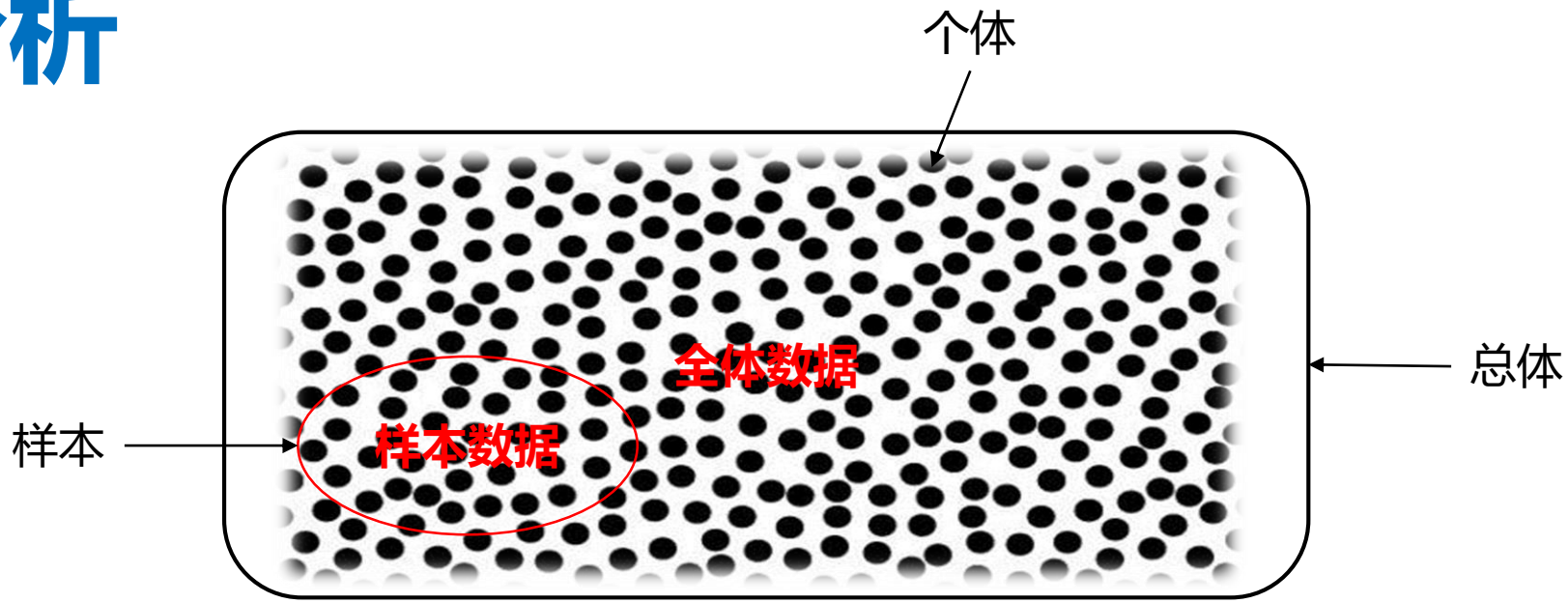
数据科学基础

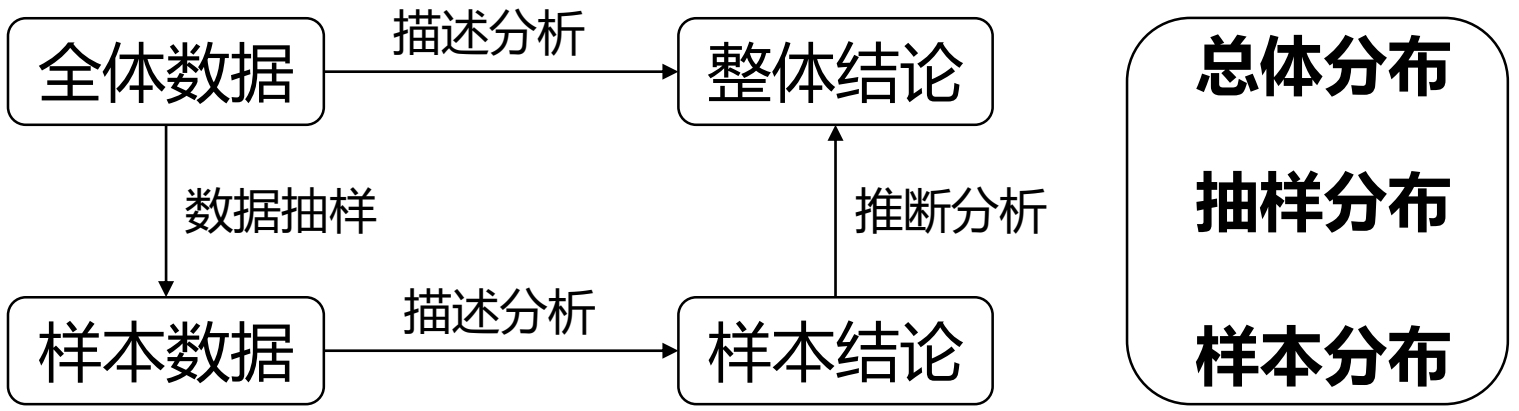
Foundations of Data Science

4.2 矩与抽样分布

陈振宇
南京大学智能软件工程实验室
www.iselab.cn

数据分析





总体分布

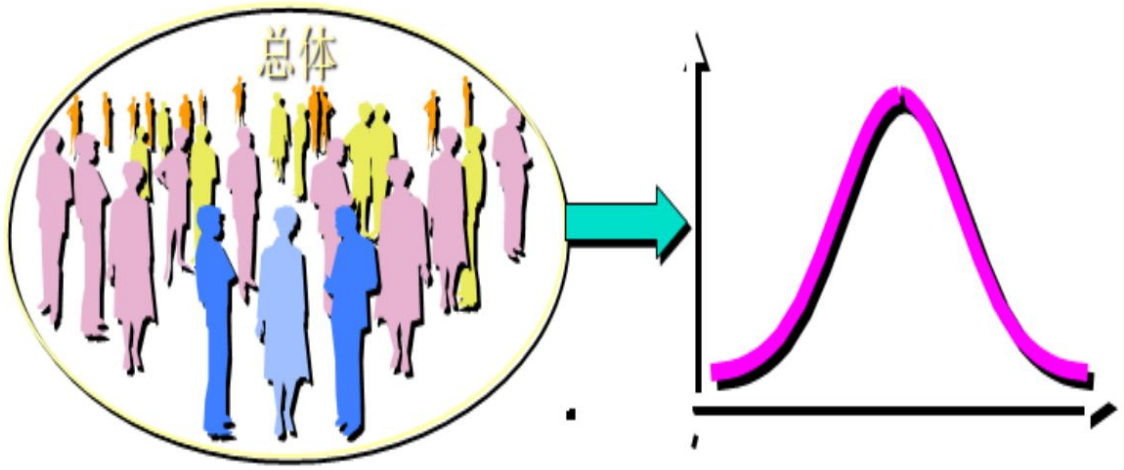
抽样分布

样本分布

更多详情，请咨询课程助教。

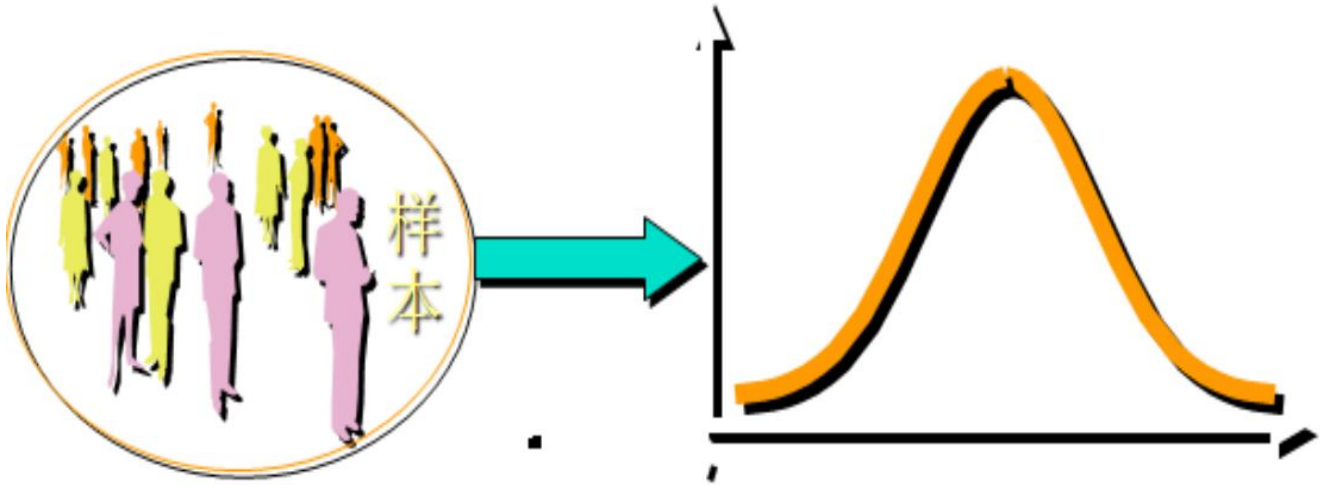
总体分布

- 总体分布：总体中所有个体观察值所构成的分布
  - 注意：不一定是正态分布
- 总体分布通常是未知的：
  - 分布形式和参数都未知
  - 分布形式已知但参数未知



样本分布

- 一个样本中个体观察值的分布，样本分布通常也称经验分布
- 当样本容量n逐渐增大时，样本分布逐渐接近总体的分布。
- 当n为总体中个体数量时与总体分布完全一致。



统计量

- 统计量是样本的函数
- 统计量具有二重性：抽样前是随机变量，抽样后是具体数据
- 统计量通常只依赖于样本，不依赖于总体分布中的未知参数
- 样本矩是最常用的样本统计量

