抽样与估计

公选课 可计量的社会

何濯羽印

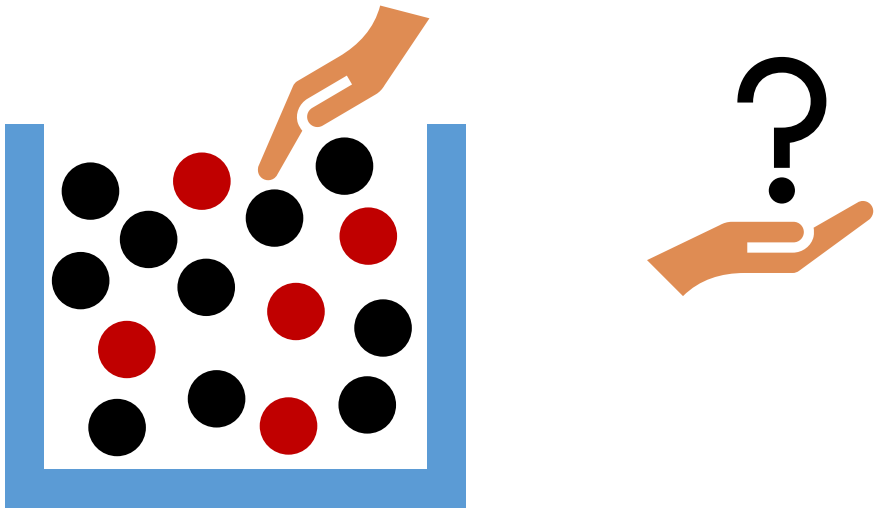
目录

1

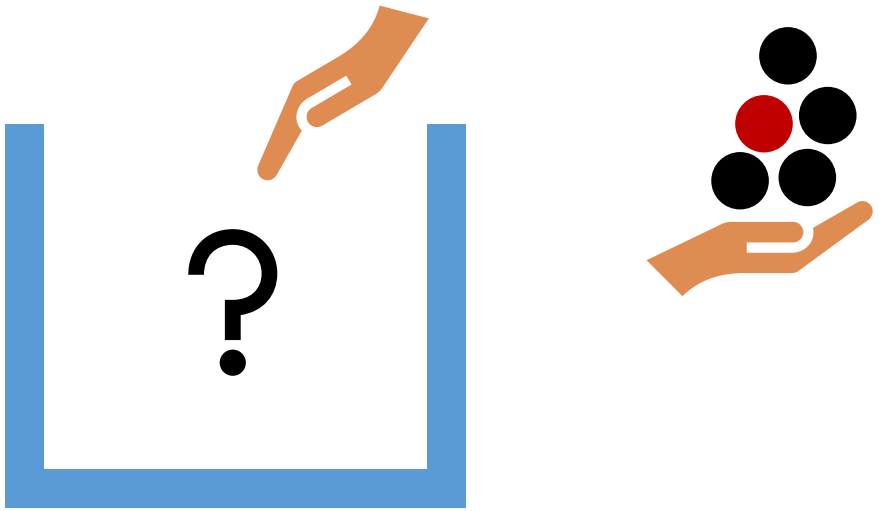
概率论 VS 统计学

Probability Theory VS Statistics

概率论 VS 统计学



概率论的问题：已知盒子中黑球与红球的个数，你随机抽出 1 个球，这个球为红球的概率是多少？



统计学的问题：你从盒子中随机抽出 5 个球后，发现是 4 个黑球和 1 个红球，请问盒子中有几个黑球和几个红球？

2

总体

Population

什么是总体?

总体（statistical population，简称 population）是指由许多拥有某种共同性质的事物组成的集合。这些事物可以是客观存在的物体（例如银河系中所有的恒星），也可以是抽象的、无法直接观测的事物（例如学生的数学天赋）。在统计学（或应用统计学的学科）中，我们的研究目标通常是了解总体中某一性质服从的概率分布或者其概率分布的部分特征。

我们可以用**随机变量**（random variable）表示总体中事物的性质，随机变量的可能取值代表事物具有的特定性质。例如，在一个由全世界人类组成的总体中，我们可以用 X 表示“性别”：

