度. 方差越小，个体向μ集中得越好. 总体方差σ^2也描述了总体中个体的分散程度或波动幅度，方差越小，个体就越整齐.

总体标准差是总体方差的算术平方根$\sigma = \sqrt{\sigma^2}$, 简称为标准差.

总体参数是描述总体特性的指标, 简称为参数(parameter).

参数表示总体的特征, 是要调查的指标. 总体均值、总体方差、总体标准差都是参数. 在讲到参数的时候, 要明确它是哪个总体的参数.

B. 样本与估计

考虑某大学一年级2000个同学的平均身高μ. 要得到这2000个同学的平均身高不是一件很困难的事情, 只要了解了每个同学的身高, 就可以利用公式

$$\mu = \frac{\text{这2000个同学身高之和}}{2000}$$

计算得到.

但是在同一时刻要了解每个同学的准确身高也不是很容易的事情. 如果让各班长在班上依次点名登记全班同学的身高, 然后汇总, 可能有些同学一时不能给出准确的回答, 也可能有些同学受到其他同学的影响后, 偏向于把自己的身高报高或报低. 用这样的数据进行计算后得到的结果可能会产生偏差.

B. 样本与估计

考虑某大学一年级2000个同学的平均身高μ. 要得到这2000个同学的平均身高不是一件很困难的事情, 只要了解了每个同学的身高, 就可以利用公式

$$\mu = \frac{\text{这2000个同学身高之和}}{2000}$$

计算得到.

但是在同一时刻要了解每个同学的准确身高也不是很容易的事情. 如果让各班长在班上依次点名登记全班同学的身高, 然后汇总, 可能有些同学一时不能给出准确的回答, 也可能有些同学受到其他同学的影响后, 偏向于把自己的身高报高或报低. 用这样的数据进行计算后得到的结果可能会产生偏差.

同一天对每个同学进行一次身高测量可以得到均值μ的准确值，但是要花费同学们较多的时间和精力．统计上解决这类问题的最好方法是进行抽样调查，例如在2000个同学中只具体测量50个同学的身高，用这50个同学的平均身高作为总体平均身高μ的近似．这时我们称这50个同学的身高为总体的样本，称50为样本量．

从总体中抽取一部分个体，称这些个体为样本(sample)，样本也叫做观测数据(observation data)．

称构成样本的个体数目为样本容量，简称为样本量(sample size)．

称从总体抽取样本的工作为抽样(sampling)．

同一天对每个同学进行一次身高测量可以得到均值μ的准确值，但是要花费同学们较多的时间和精力．统计上解决这类问题的最好方法是进行抽样调查，例如在2000个同学中只具体测量50个同学的身高，用这50个同学的平均身高作为总体平均身高μ的近似．这时我们称这50个同学的身高为总体的样本，称50为样本量．

从总体中抽取一部分个体，称这些个体为样本(sample)，样本也叫做观测数据(observation data)．

称构成样本的个体数目为样本容量，简称为样本量(sample size)．

称从总体抽取样本的工作为抽样(sampling)．

在考虑身高问题时，对于前述被选中的50个同学，用 $x_1, x_2, \cdots, x_{50}$ 分别表示第 1, 2, ⋯ , 50 个同学在调查日的身高，则这50个同学的身高

$$x_1, x_2, \cdots, x_{50}$$

是样本. 用n表示样本量，则$n = 50$.

样本均值是样本的平均值，用 $\overline{x}$ 表示.
给定n个观测数据$x_1, x_2, \cdots, x_n$，称

$$s^2 = \frac{1}{n-1}[(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2]$$

为这n个数据的样本方差.

样本方差s^2是描述观测数据关于样本均值$\overline{x}$分散程度的指标，也是描述数据的分散程度或波动幅度的指标.

在考虑身高问题时，对于前述被选中的50个同学，用 $x_1, x_2, \cdots, x_{50}$ 分别表示第 $1, 2, \cdots, 50$ 个同学在调查日的身高，则这50个同学的身高

$$x_1, x_2, \cdots, x_{50}$$

是样本．用n表示样本量，则$n = 50$．

样本均值是样本的平均值，用 $\overline{x}$ 表示．
给定n个观测数据$x_1, x_2, \cdots, x_n$，称

$$s^2 = \frac{1}{n-1}[(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2]$$

为这n个数据的样本方差．

样本方差s^2是描述观测数据关于样本均值$\overline{x}$分散程度的指标，也是描述数据的分散程度或波动幅度的指标．

样本标准差是样本方差的算术平方根 $s=\sqrt{s^2}$. 和总体均值 μ 相比较后知道, 只要抽样合理, 对于较大的样本量 n, 样本均值 $\overline{x}$ 会接近 μ. 于是, $\overline{x}$ 是总体均值 μ 的近似, 所以称为 μ 的估计 (estimator).

估计是利用样本计算出的对参数的估计值. 估计能从观测数据直接计算出来.

对相同的观测数据, 不同的方法可以给出不同的估计结果, 所以估计不是唯一的. 这种不唯一性恰恰为统计学家们寻找更好的估计留下了余地.

实际问题中, 总体的容量往往是非常大的, 这时从数据本身无法看清总体的情况. 样本均值和样本方差可以提供必要的信息.

例5.1.1 比赛中甲、乙两位射击运动员分别进行了10次射击，成绩分别如下：

甲	9.5	9.9	9.9	9.9	9.8	9.7	9.5	9.3	9.6	9.6
乙	9.4	9.3	9.5	9.0	9.1	9.8	9.7	9.5	9.3	9.4

问哪个运动员平均水平高，哪个运动员水平更稳定.

解 用 $\overline{x}$, s_x 和 $\overline{y}$, s_y 分别表示甲和乙成绩的样本均值和样本标准差，经过计算得到

$$\overline{x} = 9.67,\ s_x = 0.2058,\quad \overline{y} = 9.4,\ s_y = 0.2449.$$

甲的平均水平和稳定性都比乙好.

此题表明，知道样本标准差后，可以作出更好的比较结果.

例5.1.1 比赛中甲、乙两位射击运动员分别进行了10次射击，成绩分别如下：

甲	9.5	9.9	9.9	9.9	9.8	9.7	9.5	9.3	9.6	9.6
乙	9.4	9.3	9.5	9.0	9.1	9.8	9.7	9.5	9.3	9.4

问哪个运动员平均水平高，哪个运动员水平更稳定.

解 用 $\overline{x}$, s_x 和 $\overline{y}$, s_y 分别表示甲和乙成绩的样本均值和样本标准差，经过计算得到

$$\overline{x} = 9.67,\ s_x = 0.2058, \quad \overline{y} = 9.4,\ s_y = 0.2449.$$

甲的平均水平和稳定性都比乙好.

此题表明，知道样本标准差后，可以作出更好的比较结果.

在日常生活中人们总是自觉或不自觉地应用抽样方法, 例如在市场上买花生或瓜子时总要先尝几个看看是否饱满和新鲜; 在烧菜的过程中经常要取一点尝尝味道.

在考察锅里汤的味道时, 没有必要把汤喝完, 只要把汤 “搅拌均匀”, 从中品尝一勺就可以了. 注意无论这锅汤有多多, 只要一勺就够了.

记住上面的例子是大有好处的, 因为它提供了抽样调查方法的最重要信息.

第一，把汤“搅拌均匀”是说明抽样的随机性，没有抽样的随机性，样本就不能很好地反映总体的情况. 把刚加盐的地方舀出的汤作样本，你会作出汤太咸了的错误结论.

第二，“品尝一勺”指出了选取的样本量不能太少，也不必太大. 太少了不足以品出味道，品尝一大碗也没有必要.

第三，“无论这锅汤有多多，只要一勺就够了”. 这里体现出抽样调查的如下基本性质：总体个数增大时，样本量不必跟着增大.

有人认为，总体数目很大时，样本量也必须跟着增大. 这种认识带有片面性.

实际的情况是这样的：在随机抽样下，一开始增加样本量会很快地增加估计的准确度，但是当样本量到达一定的时候，继续增加样本量效果就不明显了，再增加样本量就只是造成浪费了.

第一，把汤“搅拌均匀”是说明抽样的随机性，没有抽样的随机性，样本就不能很好地反映总体的情况．把刚加盐的地方舀出的汤作样本，你会作出汤太咸了的错误结论．

第二，“品尝一勺”指出了选取的样本量不能太少，也不必太大．太少了不足以品出味道，品尝一大碗也没有必要．

第三，“无论这锅汤有多多，只要一勺就够了”．这里体现出抽样调查的如下基本性质：总体个数增大时，样本量不必跟着增大．

有人认为，总体数目很大时，样本量也必须跟着增大．这种认识带有片面性．

实际的情况是这样的：在随机抽样下，一开始增加样本量会很快地增加估计的准确度，但是当样本量到达一定的时候，继续增加样本量效果就不明显了，再增加样本量就只是造成浪费了．

A. 抽样调查的必要性

抽样调查是相对于普查而言的，其含义是从总体中按一定的方式抽出样本进行考察，然后用样本的情况来推断总体的情况.