样本矩统计量

【样本矩，Sample Moment】

设样本$X_1,\cdots,X_n$，样本的k阶矩定义如下：

样本k阶原点矩：

$$A_k = \frac{1}{n}\sum_{i=1}^{n} {X_i}^k$$

样本k阶中心矩：

$$B_k = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^k$$

常用样本矩统计量

- 样本均值为样本的一阶原点矩A_1，它代表样本的平均程度，记为$\bar{X}$

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

- 样本方差为修正后的二阶中心矩，即$\frac{n}{n-1}B_2$，它代表样本的分散程度，记为S^2

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

抽样分布定理

【定理】设某总体的均值为μ，方差为σ^2。X_1，…，X_n为总体一样本，$\bar{X}$为样本均值，S^2为样本方差，则

(1) $\mathrm{E}(\bar{X}) = \mu$

(2) $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$

(3) $\mathrm{E}(S^2) = \sigma^2$

总体分布的矩

【例】设一总体为四本书，四本书平均每页的错别字为$x_1 = 1$，$x_2 = 2$，$x_3 = 3$，$x_4 = 4$

则总体的均值如下：在此处键入公式。

$$\mu = \frac{\sum_{i=1}^{4} x_i}{4} = 2.5$$

总体的方差如下：

$$\sigma^2 = \frac{\sum_{i=1}^{4} (x_i - \mu)^2}{4} = 1.25$$

从总体分布到抽样分布

现从总体中抽取$n = 2$的简单随机样本，在有放回抽样条件下，共有16个样本。

所有样本的结果为

1，1	1，2	1，3	1，4
2，1	2，2	2，3	2，4
3，1	3，2	3，3	3，4
4，1	4，2	4，3	4，4

则样本均值$\bar{X}$的所有结果为

1	1.5	2	2.5
1.5	2	2.5	3
2	2.5	3	3.5
2.5	3	3.5	4

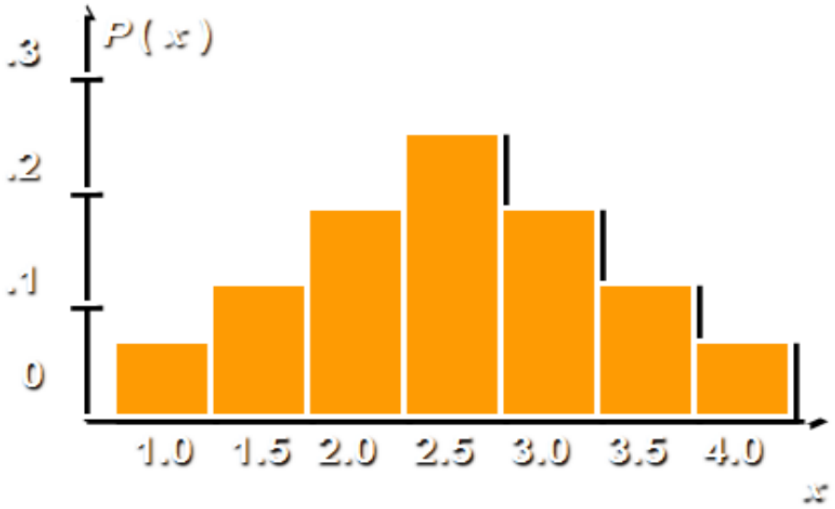
样本均值的频数

$\bar{x}$	1	1.5	2	2.5	3	3.5	4
f_i	1	2	3	4	3	2	1

抽样分布与总体分布

总体分布

抽样分布



更多详情，请咨询课程助教。

抽样分布的矩

设总体为$\{1，2，3，4\}$，则总体的均值和方差分别为$\mu = 2.5$和$\sigma^2 = 1.25$。从总体中抽取$n = 2$的简单随机样本，共有16个样本。

所有样本的结果和相应样本均值

1，1	1，2	1，3	1，4
2，1	2，2	2，3	2，4
3，1	3，2	3，3	3，4
4，1	4，2	4，3	4，4

1	1. 5	2	2. 5
1. 5	2	2. 5	3
2	2. 5	3	3. 5
2. 5	3	3. 5	4

样本均值$\bar{X}$的均值为

$$\frac{1}{16}\sum_{i=1}^{16}\bar{x_i} = 2.5$$

样本均值$\bar{X}$的方差为

$$\frac{1}{16}\sum_{i=1}^{16}(\bar{x_i} - 2.5)^2 = \frac{5}{8} = 0.625$$

抽样分布

- 样本统计量（样本均值$\bar{X}$，样本方差s^2）是随机变量。
- 样本统计量的概率分布是一种理论分布，称为抽样分布。
- 在重复选取容量为n的样本时，由该统计量的所有可能取值形成的概率分布。
- 抽样分布提供了样本统计量长远而稳定的信息，是进行推断的理论基础，也是抽样推断科学性的重要依据。

更多详情，请咨询课程助教。