- $X = 0$ 代表某人是男性；
- $X = 1$ 代表某人是女性；
- $X = 2$ 代表某人是其它性别。

注：性别是一个社会学概念而非生物学概念。性别不是二元的——这是所有将统计学应用于社会学（或相关领域）的学者应该具有的认识。

频率和概率

在总体中，我们可以用随机变量特定取值出现的**频率**（frequency rate）表示随机变量相应取值出现的**概率**（probability）。因此，如果我们的研究目标是了解总体中某一性质的概率分布，那么我们只需要计算代表该性质的随机变量的所有取值出现的频率即可。

沿用上例，在由全世界人类组成的总体中，我们可以用 X 表示“性别”，而 $X = x$（$x \in \{0, 1, 2\}$）出现的次数（或称**频数**，frequency）是 N_x，那么在该总体中，某个个体的性别是男性的概率为

$$\Pr(X = 0) = \frac{N_0}{N_0 + N_1 + N_2} = \frac{N_0}{N}$$

其中，$N = N_0 + N_1 + N_2$ 是总体中个体的总数。

总体均值

如果我们的研究只需要了解总体中某一性质所服从的概率分布的部分特征（如数学期望和方差），我们应该如何操作呢?

回忆离散型随机变量的数学期望的定义： $E(X) = \sum_{x \in \mathcal{X}} x \cdot \Pr(X = x)$

在我们可观测总体的情况下，概率即频率，故

$$\boldsymbol{\mu_X = \sum_{x \in \mathcal{X}} x \cdot \frac{N_x}{N} = \frac{1}{N} \sum_{i=1}^{N} x_i}$$

其中，x_i 代表总体中随机变量 X 的第 i 个取值。μ_X 是 X 的数学期望，也被称为**总体均值**（population mean）。在 Excel 中，我们可以用 AVERAGE( ) 函数计算总体均值。

总体方差

相似地，回忆离散型随机变量的方差的定义： $D(X)=\sum_{x\in\mathcal{X}}(x-\mu_X)^2\cdot\Pr(X=x)$

在我们可观测总体的情况下，概率即频率，故

$$\boldsymbol{\sigma_X^2=\sum_{x\in\mathcal{X}}(x-\mu_X)^2\cdot\frac{N_x}{N}=\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_X)^2}$$

σ_X^2 是 X 的方差，也被称为**总体方差**（population variance）。在 Excel 中，我们可以用 `VAR.P( )` 函数计算总体方差。

3

样本与抽样

Sample and Sampling

什么是样本?

在多数研究中，由于总体过于庞大，直接观测总体的数据会耗费许多时间与资金。于是，研究人员退而求其次，只从总体中抽取部分个体进行观测。

- 从总体中抽取部分个体的过程被称为“**抽样**”（sampling）。
- 从总体中抽取出来的个体共同组成的集合被称为“**样本**”（sample）。
- 研究人员观测到的每一个个体的数值或事实被称为“**实测值**”（observation）。
- 样本中包含的个体数量被称为“**样本容量**”（sample size），或简称“**样本量**”。样本量就相当于实测值的个数（number of observations）。

如何抽样?

抽样过程主要包括以下几个步骤:

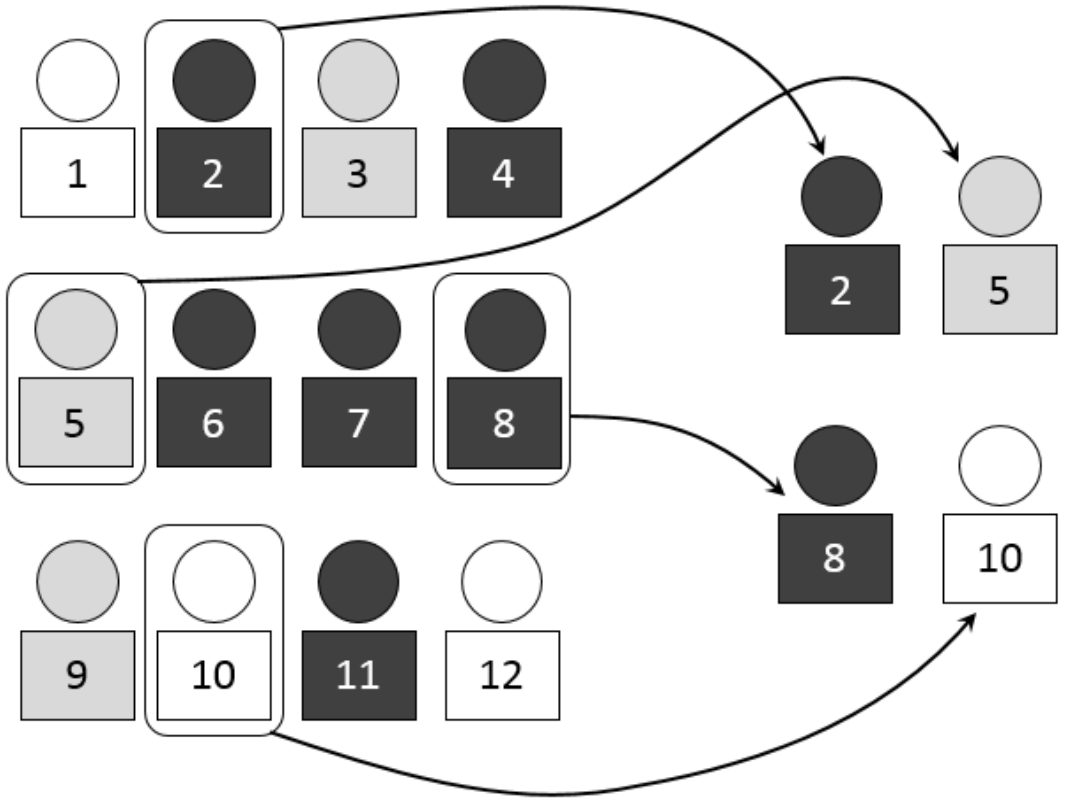
1) 定义总体;
2) 确定抽样框（对可以选择作为样本的总体单位列出名册或排序编号，以确定总体的抽样范围和结构）；
3) 确定抽样方法（简单随机抽样、分层抽样、聚类抽样、系统抽样）；
4) 决定样本量;
5) 实施抽样计划;
6) 抽样与数据收集;
7) 回顾抽样过程。

简单随机抽样

在实施**简单随机抽样**（simple random sampling）时，研究人员从总体 N 个单位中随机地抽取 n 个单位作为样本，使得每一个单位都有相同的概率被抽中。