在评价1 000个同型号的微波炉的平均工作寿命μ时，预备从中抽取n个进行工作寿命的测量试验，用这n个微波炉的平均工作寿命估计总体的平均工作寿命μ.

这里，总体是1 000个微波炉的工作寿命，样本量是n，被选中的微波炉的工作寿命构成样本. 样本平均$\bar{x}$是总体均值μ的估计.

在正确抽样的前提下，样本量越大，$\overline{x}$越接近总体均值μ. 但是，较大的样本量造成的花费也很大，因为这n个微波炉做完寿命试验后就报废了. 在本问题中要想得到真正的总体均值μ是不可能的，除非把这1 000个微波炉都拿来做工作寿命试验，报废掉这1 000个微波炉.

在很多实际问题中，采用抽样的方法来确定总体性质不仅是必要的，也是必须的.

总体很大时，抽样调查往往可以提高调查的质量. 有人认为抽样调查不如全面调查得到的结论准确，这是不客观的. 看到抽样调查是用局部推断全体，带有抽样的误差，只是看到了问题的一个方面. 实际上调查数据的质量更重要，总体很大时进行全面调查，往往因为工作量过大、时间过长等而影响数据的质量. 一项经过科学设计并严格实施的抽样调查可能得到比全面调查更可靠的结果.

在正确抽样的前提下，样本量越大，$\overline{x}$越接近总体均值μ. 但是，较大的样本量造成的花费也很大，因为这n个微波炉做完寿命试验后就报废了. 在本问题中要想得到真正的总体均值μ是不可能的，除非把这1 000个微波炉都拿来做工作寿命试验，报废掉这1 000个微波炉.

在很多实际问题中，采用抽样的方法来确定总体性质不仅是必要的，也是必须的.

总体很大时，抽样调查往往可以提高调查的质量. 有人认为抽样调查不如全面调查得到的结论准确，这是不客观的. 看到抽样调查是用局部推断全体，带有抽样的误差，只是看到了问题的一个方面. 实际上调查数据的质量更重要，总体很大时进行全面调查，往往因为工作量过大、时间过长等而影响数据的质量. 一项经过科学设计并严格实施的抽样调查可能得到比全面调查更可靠的结果.

在正确抽样的前提下，样本量越大，$\overline{x}$越接近总体均值μ. 但是，较大的样本量造成的花费也很大，因为这n个微波炉做完寿命试验后就报废了. 在本问题中要想得到真正的总体均值μ是不可能的，除非把这1 000个微波炉都拿来做工作寿命试验，报废掉这1 000个微波炉.

在很多实际问题中，采用抽样的方法来确定总体性质不仅是必要的，也是必须的.

总体很大时，抽样调查往往可以提高调查的质量. 有人认为抽样调查不如全面调查得到的结论准确，这是不客观的. 看到抽样调查是用局部推断全体，带有抽样的误差，只是看到了问题的一个方面. 实际上调查数据的质量更重要，总体很大时进行全面调查，往往因为工作量过大、时间过长等而影响数据的质量. 一项经过科学设计并严格实施的抽样调查可能得到比全面调查更可靠的结果.

B. 随机抽样

如果总体中的每个个体都有相同的机会被抽中, 就称这样的抽样方法为随机抽样方法. 人们经常用“任取”、“随机抽取”或“等可能抽取”等来表示随机抽样.

从概率论的知识知道, 如果从总体中任选一个个体, 这个个体是随机变量, 这个随机变量的数学期望是总体均值, 方差是总体方差.

随机抽样又分为无放回的随机抽样和有放回的随机抽样. 无放回的随机抽样指在总体中随机抽出一个个体后,下次在余下的个体中再进行随机抽样. 有放回的随机抽样指抽出一个个体, 记录下抽到的结果后放回, 摇匀后再进行下一次随机抽样.

B. 随机抽样

如果总体中的每个个体都有相同的机会被抽中，就称这样的抽样方法为随机抽样方法. 人们经常用“任取”、“随机抽取”或“等可能抽取”等来表示随机抽样.

从概率论的知识知道，如果从总体中任选一个个体，这个个体是随机变量，这个随机变量的数学期望是总体均值，方差是总体方差.

随机抽样又分为无放回的随机抽样和有放回的随机抽样. 无放回的随机抽样指在总体中随机抽出一个个体后，下次在余下的个体中再进行随机抽样. 有放回的随机抽样指抽出一个个体，记录下抽到的结果后放回，摇匀后再进行下一次随机抽样.

B. 随机抽样

如果总体中的每个个体都有相同的机会被抽中, 就称这样的抽样方法为随机抽样方法. 人们经常用“任取”、“随机抽取”或“等可能抽取”等来表示随机抽样.

从概率论的知识知道, 如果从总体中任选一个个体, 这个个体是随机变量, 这个随机变量的数学期望是总体均值, 方差是总体方差.

随机抽样又分为无放回的随机抽样和有放回的随机抽样. 无放回的随机抽样指在总体中随机抽出一个个体后,下次在余下的个体中再进行随机抽样. 有放回的随机抽样指抽出一个个体, 记录下抽到的结果后放回, 摇匀后再进行下一次随机抽样.

例5.2.1 设N件产品中有M件次品，N, M都是未知的．估计这批产品的次品率$p = M/N$.

解 无放回地从中依次取n件，用Y表示取得的次品数，则$Y \sim H(N, M, n)$，按照附录B，有

$$\mathrm{E}Y = np, \quad \mathrm{Var}(Y) = np(1-p)\frac{N-n}{N-1}.$$

用样本次品率$\hat{p} = Y/n$估计p时，有

$$\mathrm{E}\hat{p} = p, \quad \mathrm{Var}(\hat{p}) = \frac{1}{n}p(1-p)\frac{N-n}{N-1}. \tag{2.1}$$

如果采用有放回的随机抽样，用X表示取得的次品数，则$X \sim \mathcal{B}(n, p)$，这时有

$$\mathrm{E}X = np, \quad \mathrm{Var}(X) = np(1-p).$$

例5.2.1 设N件产品中有M件次品, N, M都是未知的. 估计这批产品的次品率$p = M/N$.

解 无放回地从中依次取n件, 用Y表示取得的次品数, 则$Y \sim H(N, M, n)$, 按照附录B, 有

$$\mathrm{E}Y = np, \quad \mathrm{Var}(Y) = np(1-p)\frac{N-n}{N-1}.$$

用样本次品率$\hat{p} = Y/n$估计p时, 有

$$\mathrm{E}\hat{p} = p, \quad \mathrm{Var}(\hat{p}) = \frac{1}{n}p(1-p)\frac{N-n}{N-1}. \tag{2.1}$$

如果采用有放回的随机抽样, 用X表示取得的次品数, 则$X \sim \mathcal{B}(n, p)$, 这时有

$$\mathrm{E}X = np, \quad \mathrm{Var}(X) = np(1-p).$$

用这时的样本次品率$\tilde{p}=X/n$估计p时，有

$$\mathrm{E}\tilde{p}=p,\quad \mathrm{Var}(\tilde{p})=\frac{1}{n}p(1-p). \qquad (2.2)$$

$\mathrm{E}\hat{p}=\mathrm{E}\tilde{p}=p$，说明这两种方法都是较好的估计方法，没有系统偏差.

由于方差$\mathrm{E}(\hat{p}-p)^2$ 描述的是$\hat{p}$向真实参数p的集中程度，因而是描述估计精度的量. 方差越小，说明估计的精度越高. $\mathrm{Var}(\hat{p})<\mathrm{Var}(\tilde{p})$ 说明无放回随机抽样的估计精度好于有放回随机抽样的估计精度. 但是当N比n大很多时，$(N-n)/(N-1)$接近于1，说明两种抽样方法差别不大.

另外$\mathrm{Var}(\tilde{p})$与N无关，说明要达到一定的估计精度，只需要适当地增加n. 并不是说总体数目N越大，就需要多抽样. 无放回随机抽样下的情况也是类似的，因为实际问题中N通常都很大，而相比之下n较小.

用这时的样本次品率$\tilde{p} = X/n$估计p时，有

$$\mathrm{E}\tilde{p} = p, \quad \mathrm{Var}(\tilde{p}) = \frac{1}{n}p(1-p). \qquad (2.2)$$

$\mathrm{E}\hat{p} = \mathrm{E}\tilde{p} = p$，说明这两种方法都是较好的估计方法，没有系统偏差.

由于方差$\mathrm{E}(\hat{p}-p)^2$ 描述的是$\hat{p}$向真实参数p的集中程度，因而是描述估计精度的量. 方差越小，说明估计的精度越高. $\mathrm{Var}(\hat{p}) < \mathrm{Var}(\tilde{p})$ 说明无放回随机抽样的估计精度好于有放回随机抽样的估计精度. 但是当N比n大很多时，$(N-n)/(N-1)$接近于1，说明两种抽样方法差别不大.

