抽样与抽样分布

目录

CONTENTS

3 案 例 说 明

1. 美国总统选取活动中，政治团体的领导需要对民意支持率做一个调查。选取得克萨斯州400名选民组成一个样本，其中有160人表示支持。因此选民总体中支持的比率就是160/400=0.4。

2. 一个轮胎制造商正在考虑生产一种新设计开发的轮胎，为了对轮胎的平均寿命做出估计，制造商生成了120个这种新型轮胎组成了样本用于检测，检测结果表明样本均值为36500英里。于是，这种新型轮胎总体的平均使用寿命估计值为36500英里。

解 释 说 明

抽样的结果提供的仅仅是相应总体特征值的估计，我们并不期望它是百分之百正确的。原因很简单，样本只包含了总体的一部分，一定会有误差产生。利用恰当的抽样方法，我们可以尽可能的给出一个关于总体特征的好的估计。

01

抽样

5 工 作 概 述

Elecronics Associates公司的抽样问题

EAI公司的人事部经理被分配一项任务，要求为公司2500名管理人员制定一份简报，内容包括管理人员的平均年薪和已完成公司管理培训计划的管理人员所占的比率。

2500名管理人员构成此项研究的总体。现在，假设我们无法从公司的数据库中获得全部EAI管理人员的必要信息。那么我们能否不用总体的2500份数据，而是用一个样本，从而获取对这些总体参数的估计呢？

假设样本的大小为30，我们现在从这30名管理人员的样本入手，探究利用样本研究EIA问题的可能性。

参　　数

根据EAI文件数据，我们可以得出总体均值和总体标准差，以及完成培训计划的总体比率

- μ =51800 美元
- σ =4000 美元
- $p = \frac{1500}{2500} = 0.6$

6 抽　　样

从有限总体的抽样

简单随机抽样

从容量为N的有限总体中抽取一个容量为 n 的样本，如果容量为 n 的每一个可能的样本都以相等的概率被抽出，则称该样本为简单随机抽样。

有放回抽样

如果在选取样本时，对已经出现过的随机数仍选入样本，某些管理人员可能在样本中被两次或两次以上的包括进来，则我们进行的是有放回抽样

无放回抽样

对已经入选的样本，不再进行第二次抽样，样本中的个人都是唯一的，没有重复

从无限总体的抽样

总投容量无限大或者总体中的个体是由一个正在运行的过程产生的，从而生成的个数数量是无限的，因此无法得到总体中所有的个体清单。比如：正在作业的产品生产线，你无法知道产品的全部数量

随机样本

1. 抽取的每个个体来自同一总体
2. 每个个体的抽取是独立的

案例：

生产过程中所生产的产品数量是无限的。抽样总体由正在运行的生产过程中生产的全部产品，而不仅仅是那些已经生产的产品组成。更具体一点，为判断生产线是正常运行还是由于机器故障使得生产线的产品不合格，抽检员要关心的条件是“抽取的每个个体来自同一样本”是否成立。因此他必须在近似相同的时点选择产品，这样才能避免抽取的某些产品是生产线正常运行产生的，而另外一个是由于故障产生的

02

点估计

8 点 估 计

30名管理人员样本数据

Annual Salary ($)	Management Training Program	Annual Salary ($)	Management Training Program
$x_1 = 49{,}094.30$	Yes	$x_{16} = 51{,}766.00$	Yes
$x_2 = 53{,}263.90$	Yes	$x_{17} = 52{,}541.30$	No
$x_3 = 49{,}643.50$	Yes	$x_{18} = 44{,}980.00$	Yes
$x_4 = 49{,}894.90$	Yes	$x_{19} = 51{,}932.60$	Yes
$x_5 = 47{,}621.60$	No	$x_{20} = 52{,}973.00$	Yes
$x_6 = 55{,}924.00$	Yes	$x_{21} = 45{,}120.90$	Yes
$x_7 = 49{,}092.30$	Yes	$x_{22} = 51{,}753.00$	Yes
$x_8 = 51{,}404.40$	Yes	$x_{23} = 54{,}391.80$	No
$x_9 = 50{,}957.70$	Yes	$x_{24} = 50{,}164.20$	No
$x_{10} = 55{,}109.70$	Yes	$x_{25} = 52{,}973.60$	No
$x_{11} = 45{,}922.60$	Yes	$x_{26} = 50{,}241.30$	No
$x_{12} = 57{,}268.40$	No	$x_{27} = 52{,}793.90$	No
$x_{13} = 55{,}688.80$	Yes	$x_{28} = 50{,}979.40$	Yes
$x_{14} = 51{,}564.70$	No	$x_{29} = 55{,}860.90$	Yes
$x_{15} = 56{,}188.20$	No	$x_{30} = 57{,}309.10$	No