通过简单随机抽样得到的样本具有的特点是：

- 每个样本单位被抽中的**概率相等**；
- 每个样本单位**相互独立**（statistically independent）。

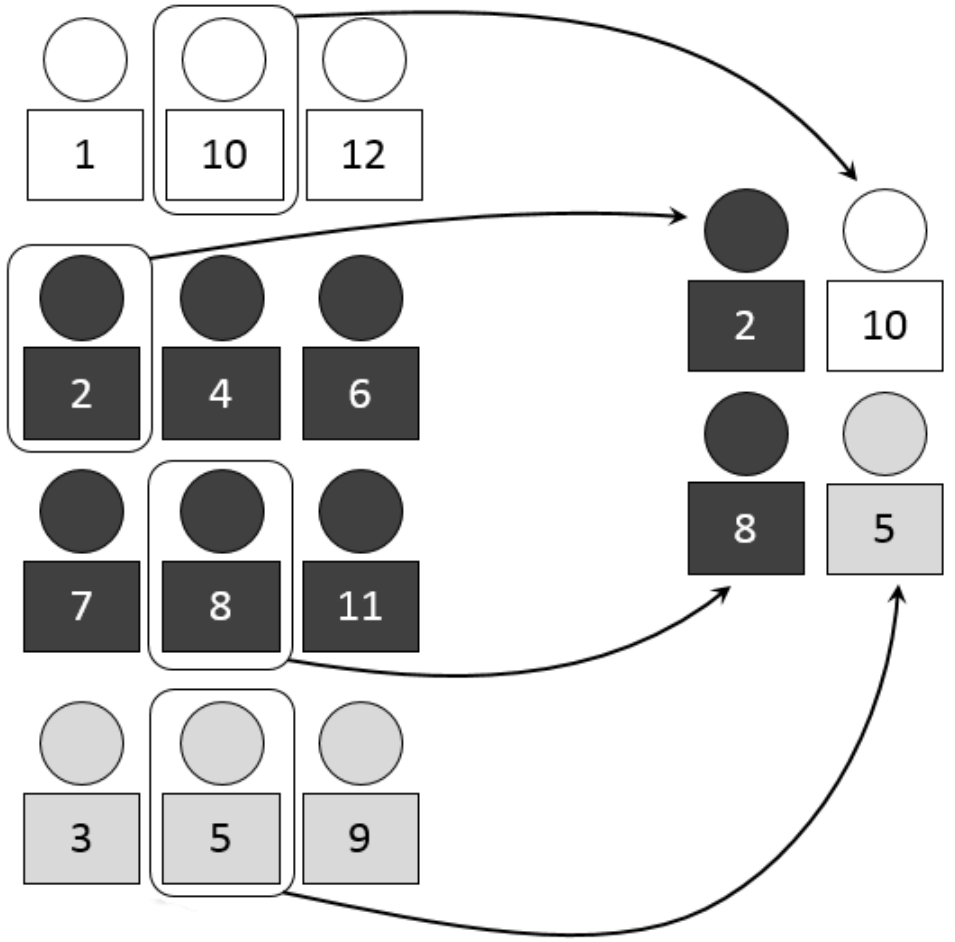


分层抽样

在实施**分层抽样**（stratified sampling）时，研究人员将总体按某种特征划分为不同的层（stratum），然后按照一定比例从每个层中进行简单随机抽样。

不同层中抽取的个体个数可能不同。例如，**等比分配法**要求

$$\frac{\text{来自A层的样本量}}{\text{总样本量}} = \frac{\text{总体中位于A层的个体数}}{\text{总体个体数}}$$

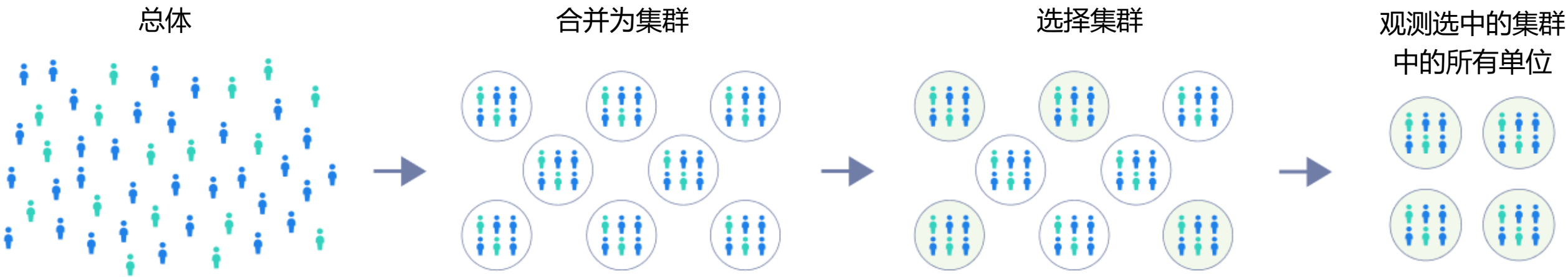


分层抽样的**优点**是：

- 可对不同的层进行独立的分析；
- 可以根据各层的具体情况对不同的层采用不同的抽样方法；
- 分层后每一层的同质性增加，因而使得实测值的变异度减小，各层的抽样误差随之减小。

聚类抽样

在实施**聚类抽样**（cluster sampling）时，研究人员将总体中若干个单位合并为集群（cluster），然后直接选择集群，对选中集群的所有单位进行观测。