另外$\mathrm{Var}(\tilde{p})$与N无关，说明要达到一定的估计精度，只需要适当地增加n. 并不是说总体数目N越大，就需要多抽样. 无放回随机抽样下的情况也是类似的，因为实际问题中N通常都很大，而相比之下n较小.

用这时的样本次品率$\tilde{p} = X/n$估计p时，有

$$\mathrm{E}\tilde{p} = p, \quad \mathrm{Var}(\tilde{p}) = \frac{1}{n}p(1-p). \qquad (2.2)$$

$\mathrm{E}\hat{p} = \mathrm{E}\tilde{p} = p$，说明这两种方法都是较好的估计方法，没有系统偏差.

由于方差$\mathrm{E}(\hat{p}-p)^2$ 描述的是$\hat{p}$向真实参数p的集中程度，因而是描述估计精度的量. 方差越小，说明估计的精度越高.
$\mathrm{Var}(\hat{p}) < \mathrm{Var}(\tilde{p})$ 说明无放回随机抽样的估计精度好于有放回随机抽样的估计精度. 但是当N比n大很多时，$(N-n)/(N-1)$接近于1，说明两种抽样方法差别不大.

另外$\mathrm{Var}(\tilde{p})$与N无关，说明要达到一定的估计精度，只需要适当地增加n. 并不是说总体数目N越大，就需要多抽样. 无放回随机抽样下的情况也是类似的，因为实际问题中N通常都很大，而相比之下n较小.

在相同的总体中和相同的样本量下，无放回随机抽样得到的结果比有放回随机抽样得到的结果要好. 但是当总体的数量很大，样本量相对总体的数量又很小时，这两种抽样方法得到的结果是相近的.

试验和理论都证明：在随机抽样下，样本均值 $\overline{x}$ 是总体均值 μ 很好的估计，样本标准差 s 是总体标准差 σ 很好的估计. 在样本量不大时，增加样本量可以比较好地提高估计的精确度.

考虑某大学一年级2000个同学的平均身高 μ 时，需要调查 50 个同学的身高. 实现无放回的随机抽样的方法是先将2000个同学的学号分别写在2000张小纸片上，然后放入一个大纸箱进行充分地摇匀，最后从纸箱中无放回地抽取50个纸片，纸片上的学号就是被选中的同学的学号.

在相同的总体中和相同的样本量下，无放回随机抽样得到的结果比有放回随机抽样得到的结果要好．但是当总体的数量很大，样本量相对总体的数量又很小时，这两种抽样方法得到的结果是相近的．

试验和理论都证明：在随机抽样下，样本均值 $\bar{x}$ 是总体均值 μ 很好的估计，样本标准差 s 是总体标准差 σ 很好的估计．在样本量不大时，增加样本量可以比较好地提高估计的精确度．

考虑某大学一年级2000个同学的平均身高 μ 时，需要调查 50 个同学的身高．实现无放回的随机抽样的方法是先将2000个同学的学号分别写在2000张小纸片上，然后放入一个大纸箱进行充分地摇匀，最后从纸箱中无放回地抽取50个纸片，纸片上的学号就是被选中的同学的学号．

在相同的总体中和相同的样本量下，无放回随机抽样得到的结果比有放回随机抽样得到的结果要好. 但是当总体的数量很大，样本量相对总体的数量又很小时，这两种抽样方法得到的结果是相近的.

试验和理论都证明：在随机抽样下，样本均值 $\bar{x}$ 是总体均值 μ 很好的估计，样本标准差 s 是总体标准差 σ 很好的估计. 在样本量不大时，增加样本量可以比较好地提高估计的精确度.

考虑某大学一年级2000个同学的平均身高 μ 时，需要调查 50 个同学的身高. 实现无放回的随机抽样的方法是先将2000个同学的学号分别写在2000张小纸片上，然后放入一个大纸箱进行充分地摇匀，最后从纸箱中无放回地抽取50个纸片，纸片上的学号就是被选中的同学的学号.

C. 随机抽样的无偏性

样本均值是对总体均值的估计. 在总体中任取一个个体X, X是随机变量, 从数学期望的定义知道$\mathrm{E}X=\mu$是总体均值. 这说明随机抽样是无偏的. 如果用$X_1, X_2, \cdots, X_n$表示依次随机抽取的样本, 则样本均值

$$\overline{X}=\frac{1}{n}\sum_{j=1}^{n}X_j$$

是总体均值μ的估计. 下面证明$\mathrm{E}\overline{X}=\mu$.

在有放回的随机抽样下，$X_1, X_2, \cdots, X_n$有相同的数学期望μ，于是有

$$\mathrm{E}\overline{X} = \frac{1}{n}\sum_{j=1}^{n}\mathrm{E}X_j = \frac{1}{n}\sum_{j=1}^{n}\mu = \mu.$$

在无放回的随机抽样下，根据抽签的原理，第j次抽到每个个体的概率是相同的，所以X_1和X_j是同分布的，因而有相同的数学期望，于是也有$\mathrm{E}\overline{X} = \mu$. 说明在随机抽样下，样本均值$\overline{X}$总是总体均值$\mu$的无偏估计.

下面是没有采取正确的抽样方案导致调查结论严重失真的著名案例.

例5.2.2 1936年是美国总统选举年. 这年罗斯福 (Roosevelt) 任美国总统期满, 参加第二届的连任竞选, 对手是堪萨斯州州长兰登 (Landon). 当时美国刚从经济大萧条中恢复过来, 失业人数仍高达900多万, 人们的经济收入下降了三分之一后开始逐步回升. 当时, 观察家们普遍认为罗斯福会当选. 而美国的《文学摘要》杂志的调查却预测兰登会以57% 对43%的压倒优势获胜.

《文学摘要》的预测是基于对240万选民的民意调查得出的. 自1916年以来, 在历届美国总统的选举中《文学摘要》都作了正确的预测. 《文学摘要》的威信有力地支持着它的这次预测.

但是选举的结果是罗斯福以62% 对38%的压倒优势获胜, 此后不久《文学摘要》杂志就破产了.

要了解《文学摘要》预测失败的原因就必须检查他们的抽样调查方案.《文学摘要》是将问卷寄给了1 000万个选民,基于收回的240万份问卷得出的判断. 这些选民的地址是在诸如电话簿、俱乐部会员名单等上查到的.

分析:1936年只有大约四分之一的家庭安装了电话. 由于有钱人才更有可能安装家庭电话和参加俱乐部,所以《文学摘要》的调查方案漏掉了那些不属于俱乐部的穷人和没有安装电话的穷人,这就导致了调查结果有排除穷人的偏向.

在1936年,由于经济开始好转,穷人普遍有赞同罗斯福当选的倾向,富人有赞同兰登当选的倾向.《文学摘要》的调查结果更多地代表了富人的意愿,导致了预测的失败.

抽样的方案应当公平地对待每一位选民和每一个群体,以便得到选民的真实情况. 将哪一个群体排除在外的抽样方案都可能导致有偏的样本,从而导致错误的结论.

如果X是从总体中随机抽样得到的个体，则X是随机变量，X的分布就是总体的分布. 如果对总体进行有放回的随机抽样，则得到独立同分布且和X同分布的随机变量$X_1, X_2, \cdots, X_n$. 这时称$X_1, X_2, \cdots, X_n$是总体 X 的简单随机样本，简称为总体X的样本.

在观测放射性物质钋放射α粒子的试验中，用X表示7.5 s内观测到的粒子数. 独立重复观测时，用X_i表示第i个7.5 s的观测结果，则$X_1, X_2, \cdots, X_n$独立同分布且和X同分布，$X_1, X_2, \cdots, X_n$ 是总体X的样本.

定义5.3.1 如果$X_1, X_2, \cdots, X_n$独立同分布且和X同分布，则称X是总体，称$X_1, X_2, \cdots, X_n$是总体X的样本，称观测数据的个数n为样本量.

在实际问题中得到的总是样本$X_1, X_2, \cdots, X_n$的观测值 $x_1, x_2, \cdots, x_n$，这时也称$x_1, x_2, \cdots, x_n$ 是总体X的样本. 在统计学中，常常不把 $X_1, X_2, \cdots, X_n$与它们的观测值$x_1, x_2, \cdots, x_n$严格区分，这是为了符号使用的方便. 当对数据进行统计分析时，用大写的$X_1, X_2, \cdots, X_n$，实际计算时更多地用小写的 $x_1, x_2, \cdots, x_n$.

在统计问题中，总体X的分布形式往往是已知的．例如重复测量一个物体的重量时，认为总体X服从正态分布 $N(\mu, \sigma^2)$，未知参数是μ，σ^2，问题是根据总体X的样本$X_1, X_2, \cdots, X_n$ 估计总体参数μ，σ^2．观测放射物钋放射α粒子时，总体X服从泊松分布$\mathcal{P}(\lambda)$，未知参数是λ，问题是根据总体X的样本$X_1, X_2, \cdots, X_n$ 估计λ．

设$X_1, X_2, \cdots, X_n$是总体X的样本，θ 是总体X的未知参数．如果

$$g_n(x_1, x_2, \cdots, x_n)$$

是已知函数，则称

$$\hat{\theta}_n = g_n(X_1, X_2, \cdots, X_n)$$

是θ的估计量,简称为估计(estimator)．换句话说，估计或估计量是从观测数据$X_1, X_2, \cdots, X_n$能够直接计算的量．计算后得到的值称为估计值．估计量也称为统计量(statistic)．为了符号使用的简便，以后会把统计量$\hat{\theta}_n$简写成$\hat{\theta}$，于是

$$\hat{\theta} \equiv \hat{\theta}_n.$$

设$\hat{\theta}$是总体参数θ的估计，作为随机变量$X_1, X_2, \cdots, X_n$的函数，估计量$\hat{\theta}$也是随机变量．估计量是样本的函数．关于估计量有如下的定义．

定义5.3.2 设$\hat{\theta}$ 是θ的估计．

(1) 如果$\mathrm{E}\hat{\theta} = \theta$，则称$\hat{\theta}$ 是θ的**无偏估计**；

(2) 如果当样本量$n \to \infty$，$\hat{\theta}$依概率收敛到θ，则称$\hat{\theta}$是θ的**相合估计**；

(3) 如果当样本量$n \to \infty$，$\hat{\theta}$以概率1收敛到θ，则称$\hat{\theta}$是θ的**强相合估计**．

由于以概率1收敛可以推出依概率收敛，所以强相合估计一定是相合估计．一个估计起码应当是相合的，否则我们不知道这个估计有什么用，也不知道它到底估计谁．

A. 样本均值

设总体均值$\mu = EX$存在, $X_1, X_2, \cdots, X_n$是总体X的样本. 均值μ的估计定义为

$$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

由于$\overline{X}_n$是从样本计算出来的, 所以是样本均值. 样本均值$\overline{X}_n$有如下的性质:

(1) $\overline{X}_n$是μ的无偏估计, 这是因为$E\overline{X}_n = \mu$;

(2) $\overline{X}_n$是μ的强相合估计, 从而是相合估计. 这是因为从强大数律得到

$$\lim_{n\to\infty} \overline{X}_n = \mu \text{ a.s.}.$$

给定总体X的样本$X_1, X_2, \cdots, X_n$, 也可以用

$$\overline{X}_{n-1} = \frac{X_1 + X_2 + \cdots + X_{n-1}}{n-1}$$

估计总体均值μ. 这时仍然有

$$\mathrm{E}\overline{X}_{n-1} = \mu, \quad \lim_{n\to\infty} \overline{X}_{n-1} = \mu \text{ a.s.},$$

说明$\overline{X}_{n-1}$也是μ的无偏估计和强相合估计.
但是因为少用了一个数据, 所以

$$\operatorname{Var}(\overline{X}_{n-1}) = \frac{\sigma^2}{n-1} > \operatorname{Var}(\overline{X}_n) = \frac{\sigma^2}{n}.$$

说明在均方误差的意义下, $\overline{X}_{n-1}$没有$\overline{X}_n$的估计精度高. 这时称$\overline{X}_n$比$\overline{X}_{n-1}$ 更有效.

B. 样本方差

给定总体X的样本$X_1, X_2, \cdots, X_n$, 以下用$\hat{\mu}$ 表示样本均值, 于是

$$\hat{\mu} = \overline{X}_n.$$

总体方差$\sigma^2 = \mathrm{Var}(X)$的估计由

$$S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \hat{\mu})^2$$

定义. 由于S^2是从样本计算出来的,所以称为样本方差.

取定j. 因为$\mathrm{E}(X_j-\hat{\mu})=\mu-\mu=0$, 所以从$X_1,X_2,\cdots,X_n$的独立性得到

$$\begin{aligned}
\mathrm{E}(X_j-\hat{\mu})^2 &= \mathrm{Var}\left(X_j-\frac{1}{n}\sum_{i=1}^{n}X_i\right)\\
&= \mathrm{Var}\left[\left(1-\frac{1}{n}\right)X_j-\frac{1}{n}\sum_{i\neq j}X_i\right]\\
&= \left(\frac{n-1}{n}\right)^2\sigma^2+\frac{1}{n^2}\sum_{i\neq j}\sigma^2\\
&= \left[\left(\frac{n-1}{n}\right)^2+\frac{n-1}{n^2}\right]\sigma^2\\
&= \frac{n-1}{n}\sigma^2.
\end{aligned}$$

于是得到

$$\mathrm{E}S^2 = \frac{1}{n-1}\sum_{j=1}^{n}\mathrm{E}\,(X_j - \hat{\mu})^2 = \frac{n}{n-1}\cdot\frac{n-1}{n}\sigma^2 = \sigma^2.$$

说明样本方差S^2是总体方差的无偏估计.
利用强大数律$\hat{\mu} \to \mu$ a.s. 和

$$\frac{1}{n-1}\sum_{j=1}^{n} X_j^2 \to \mathrm{E}\,X^2 \text{ a.s.},$$

得到

$$\begin{aligned}
\frac{1}{n-1}\sum_{j=1}^{n}(X_j-\hat{\mu})^2 &= \frac{1}{n-1}\sum_{j=1}^{n}(X_j^2-2X_j\hat{\mu}+\hat{\mu}^2) \\
&= \frac{1}{n-1}\Big(\sum_{j=1}^{n}X_j^2-2n\frac{1}{n}\sum_{j=1}^{n}X_j\hat{\mu}+n\hat{\mu}^2\Big) \\
&= \frac{1}{n-1}\sum_{j=1}^{n}X_j^2-\frac{n}{n-1}\hat{\mu}^2 \\
&\to \mathrm{E}\,X^2-\mu^2=\sigma^2 \text{ a.s.},
\end{aligned}$$

说明样本方差S^2是总体方差σ^2的强相合估计.

C. 样本标准差

由于S^2是σ^2的估计，所以定义标准差σ的估计为

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(X_j - \hat{\mu})^2},$$

称S为样本标准差. 由于$S^2 \to \sigma^2$ a.s., 所以$S \to \sigma$ a.s. 成立, 说明S是σ的强相合估计.

当$\sigma > 0$, S不是σ的无偏估计, 也就是说 $\mathrm{E}S = \sigma$ 不成立. 这是因为没有不全为零的常数a, b, 使得$P(aS + b = 0) = 1$, 所以由内积不等式得到

$$\mathrm{E}S = \mathrm{E}(S \cdot 1) < \sqrt{\mathrm{E}S^2 \cdot \mathrm{E}1^2} = \sqrt{\sigma^2} = \sigma.$$

于是$\mathrm{E}S < \sigma$. 这时称S低估了σ.

我们把上面的结果总结如下.

定理5.3.1 设$X_1, X_2, \cdots, X_n$是总体X的样本, $\mu = \mathrm{E}X$, $\sigma^2 = \mathrm{Var}(X) > 0$, 则

(1) 样本均值$\overline{X}_n$是总体均值μ的强相合无偏估计;

(2) 样本方差S^2 是总体方差σ^2的强相合无偏估计;

(3) 样本标准差S是总体标准差σ的强相合估计, 但是$\mathrm{E}\,S < \sigma$.

例5.3.1 设$X_1, X_2, \cdots, X_n$是总体X的样本. 当$\mu_k = \mathrm{E}X^k$存在时, 试给出μ_k的强相合无偏估计.

解 因为$X_1^k, X_2^k, \cdots, X_n^k$独立同分布, 且和$X^k$同分布, 所以是总体$X^k$的样本. 并且

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \qquad (3.1)$$

是μ_k的估计. 从定理5.3.1知道$\hat{\mu}_k$是μ_k的强相合无偏估计.

例5.3.1 设$X_1, X_2, \cdots, X_n$是总体X的样本. 当$\mu_k = \mathrm{E}X^k$存在时, 试给出μ_k的强相合无偏估计.

解 因为$X_1^k, X_2^k, \cdots, X_n^k$独立同分布, 且和$X^k$同分布, 所以是总体$X^k$的样本. 并且

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \qquad (3.1)$$

是μ_k的估计. 从定理5.3.1知道$\hat{\mu}_k$是μ_k的强相合无偏估计.

在例5.3.1中，称$\mu_k = \mathrm{E}\, X^k$为X的k阶原点矩，称$\hat{\mu}_k$为k阶样本原点矩.