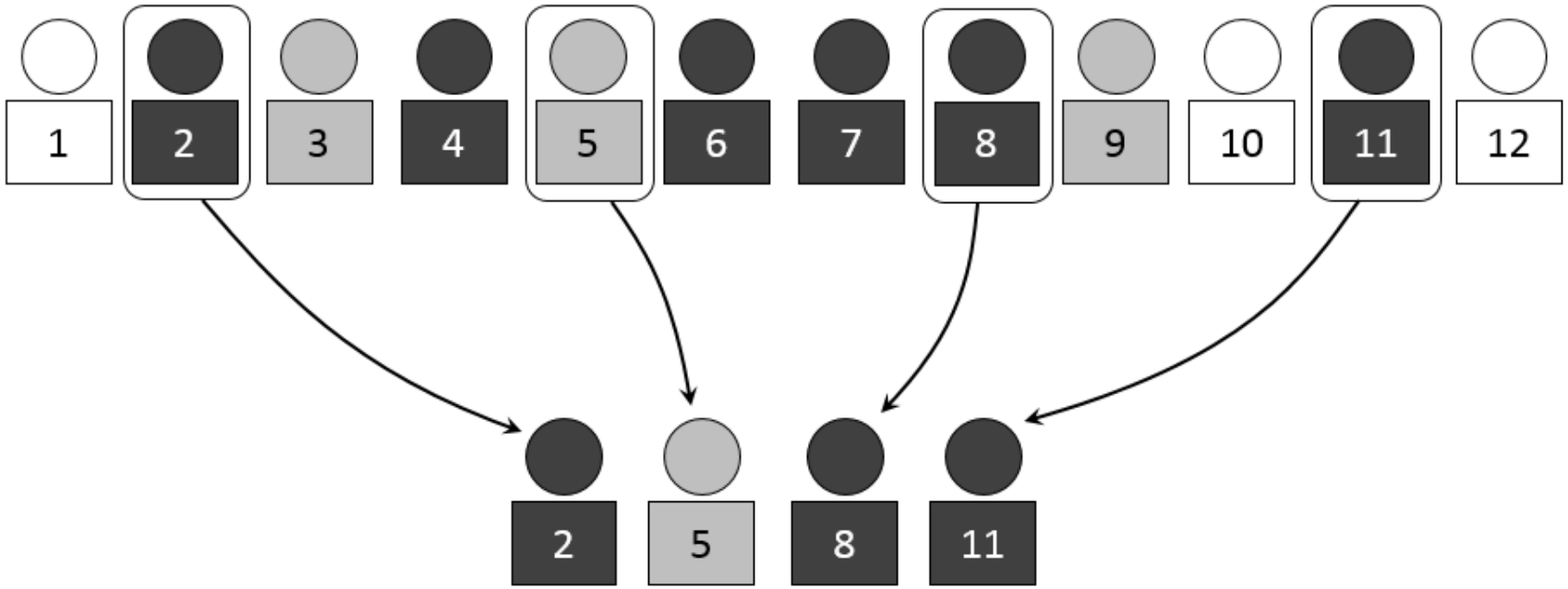
聚类抽样的**优点**是：抽样工作被极大地简化。

聚类抽样的**缺点**是：估计的精度较差。

系统抽样

在实施**系统抽样**（systematic sampling）时，研究人员将总体中的所有单位按一定顺序排列，在规定的范围内随机抽取一个单位作为初始单位，然后按事先规定好的规则确定其他样本单位。

例：我们先在数字 1 到 N 中随机选择了数字 r 作为初始单位，然后依次抽取序号为 $r+k$、$r+2k$、$r+3k$ 等的单位。下图中，$r=2$，$k=3$，$N=12$。



独立同分布的假设

在统计学中，有一个十分常见的假设：**独立同分布**（independently and identically distributed，简记为 i.i.d.）。

这一假设要求我们的样本数据 $\{X_1, X_2, \dots, X_n\}$ 同时具有以下两个特征：

- X_i 与 X_j 相互独立，$\forall i \neq j$。即所有样本单位的随机变量 X 相互独立。
- $X_i \sim F_X$，$\forall i$。即所有样本单位的随机变量 X 服从同一个概率分布。

使得这个假设成立的要求十分严苛，在现实中并不常见，但它可以简化统计学中许多的计算与证明。

4

估计

Estimation

推断性统计

如前所述，我们通常无法直接观测总体，所以只能通过抽样，利用样本的特征去推测总体的特征。

这一统计学的分支被称为“**推断性统计**”（inferential statistics）。

在推断性统计中，

- 我们把推测的目标称为“**被估量**”（estimand）。
- 我们把推测出被估量取值的方法称为“**估计量**”（estimator）。
- 我们把估计量采用实际数据计算得到的数值称为“**估计值**”（estimate）。