在实际数据的计算中，常用$\overline{x}_n$, s^2 和s 分别表示样本均值、样本方差和样本标准差:

$$\overline{x}_n = \frac{1}{n}\sum_{j=1}^{n} x_j,\ s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \overline{x}_n)^2,\ s = \sqrt{s^2}.$$

如果$x_1, x_2, \cdots, x_n$是总体X的样本，则定理5.3.1保证了如下事实:

$$\lim_{n\to\infty} \overline{x}_n = \mu,\ \lim_{n\to\infty} s^2 = \sigma^2,\ \lim_{n\to\infty} s = \sigma.$$

为了解样本均值的表现，下面用计算机产生10^7个正态总体$N(\mu,\sigma^2)$的样本 $x_1,x_2,\cdots,x_n$.利用前 n个观测数据计算的$\hat{\mu}_1$, $\hat{\mu}_2$, s^2, s如下，真值是 $\mu_1=1.8$, $\mu_2=3.28$, $\sigma^2=0.04$, $\sigma=0.2$.

n	10	10^2	10^3	10^4	10^5	10^6	10^7
$\hat{\mu}_1$	1.8794	1.8243	1.8023	1.8002	1.8006	1.8004	1.8000
$\hat{\mu}_2$	3.5688	3.3623	3.2867	3.2808	3.2826	3.2813	3.2798
s^2	0.0452	0.0350	0.0386	0.0401	0.0403	0.0400	0.0400
s	0.2126	0.1871	0.1964	0.2002	0.2007	0.2000	0.2000

计算结果支持强相合结论:
$\hat{\mu}_1\to\mu_1$, $\hat{\mu}_2\to\mu_2$, $s^2\to\sigma^2$, $s\to\sigma$.

为了解样本均值的表现，下面用计算机产生10^7个正态总体$N(\mu,\sigma^2)$的样本 $x_1, x_2, \cdots, x_n$.利用前 n个观测数据计算的$\hat{\mu}_1$, $\hat{\mu}_2$, s^2, s如下，真值是 $\mu_1 = 1.8$, $\mu_2 = 3.28$, $\sigma^2 = 0.04$, $\sigma = 0.2$.

n	10	10^2	10^3	10^4	10^5	10^6	10^7
$\hat{\mu}_1$	1.8794	1.8243	1.8023	1.8002	1.8006	1.8004	1.8000
$\hat{\mu}_2$	3.5688	3.3623	3.2867	3.2808	3.2826	3.2813	3.2798
s^2	0.0452	0.0350	0.0386	0.0401	0.0403	0.0400	0.0400
s	0.2126	0.1871	0.1964	0.2002	0.2007	0.2000	0.2000

计算结果支持强相合结论:
$\hat{\mu}_1 \to \mu_1$, $\hat{\mu}_2 \to \mu_2$, $s^2 \to \sigma^2$, $s \to \sigma$.

设$x_1, x_2, \cdots, x_n$是总体X的样本. 对于$k \geqslant 1$, 以后称

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k \tag{4.1}$$

为$\mu_k = \mathrm{E}\, X^k$的矩估计.

下面通过举例介绍参数的矩估计.

例5.4.1 某高校在一年中组织了12次科普报告会，每次报告会的听众人数如下:

169 183 167 157 163 151 154 157 163 154 162 165

如果每次报告会的听众人数相互独立，都服从泊松分布$\mathcal{P}(\lambda)$，试估计参数λ.

解　用X表示服从泊松分布$\mathcal{P}(\lambda)$的随机变量，则

$$\lambda = \mathrm{E}\,X = \mu_1.$$

因为μ_1的矩估计是 $\hat{\mu}_1$，所以λ的矩估计也是 $\hat{\mu}_1$. 将上面的数据代入$\hat{\mu}_1$，得到λ 的估计

$$\hat{\lambda} = \hat{\mu}_1 = \frac{1}{12}\sum_{i=1}^{12} x_i = 162.083.$$

本例中称$\hat{\lambda}$为λ的矩估计. $\hat{\lambda}$正是这12次报告会的平均听众数.

例5.4.2 网民甲在搜狐网上发贴后就开始观察跟贴情况. 并记录了以下的跟贴间隔时间(单位:s):

1	2	9	33	28	62	17	46	1	12
35	11	53	33	2	15	62	7	18	81
63	20	18	40	22	32	126	145	47	176
38	15	96	135	19	20	67	166	67	21

假设跟贴的间隔时间服从指数分布$\mathcal{E}(\lambda)$, 试估计参数λ.

解　设X服从指数分布$\mathcal{E}(\lambda)$. 由$\mu = \mathrm{E}X = 1/\lambda$得到$\lambda = 1/\mu$. 因为$\hat{\mu}_1$是$\mu$的矩估计，于是可以用

$$\hat{\lambda} = 1/\hat{\mu}_1 = 1/\overline{x}_n$$

估计参数λ. 经计算得到

$$\overline{x}_n = 46.525,\ \hat{\lambda} = 0.021\,5.$$

在例5.4.2中，称$\hat{\lambda} = 0.021\,5$为λ的矩估计.

例5.4.3 单晶硅太阳能电池以高纯度单晶硅棒为原料. 制作时需要将单晶硅棒进行切片, 每片的厚度在0.3 mm 左右. 现在通过随机抽样的方法测量了某厂家的n片单晶硅的厚度, 得到测量数据$x_1, x_2, \cdots, x_n$. 假设这批单晶硅厚度的总体分布是正态分布, 试估计这批单晶硅的总体均值和总体方差.

解 设$x_1, x_2, \cdots, x_n$是总体X的样本观测值, 则$X \sim N(\mu, \sigma^2)$, 并且

$$\mu = \mathrm{E}X, \ \sigma^2 = \mathrm{E}X^2 - (\mathrm{E}X)^2 = \mu_2 - \mu^2.$$

因为$\hat{\mu}_1$, $\hat{\mu}_2$分别是μ, μ_2的矩估计, 于是分别用

$$\hat{\mu} = \hat{\mu}_1, \quad \hat{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}_1^2 \qquad (4.2)$$

估计μ, σ^2.

例5.4.3 单晶硅太阳能电池以高纯度单晶硅棒为原料. 制作时需要将单晶硅棒进行切片, 每片的厚度在0.3 mm 左右. 现在通过随机抽样的方法测量了某厂家的n片单晶硅的厚度, 得到测量数据$x_1, x_2, \cdots, x_n$. 假设这批单晶硅厚度的总体分布是正态分布, 试估计这批单晶硅的总体均值和总体方差.

解 设$x_1, x_2, \cdots, x_n$是总体X的样本观测值, 则$X \sim N(\mu, \sigma^2)$, 并且

$$\mu = \mathrm{E}X,\ \sigma^2 = \mathrm{E}X^2 - (\mathrm{E}X)^2 = \mu_2 - \mu^2.$$

因为$\hat{\mu}_1, \hat{\mu}_2$分别是μ, μ_2的矩估计, 于是分别用

$$\hat{\mu} = \hat{\mu}_1, \quad \hat{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}_1^2 \qquad (4.2)$$

估计μ, σ^2.

在例5.4.3 中，称$\hat{\mu}$ 为μ的矩估计，称$\hat{\sigma}^2$为σ^2的矩估计. 容易看出，$\hat{\mu}$就是μ 的样本均值. 但是从

$$\begin{aligned}\hat{\sigma}^2 &= \hat{\mu}_2 - (\hat{\mu}_1)^2 \\ &= \frac{1}{n}\sum_{j=1}^{n} x_j^2 - \overline{x}_n^2 \\ &= \frac{1}{n}\sum_{j=1}^{n}(x_j - \hat{\mu})^2\end{aligned}$$

知道，σ^2的矩估计$\hat{\sigma}^2$比样本方差s^2略小.

从上面的例子看出，如果总体X的分布函数$F(x;\theta)$ 只有一个未知参数θ，则$\mu_1 = \mathrm{E}\,X$常和θ有关．如果$g(s)$是已知函数，并且能从

$$\mu_1 = \mathrm{E}\,X \quad \text{得到} \quad \theta = g(\mu_1),$$

则$\hat{\theta} = g(\hat{\mu}_1)$ 是θ的矩估计，其中$\hat{\mu}_1$是样本均值．

如果总体X的分布函数$F(x;\theta_1,\theta_2)$ 有2个未知参数θ_1,θ_2，则$\mu_1=\mathrm{E}\,X$和$\mu_2=\mathrm{E}\,X^2$常和θ_1, θ_2有关. 如果$g_1(s,t)$, $g_2(s,t)$是已知函数, 并且能从

$$\begin{cases}\mu_1=\mathrm{E}\,X,\\ \mu_2=\mathrm{E}\,X^2\end{cases} \quad \text{得到} \quad \begin{cases}\theta_1=g_1(\mu_1,\mu_2),\\ \theta_2=g_2(\mu_1,\mu_2),\end{cases}$$

则

$$\hat\theta_1=g_1(\hat\mu_1,\hat\mu_2),\ \hat\theta_2=g_2(\hat\mu_1,\hat\mu_2)$$

分别是θ_1,θ_2 的矩估计, 其中$\hat\mu_1$, $\hat\mu_2$由(4.1)定义.