举例

我们想了解成都东软学院 2023 届毕业生的平均年收入。总体是成都东软学院 2023 届所有毕业生的年收入（用随机变量 X 表示），样本是我们随机抽取的 100 名成都东软学院 2023 届毕业生的年薪（用 $X_1, X_2, \ldots, X_{100}$ 表示）。在这个研究中，

- **被估量（estimand）**是成都东软学院 2023 届所有毕业生的平均年薪（总体均值 μ_X）。
- **估计量（estimator）**可以是样本中毕业生年薪的平均数，也可以是中位数，还可以是最大值。实际上，我们拥有无数种估计量。
- 假设我们选择样本平均数，即 $\frac{1}{100}\sum_{i=1}^{100} X_i$，作为我们的估计量。将100个实测值代入该估计量后，我们得到一个数值：60,000元——这个数值就是我们的**估计值（estimate）**。

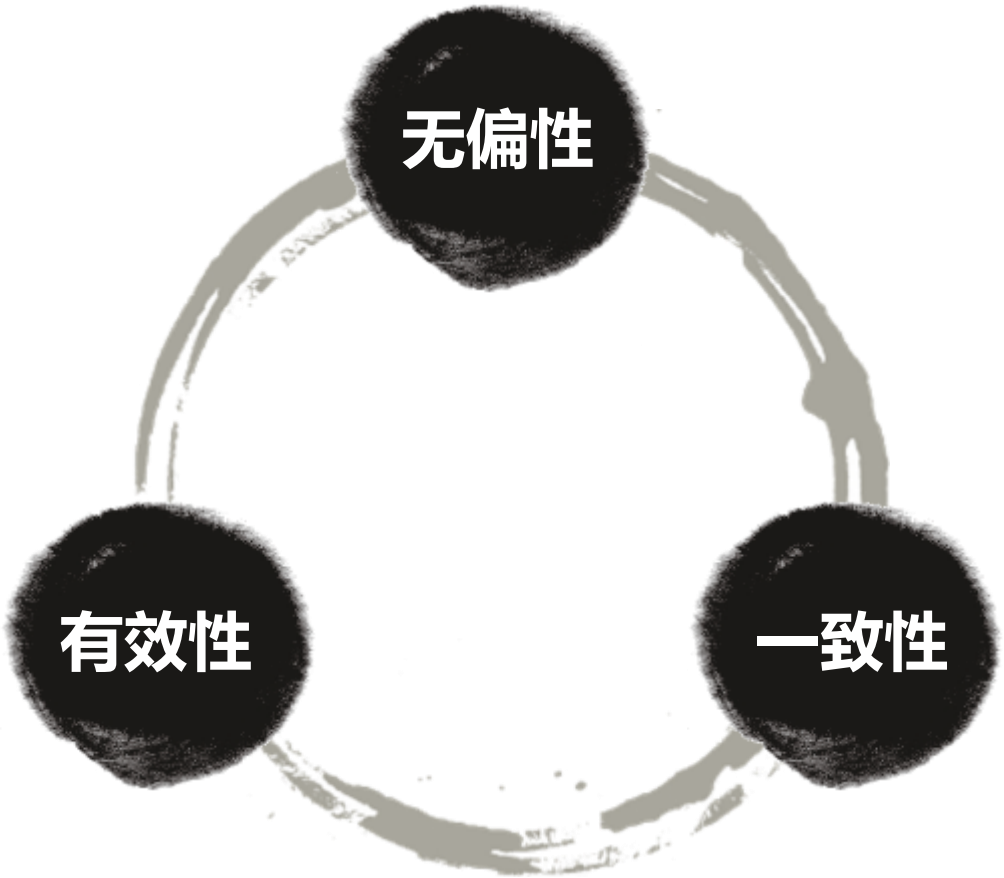
如上所述，面对任何一个被估量，我们都可以构造出无数种估计量。

估计量的构造法

有一种被广泛使用的估计量构造法：**样本类似原则**（sample analogue principle）——使用一个具有与总体参数相同性质的样本统计量。例如，

	总体参数（无法直接观测）	估计量（样本的函数，可直接观测）
数学期望	$E(X)=\mu_X=\sum_{x\in\mathcal{X}} x\cdot \Pr(X=x)$	
样本均值		$\overline{X}=\frac{1}{n}\sum_{i=1}^{n} X_i$
方差	$D(X)=\sum_{x\in\mathcal{X}}(x-\mu_X)^2\cdot \Pr(X=x)$	
样本方差		$\hat{\sigma}_X^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$
协方差	$Cov(X,Y)=E\{[X-E(X)]\cdot[Y-E(Y)]\}$ $=\sum_{(x,y)}(x-\mu_X)\cdot(y-\mu_Y)\cdot \Pr(X=x,Y=y)$	
样本协方差		$\hat{\sigma}_{XY}=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})$

“好”的估计量



既然我们可以构建无数种估计量，那么是否存在最好的估计量呢？或者说，如何判断一个估计量是“好”的估计量？统计学家们从三个角度评判一个估计量是否足够“好”：

- **无偏性**（unbiasedness）：估计量的数学期望等于被估量的真实值。
- **有效性**（efficiency）：一个估计量的方差越小说明该估计量越有效。
- **一致性**（consistency）：当样本容量趋向于无穷大时，估计量无限趋近于被估量的真实值。

注： “一致性”的严格定义以及相关讨论涉及概率极限（probability limit）和渐近理论（asymptotic theory），国内外多数高校会在硕士/博士课程中教授相关知识。因此，在此不做任何介绍。

估计量的本质

估计量的本质是随机变量！