例5.4.4 设数据$x_1, x_2, \cdots, x_n$是总体$U[a, b]$的样本观测值，其中的a, b是未知参数，求a, b的矩估计.

解 设X在$[a, b]$中均匀分布，则X是所述的总体. 因为X的概率密度关于$(a+b)/2$对称，所以

$$\mu_1 = \mathrm{E}\,X = \frac{a+b}{2}. \qquad (4.3)$$

X的方差是

$$\mathrm{Var}\,(X) = \frac{(b-a)^2}{12}.$$

由公式$\mathrm{Var}\,(X) = \mathrm{E}\,X^2 - (\mathrm{E}\,X)^2$得到

$$\mu_2 - \mu_1^2 = \frac{(b-a)^2}{12}. \qquad (4.4)$$

例5.4.4 设数据$x_1, x_2, \cdots, x_n$是总体$\mathcal{U}[a, b]$的样本观测值，其中的a, b是未知参数，求a, b的矩估计.

解 设X在$[a, b]$中均匀分布，则X是所述的总体. 因为X的概率密度关于$(a+b)/2$对称，所以

$$\mu_1 = \mathrm{E}\,X = \frac{a+b}{2}. \qquad (4.3)$$

X的方差是

$$\mathrm{Var}\,(X) = \frac{(b-a)^2}{12}.$$

由公式$\mathrm{Var}\,(X) = \mathrm{E}\,X^2 - (\mathrm{E}\,X)^2$得到

$$\mu_2 - \mu_1^2 = \frac{(b-a)^2}{12}. \qquad (4.4)$$

解由(4.3) 和(4.4)构成的方程组得到

$$a = \mu_1 - \sqrt{3(\mu_2 - \mu_1^2)}, \ b = \mu_1 + \sqrt{3(\mu_2 - \mu_1^2)}.$$

于是, a, b的矩估计分别为

$$\hat{a} = \hat{\mu}_1 - \sqrt{3(\hat{\mu}_2 - \hat{\mu}_1^2)}, \ \hat{b} = \hat{\mu}_1 + \sqrt{3(\hat{\mu}_2 - \hat{\mu}_1^2)}. \tag{4.5}$$

为了解矩估计(4.5)的表现，用计算机产生总体$\mathcal{U}(0.8, 5.2)$的样本量为10^7的样本观测值，利用公式(2.5)和前n个观测值计算的矩估计如下：

n	10	10^2	10^3	10^4	10^5	10^6	10^7
$\hat{a}$	1.0336	0.9891	0.8053	0.8124	0.8001	0.8016	0.8001
$\hat{b}$	5.4547	5.2623	5.2132	5.2330	5.2067	5.2018	5.1997

计算结果支持强相合结论：$\hat{a} \to a$, $\hat{b} \to b$.

因为矩估计没有充分利用总体分布的信息，所以在已知X的分布信息时，矩估计一般不如下一节给出的最大似然估计好.

A. 离散分布的情况

如果袋中有红球和黑球共3个，现从中任取一个，得到红球. 你会判断袋中有2个红球，一个黑球. 这是因为当红球多于黑球时，才更可能取得红球.

若用A表示取得红球，用p表示袋中红球的比例，则$P(A) = p$. 从已知条件知道$p = 2/3$或$p = 1/3$. 因为现在A发生了，所以判断$p = 2/3$. 这时称$\hat{p} = 2/3$为p的最大似然估计. 这种思考问题的方法被称为最大似然方法.

甲、乙两人下棋，用p表示甲在每局中获胜的概率. 如果5局中甲胜了3局，你会判断$p > 1/2$. 这是因为当$p > 1/2$时，甲才更可能5局3胜. 但是p到底应当是多大呢? 让我们把这个问题数学化.

用p表示甲在每局中获胜的概率，设5局中甲胜X局，并假设各局的胜负相互独立，则

$$P(X=3)=C_5^3 p^3(1-p)^2.$$

现在已知$X=3$，所以p应当使得$X=3$发生的概率

$$L(p)=C_5^3 p^3(1-p)^2$$

达到最大. 对于$L(p)$求导数，令

$$L'(p)=C_5^3[3p^2(1-p)^2-2p^3(1-p)]=C_5^3 p^2(1-p)[3(1-p)-2p]=0,$$

得到$p=3/5(p=0,1$不合题意). 这时称$\hat{p}=3/5$为p的最大似然估计.

定义5.5.1 设离散随机变量$X_1, X_2, \cdots, X_n$有联合分布

$$p(x_1, x_2, \cdots, x_n; \theta) = P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n),$$

其中θ是未知参数，给定观测数据$x_1, x_2, \cdots, x_n$后，称θ的函数

$$L(\theta) = p(x_1, x_2, \cdots, x_n; \theta)$$

为似然函数，称$L(\theta)$的最大值点$\hat{\theta}$为θ的最大似然估计.

注： 定义5.5.1中的θ也可以是向量$\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_m)$. 最大似然估计通常被缩写成MLE(maximum likelihood estimator).

定义5.5.1 设离散随机变量$X_1, X_2, \cdots, X_n$有联合分布

$$p(x_1, x_2, \cdots, x_n; \theta) = P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n),$$

其中θ是未知参数，给定观测数据$x_1, x_2, \cdots, x_n$后，称θ的函数

$$L(\theta) = p(x_1, x_2, \cdots, x_n; \theta)$$

为似然函数，称$L(\theta)$的最大值点$\hat{\theta}$为θ的最大似然估计.

注： 定义5.5.1中的θ也可以是向量$\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_m)$. 最大似然估计通常被缩写成MLE(maximum likelihood estimator).

因为$\ln x$ 是严格单调的增函数，所以$l(\theta) = \ln L(\theta)$ 和$L(\theta)$有相同的最大值点. 通常称$l(\theta)$为对数似然函数. 在许多情况下，最大似然估计可由似然方程

$$l'(\theta) = 0$$

解出.

注意在定义5.5.1中，$x_1, x_2, \cdots, x_n$是观测数据，不再变动，所以$L(\theta)$和$l(\theta)$都是θ的函数.

例5.5.1 设$X_1, X_2, \cdots, X_n$独立同分布，都服从泊松分布$\mathcal{P}(\lambda)$.

(a) 给定X_1的观测值$x_1 = 169$，计算λ的MLE;

(b) 给定$X_1, X_2, \cdots, X_{12}$的观测值:

169 167 157 196 163 151 154 157 163 154 162 165

计算λ的MLE.

解 (a) X_1有概率分布

$$P(X_1 = x) = \frac{\lambda^x}{x!}\mathrm{e}^{-\lambda},\ x = 0, 1, \cdots.$$

因为$X_1 = x_1$, 所以λ的似然函数为

$$L(\lambda) = \frac{\lambda^{x_1}}{x_1!}\mathrm{e}^{-\lambda}.$$

对数似然函数为

$$l(\lambda) = \ln L(la) = x_1 \ln \lambda - \lambda - \ln(x_1!).$$

解似然方程 $l'(\lambda) = x_1/\lambda - 1 = 0$, 得到 λ 的MLE为 $\hat{\lambda} = x_1 = 169$.

(b) λ的似然函数为

$$\begin{aligned} L(\lambda) &= P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) \\ &= \frac{\lambda^{x_1}}{x_1!}\mathrm{e}^{-\lambda}\frac{\lambda^{x_2}}{x_2!}\mathrm{e}^{-\lambda}\cdots\frac{\lambda^{x_n}}{x_n!}\mathrm{e}^{-\lambda} \\ &= \frac{\lambda^{(x_1+x_2+\cdots+x_n)}}{x_1!x_2!\cdots x_n!}\mathrm{e}^{-n\lambda}. \end{aligned}$$

对数似然函数为

$$l(\lambda) = \ln L(\lambda) = (x_1 + x_2 + \cdots + x_n)\ln\lambda - n\lambda - c_0,$$

其中的c_0和参数λ 无关. 解似然方程$l'(\lambda) = 0$, 得到λ的MLE

$$\hat{\lambda} = \frac{x_1 + x_2 + \cdots + x_n}{n} = 163.167.$$

例5.5.2 设$X_1, X_2, \cdots, X_n$独立同分布，都服从两点分布$\mathcal{B}(1, p)$. 给定观测值$x_1, x_2, \cdots, x_n$，计算p的MLE.

解 因为$P(X_i = x_i) = p^{x_i}(1-p)^{1-x_i}$，所以$p$的似然函数为

$$\begin{aligned} L(p) &= P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) \\ &= P(X_1 = x_1)P(X_2 = x_2)\cdots P(X_n = x_n) \\ &= p^{n\overline{x}_n}(1-p)^{n-n\overline{x}_n}. \end{aligned}$$

对数似然函数为

$$l(p) = \ln L(p) = n\overline{x}_n \ln p + (n - n\overline{x}_n)\ln(1-p).$$

解似然方程$l'(p) = n\overline{x}_n/p - (n - n\overline{x}_n)/(1-p) = 0$，得到$p$的MLE

$$\hat{p} = \overline{x}_n.$$

例5.5.2　设$X_1, X_2, \cdots, X_n$独立同分布，都服从两点分布$\mathcal{B}(1, p)$. 给定观测值$x_1, x_2, \cdots, x_n$，计算p的MLE.

解　因为$P(X_i = x_i) = p^{x_i}(1-p)^{1-x_i}$，所以$p$的似然函数为

$$\begin{aligned} L(p) &= P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) \\ &= P(X_1 = x_1)P(X_2 = x_2)\cdots P(X_n = x_n) \\ &= p^{n\overline{x}_n}(1-p)^{n-n\overline{x}_n}. \end{aligned}$$

对数似然函数为

$$l(p) = \ln L(p) = n\overline{x}_n \ln p + (n - n\overline{x}_n)\ln(1-p).$$

解似然方程$l'(p) = n\overline{x}_n/p - (n - n\overline{x}_n)/(1-p) = 0$，得到$p$的MLE

$$\hat{p} = \overline{x}_n.$$

B. 连续分布的情况

例5.5.3 设$X \sim N(\mu, 1)$, 给定观测数据$X = x$, 试估计μ.

解 根据最大似然方法, μ应当使得$X = x$发生的概率

$$P(X = x) = f(x; \mu)\mathrm{d}x$$

达到最大. 等价地说μ应当使得

$$f(x; \mu) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-(x-\mu)^2/2}$$

达到最大. 于是得到μ的最大似然估计$\hat{\mu} = x$.

例5.5.4 设X_1, X_2独立同分布，都服从正态分布$N(\mu, \sigma^2)$. 给定观测数据x_1, x_2，试估计μ, σ^2.

解 (X_1, X_2)的联合密度

$$f(x_1, x_2; \mu, \sigma^2) = \frac{1}{2\pi\sigma^2} \exp\left\{ -\frac{1}{2\sigma^2}\left[(x_1-\mu)^2 + (x_2-\mu)^2\right] \right\}$$

是一个扣在xOy平面上的单峰曲面. 根据最大似然思想， μ, σ^2的选择应当使得曲面在(x_1, x_2)处达到最大. 于是使得函数

$$L(\mu, \sigma^2) = f(x_1, x_2; \mu, \sigma^2)$$

达到最大值的$(\hat{\mu}, \hat{\sigma}^2)$就是参数$(\mu, \sigma^2)$的最大似然估计.

现在问题转化为求$L(\mu, \sigma^2)$的最大值点的问题. 注意上面的x_1, x_2已经是常数了，自变元是(μ, σ^2).

例5.5.4 设X_1, X_2独立同分布，都服从正态分布$N(\mu, \sigma^2)$. 给定观测数据x_1, x_2，试估计μ, σ^2.

解 (X_1, X_2)的联合密度

$$f(x_1, x_2; \mu, \sigma^2) = \frac{1}{2\pi\sigma^2} \exp\left\{ -\frac{1}{2\sigma^2}\left[(x_1-\mu)^2 + (x_2-\mu)^2\right]\right\}$$

是一个扣在xOy平面上的单峰曲面. 根据最大似然思想， μ, σ^2的选择应当使得曲面在(x_1, x_2)处达到最大. 于是使得函数

$$L(\mu, \sigma^2) = f(x_1, x_2; \mu, \sigma^2)$$

达到最大值的$(\hat{\mu}, \hat{\sigma}^2)$就是参数$(\mu, \sigma^2)$的最大似然估计.

现在问题转化为求$L(\mu, \sigma^2)$的最大值点的问题. 注意上面的x_1, x_2已经是常数了，自变元是(μ, σ^2).

定义5.5.2　设随机向量$\boldsymbol{X}=(X_1,X_2,\cdots,X_n)$有联合密度$f(\boldsymbol{x};\boldsymbol{\theta})$，其中$\boldsymbol{\theta}=(\theta_1,\theta_2,\cdots,\theta_m)$是未知参数. 得到$\boldsymbol{X}$的观测值$\boldsymbol{x}=(x_1,x_2,\cdots,x_n)$后，称

$$L(\boldsymbol{\theta})=f(\boldsymbol{x};\boldsymbol{\theta})$$

为$\boldsymbol{\theta}$的似然函数，称$L(\boldsymbol{\theta})$ 的最大值点$\hat{\boldsymbol{\theta}}$ 为 $\boldsymbol{\theta}$ 的最大似然估计(MLE).

设总体X有概率密度$f(x;\boldsymbol{\theta})$，则X的样本 $X_1,X_2,\cdots,X_n$ 有联合密度

$$f(x_1,x_2,\cdots,x_n;\boldsymbol{\theta})=\prod_{j=1}^{n} f(x_j;\boldsymbol{\theta}),$$

基于观测值$\boldsymbol{x}=(x_1,x_2,\cdots,x_n)$的似然函数是

$$L(\boldsymbol{\theta})=\prod_{j=1}^{n} f(x_j;\boldsymbol{\theta}). \tag{5.1}$$

由于

$$l(\boldsymbol{\theta}) = \ln L(\boldsymbol{\theta}) \qquad (5.2)$$

和似然函数(5.1)有相同的最大值点，所以称(5.2)为对数似然函数.
许多问题中，求$L(\boldsymbol{\theta})$的最大值可以通过解似然方程组

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \theta_j} = 0, \quad j = 1, 2, \cdots, m \qquad (5.3)$$

得到.

例5.5.5 设$x_1, x_2, \cdots, x_n$是总体$\mathcal{E}(\lambda)$的样本，求λ的MLE.

解 因为指数分布$\mathcal{E}(\lambda)$的概率密度是

$$f(x;\lambda) = \lambda e^{-\lambda x}, \ x \geqslant 0.$$

基于观测值$\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$的似然函数是

$$L(\lambda) = \lambda^n \exp\Big(-\lambda \sum_{j=1}^{n} x_j\Big).$$

例5.5.5 设$x_1, x_2, \cdots, x_n$是总体$\mathcal{E}(\lambda)$的样本，求λ的MLE.

解 因为指数分布$\mathcal{E}(\lambda)$的概率密度是

$$f(x;\lambda) = \lambda \mathrm{e}^{-\lambda x}, \ \ x \geqslant 0.$$

基于观测值$\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$的似然函数是

$$L(\lambda) = \lambda^n \exp\Big(-\lambda \sum_{j=1}^{n} x_j\Big).$$

对数似然函数是

$$l(\lambda) = n \ln \lambda - \lambda \sum_{j=1}^{n} x_j.$$

由

$$\frac{\partial l}{\partial \lambda} = \frac{n}{\lambda} - \sum_{j=1}^{n} x_j = 0,$$

得到参数λ的MLE 为$\hat{\lambda} = 1/\overline{x}_n$.

通过和例5.4.2比较看出,对于指数总体来讲, MLE和矩估计是一致的.

例5.5.6 设$x_1, x_2, \cdots, x_n$是正态总体$N(\mu, \sigma^2)$的样本观测值，求μ, σ^2的MLE.

解 因为正态分布 $N(\mu, \sigma^2)$ 的概率密度是

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi a}} \exp\left[-\frac{(x-\mu)^2}{2a}\right], \text{ 其中 } a = \sigma^2.$$

所以基于观测值$\mathbf{x} = (x_1, x_2, \cdots, x_n)$的似然函数是

$$L(\mu, a) = \frac{1}{(\sqrt{2\pi a})^n} \exp\left[-\sum_{j=1}^{n} \frac{(x_j-\mu)^2}{2a}\right].$$

例5.5.6 设$x_1, x_2, \cdots, x_n$是正态总体$N(\mu, \sigma^2)$的样本观测值，求μ, σ^2的MLE.

解 因为正态分布 $N(\mu, \sigma^2)$ 的概率密度是

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi a}} \exp\Big[-\frac{(x-\mu)^2}{2a}\Big], \text{ 其中 } a = \sigma^2.$$

所以基于观测值$\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$的似然函数是

$$L(\mu, a) = \frac{1}{(\sqrt{2\pi a})^n} \exp\Big[-\sum_{j=1}^{n} \frac{(x_j-\mu)^2}{2a}\Big].$$

对数似然函数是

$$l(\mu, a) = -\frac{n}{2}\ln a - \sum_{j=1}^{n}\frac{(x_j-\mu)^2}{2a} + c,$$

其中c是常数. 求$l(\mu, a)$的最大值点可以通过解似然方程组

$$\begin{cases} \frac{\partial l}{\partial \mu} = \frac{1}{a}\sum_{j=1}^{n}(x_j-\mu) = 0, \\ \frac{\partial l}{\partial a} = -\frac{n}{2a} + \frac{1}{2a^2}\sum_{j=1}^{n}(x_j-\mu)^2 = 0 \end{cases} \tag{5.4}$$

得到.