例如，我们常把随机变量 X 的**样本平均数**（或称**样本均值**，sample mean）用作其总体均值的估计量。

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

其中，$\{X_i\}_{i=1}^{n}$ 是我们抽样得到的样本，n 是样本容量。在未代入实际数据之前，$X_1, X_2, \ldots, X_n$ 是 n 个代表不同样本个体的同一性质的随机变量。很明显，$\overline{X}$ 是样本 $\{X_i\}_{i=1}^{n}$ 的函数——$\overline{X}$ 是 n 个随机变量之和除以样本容量得到的商，所以 $\overline{X}$ 也是随机变量。

注：样本的函数被统称为“统计量”（statistic）。估计量是用于估计总体参数的统计量。

无偏估计量

我们用 θ 表示被估量的真实值，用 $\hat{\theta}$ 表示 θ 的一个估计量。我们定义**偏度**（bias）为估计量的数学期望与被估量真实值的差：

$$\text{bias} = E(\hat{\theta}) - \theta$$

如果 $E(\hat{\theta}) = \theta$，即偏度为 0，那么我们称 $\hat{\theta}$ 是 θ 的**无偏估计量**（unbiased estimator）。

例：在**独立同分布**（i.i.d.）的假设下，样本均值是总体均值的无偏估计量。

$$E(\overline{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n} \cdot E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} \cdot n\mu_X = \mu_X$$

有效估计量

我们用 $\hat{\theta}_1$ 和 $\hat{\theta}_2$ 表示 θ 的两个不同估计量。如果 $E(\hat{\theta}_1) = E(\hat{\theta}_2) = \theta$ 且 $D(\hat{\theta}_1) < D(\hat{\theta}_2)$，那么我们称 $\hat{\theta}_1$ 比 $\hat{\theta}_2$ **更有效**（more efficient）。

例：在**独立同分布**（i.i.d.）的假设下，比较样本均值 $\overline{X}$ 和第一个样本单位 X_1，哪一个作为总体均值的估计量更有效?

首先，很容易证明两个估计量均是总体均值的无偏估计量：$E(\overline{X}) = E(X_1) = \mu_X$。

其次，

$$D(\overline{X}) = D\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} \cdot D\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} D(X_i) = \frac{1}{n^2} \cdot n\sigma_X^2 = \frac{\sigma_X^2}{n}$$