从(5.4)解得 μ, $\sigma^2 = a$ 的MLE为

$$\begin{cases} \hat{\mu} = \overline{x}_n, \\ \hat{\sigma^2} = \hat{a} = \frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x}_n)^2. \end{cases}$$

和例5.4.3比较后看出,对于正态总体来讲, MLE和矩估计也是一致的.

例5.5.7 设$x_1, x_2, \cdots, x_n$是总体$\mathcal{U}[a, b]$的样本，求a, b的MLE.

解 均匀分布$\mathcal{U}[a, b]$的概率密度是

$$f(x; a, b) = \frac{1}{b-a}\mathrm{I}[a \leqslant x \leqslant b].$$

给定观测数据$x_1, x_2, \cdots, x_n$，定义

$$x_{(1)} = \min\{x_1, x_2, \cdots, x_n\}, \ x_{(n)} = \max\{x_1, x_2, \cdots, x_n\}. \tag{5.5}$$

例5.5.7 设$x_1, x_2, \cdots, x_n$是总体$\mathcal{U}[a, b]$的样本，求a, b的MLE.

解 均匀分布$\mathcal{U}[a, b]$的概率密度是

$$f(x; a, b) = \frac{1}{b-a}\mathrm{I}[a \leqslant x \leqslant b].$$

给定观测数据$x_1, x_2, \cdots, x_n$，定义

$$x_{(1)} = \min\{x_1, x_2, \cdots, x_n\}, \ x_{(n)} = \max\{x_1, x_2, \cdots, x_n\}. \tag{5.5}$$

可以把a, b的似然函数写成

$$L(a,b) = \frac{1}{(b-a)^n}\prod_{j=1}^{n} \mathrm{I}[a \leqslant x_j \leqslant b]$$
$$= \frac{1}{(b-a)^n}\mathrm{I}[a \leqslant x_{(1)} \leqslant x_{(n)} \leqslant b].$$

要$L(a,b)$达到最大，首先要示性函数 $\mathrm{I}[a \leqslant x_{(1)} \leqslant x_{(n)} \leqslant b] = 1$，这等于要 $a \leqslant x_{(1)}$，$b \geqslant x_{(n)}$. 然后再要求 $1/(b-a)^n$ 最大. 不难看出，这时必须取$a = x_{(1)}$，$b = x_{(n)}$. 所以a, b的MLE分别是

$$\hat{a} = x_{(1)},\ \hat{b} = x_{(n)}. \qquad (5.6)$$

与例5.4.4比较后发现，对于均匀分布来讲，矩估计和MLE不同，为了解MLE (5.6)的表现，利用计算机产生10^7个总体$\mathcal{U}[0.8, 5.2]$ 的样本，利用公式(5.6)和前n个观测数据计算的MLE如下.

n	10	10^2	10^3	10^4	10^5	10^6	10^7
$\hat{a}$	0.8814	0.8434	0.8010	0.8001	0.8000	0.8000	0.8000
$\hat{b}$	4.9806	5.1487	5.1979	5.1991	5.2000	5.2000	5.2000

计算结果也支持强相合结论: $\hat{a} \to a$, $\hat{b} \to b$.

比较例5.4.4和例5.5.7的计算结果后看出，对这批数据来讲，MLE比矩估计表现要好. 但是由于观测数据带有随机性，所以还不能仅从上述计算得出MLE比矩估计好的一般结论. 尽管如此，因为MLE的计算利用了分布信息，而矩估计没有充分利用分布的信息，所以MLE的估计精度一般不低于矩估计的估计精度.

对于均匀分布$U[0,b]$来讲，矩估计是$\tilde{b}=2\overline{x}_n$，最大似然估计是$\hat{b}=x_{(n)}$. 这两者明显是不一样的. 为了解哪个估计更准确一些，我们用计算机产生10 000个在区间$[0,2.8]$上均匀分布的随机数(独立同分布随机变量的观测值)，使用前n个随机数计算出矩估计$\tilde{b}$和MLE $\hat{b}$的估计误差如下($b=2.8$):

n	10	30	60	100	1 000	10 000
$\lvert\tilde{b}-b\rvert$	0.231	0.315	0.313	0.063	0.096	0.003
$\lvert\hat{b}-b\rvert$	0.249	0.249	0.214	0.008	0.004	0.000 02

我们称上述的方法为计算机模拟试验方法. 从上述模拟试验看出，MLE $\hat{b}$ 的表现似乎要比矩估计$\tilde{b}$好. 但是数据的产生带有随机性，所以一次模拟试验显然不够.

对于均匀分布$\mathcal{U}[0,b]$来讲，矩估计是$\tilde{b}=2\overline{x}_n$，最大似然估计是$\hat{b}=x_{(n)}$. 这两者明显是不一样的. 为了解哪个估计更准确一些，我们用计算机产生10 000个在区间[0, 2.8]上均匀分布的随机数(独立同分布随机变量的观测值)，使用前n个随机数计算出矩估计$\tilde{b}$和MLE $\hat{b}$的估计误差如下($b=2.8$):

n	10	30	60	100	1 000	10 000
$\lvert\tilde{b}-b\rvert$	0.231	0.315	0.313	0.063	0.096	0.003
$\lvert\hat{b}-b\rvert$	0.249	0.249	0.214	0.008	0.004	0.000 02

我们称上述的方法为计算机模拟试验方法. 从上述模拟试验看出，MLE $\hat{b}$ 的表现似乎要比矩估计$\tilde{b}$好. 但是数据的产生带有随机性，所以一次模拟试验显然不够.

为了克服数据的随机性，我们将上述模拟试验独立重复1 000次. 用$\tilde{b}_j$ 表示第j次模拟计算得到的矩估计$\tilde{b}$, 用$\hat{b}_j$表示第j次模拟计算得到的最大似然估计$\hat{b}$. 再定义$m = 1\,000$次模拟的平均

$$M(\tilde{b}) = \frac{1}{m}\sum_{j=1}^{m}\tilde{b}_j, \quad M(\hat{b}) = \frac{1}{m}\sum_{j=1}^{m}\hat{b}_j$$

和$m = 1\,000$次模拟的样本标准差

$$\text{std}(\tilde{b}) = \sqrt{\frac{1}{m-1}\sum_{j=1}^{m}\left[\tilde{b}_j - M(\tilde{b})\right]^2},$$

$$\text{std}(\hat{b}) = \sqrt{\frac{1}{m-1}\sum_{j=1}^{m}\left[\hat{b}_j - M(\hat{b})\right]^2}.$$

结论如下表所示.

n	10	30	60	100	1000	10000
$\lvert M(\tilde{b})-b\rvert$	0.007	0.014	0.011	0.008	0.002	0.0006
$\mathrm{std}(\tilde{b})$	0.511	0.293	0.203	0.162	0.048	0.016
$\lvert M(\hat{b})-b\rvert$	0.270	0.096	0.048	0.029	0.003	0.0002
$\mathrm{std}(\hat{b})$	0.2398	0.0989	0.0466	0.0281	0.0028	0.0003

从上述计算结果中看出, $|M(\tilde{b})-b|$普遍小于$|M(\hat{b})-b|$, 所以从偏差的角度讲, 矩估计好一些. 这个结果是和$\mathrm{E}\tilde{b}=b$, $\mathrm{E}\,\hat{b}<b$一致的. 但是另一方面, $\mathrm{std}(\hat{b})$普遍小于$\mathrm{std}(\tilde{b})$, 说明最大似然估计的稳定性更好一些. 由于1000次重复试验已经能够较好地克服随机因素的影响, 所以可以认为上述模拟试验的结果是可信的.

结论如下表所示.

n	10	30	60	100	1000	10000
$\lvert M(\tilde{b})-b\rvert$	0.007	0.014	0.011	0.008	0.002	0.0006
$\mathrm{std}(\tilde{b})$	0.511	0.293	0.203	0.162	0.048	0.016
$\lvert M(\hat{b})-b\rvert$	0.270	0.096	0.048	0.029	0.003	0.0002
$\mathrm{std}(\hat{b})$	0.2398	0.0989	0.0466	0.0281	0.0028	0.0003

从上述计算结果中看出, $|M(\tilde{b})-b|$普遍小于$|M(\hat{b})-b|$, 所以从偏差的角度讲, 矩估计好一些. 这个结果是和$\mathrm{E}\tilde{b}=b$, $\mathrm{E}\,\hat{b}<b$一致的. 但是另一方面, $\mathrm{std}(\hat{b})$普遍小于$\mathrm{std}(\tilde{b})$, 说明最大似然估计的稳定性更好一些. 由于1 000次重复试验已经能够较好地克服随机因素的影响, 所以可以认为上述模拟试验的结果是可信的.