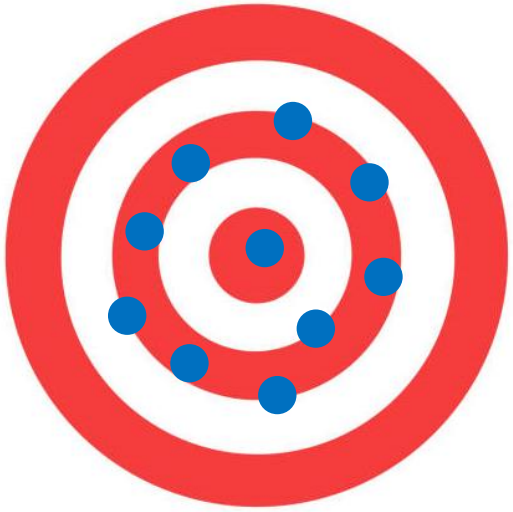
$$D(X_1) = \sigma_X^2 < \frac{\sigma_X^2}{n}$$

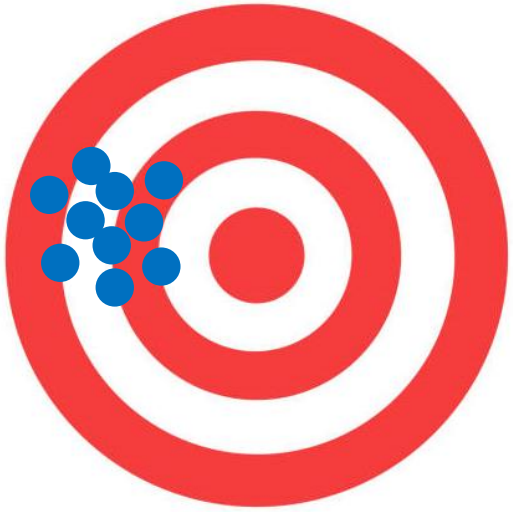
因此，$\overline{X}$ 比 X_1 更有效。我们还发现：当样本量 n 增大时，$\overline{X}$ 的有效性会增强（即方差会减小）。

无偏 VS 有效

当我们在判定一个估计量的无偏性和有效性时，一定是先检查无偏性。再确认其是无偏估计量后，我们才会继续检查有效性。换言之，如果一个估计量是有偏的，那么就没有必要再检查它的有效性了。

我们可以把估计量的无偏性类比为一名弓箭手的瞄准能力，而有效性对应的则是弓箭手的双手颤抖程度。如果一名弓箭手的瞄准能力非常差，即使他的双手完全不会颤抖，他也很难射中靶心。





总体方差的估计量

总体均值的无偏估计量是样本均值。那么，总体方差的无偏估计量是什么？样本方差是它的无偏估计量吗？

$$\hat{\sigma}_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

很遗憾，即使在独立同分布的假设下，这个非常符合人类直觉的估计量也不是总体方差 σ_X^2 的无偏估计量。实际上，

$$E(\hat{\sigma}_X^2) = \frac{n-1}{n}\sigma_X^2 < \sigma_X^2$$

证明过程较复杂，在此不做展示。

即样本方差是总体方差的有偏估计量，且偏度为负数。

根据上面的结果，我们可以轻松地构建出一个无偏估计量：

$$s_X^2 = \frac{n}{n-1}\hat{\sigma}_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

这个式子也是多数统计软件（包括 R、Python、Stata）默认的方差估计公式。在 Excel 中，`VAR.S( )` 函数使用的就是这个公式。

方差估计

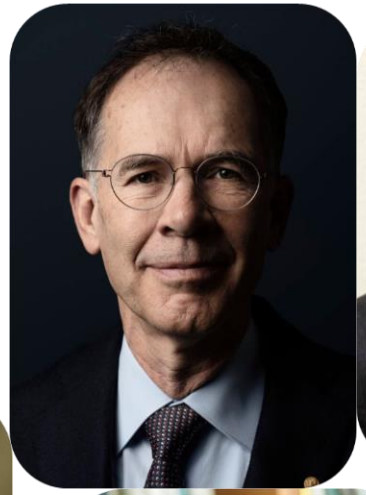
“如何构建（不同情况中）方差的无偏、有效、一致估计量”是统计学界的一个长期难题。

向所有曾经以及继续在方差估计（variance estimation）领域做贡献的学者致敬！

Alberto Abadie

Bruce E. Hansen

Guido W. Imbens



Halbert White

Manuel Arellano

James G. MacKinnon

Matthew D. Webb

Kenneth D. West

Whitney K. Newey

Seojeong Lee

谢谢

何濯羽印