样本统计量

$\bar{x} = 51814$ 美元

$s = 3348$ 美元

$p = 0.63$

点估计量

我们称上述三个值为对应总体值的点估计量。
点估计量与总体参数的真实值是有一定差异的，
这个差异是可预期的

应用建议

当利用样本去推断总体时，我们应该确保所设计的研究中抽样总体与目标总体时高度一致的

抽样分布

10 抽 样 分 布

我们对选取30名管理人员组成一个简单随机样本的过程不断重复，假设重复了500次，我们就会得到500个这样的简单随机样本所计算的平均值与标准差等数据。表中给出的是500个$\bar{x}$值的频数及频率分布。

500个简单随机样本的$\bar{x}$与p

Sample Number	Sample Mean ($\bar{x}$)	Sample Proportion ($\bar{p}$)
1	51,814	.63
2	52,670	.70
3	51,780	.67
4	51,588	.53
.	.	.
.	.	.
.	.	.
500	51,752	.50

500个简单随机样本的频数与相对频率

Mean Annual Salary ($)	Frequency	Relative Frequency
49,500.00–49,999.99	2	.004
50,000.00–50,499.99	16	.032
50,500.00–50,999.99	52	.104
51,000.00–51,499.99	101	.202
51,500.00–51,999.99	133	.266
52,000.00–52,499.99	110	.220
52,500.00–52,999.99	54	.108
53,000.00–53,499.99	26	.052
53,500.00–53,999.99	6	.012
Totals	500	1.000

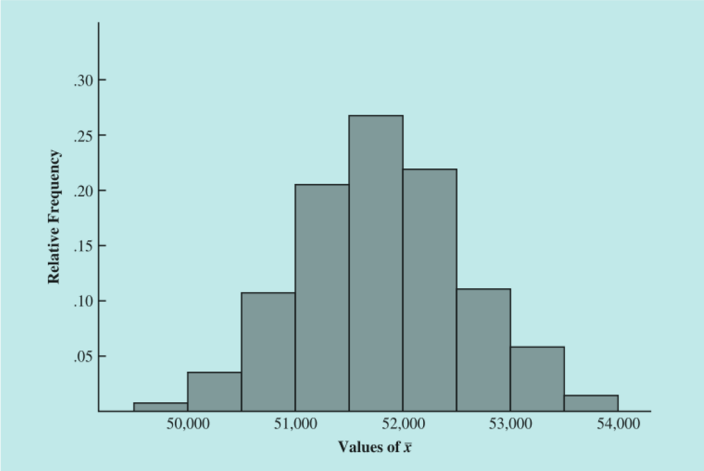
随 机 变 量

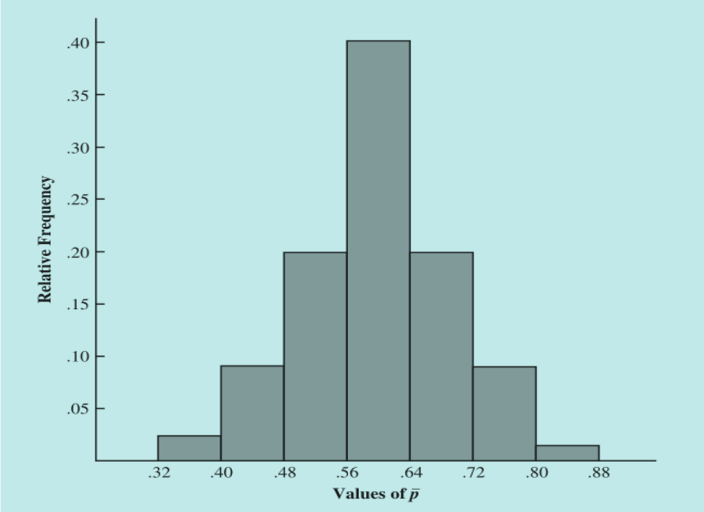
将抽取一个简单随机样本的过程看作一个试验，则样本的均值$\bar{x}$就是对试验结果的一个数值描述，从而样本均值$\bar{x}$是一个随机变量，它会拥有均值或数学期望、标准差和概率分布

抽 样 分 布

在不同的简单随机样本中，$\bar{x}$的取值也有各种可能的结果，我们称$\bar{x}$的概率分布为$\bar{x}$的抽样分布

11 抽 样 分 布





$\bar{x}$的直方图

分布形状近似是钟型的，500个$\bar{x}$的均值在总体均值51800美元附近。

p的直方图

统计量抽样分布

我们将抽样过程重复了500次，得到了不同的样本，计算出的统计量$\bar{x}$与p也不同。任何特定的样本统计量的概率分布称为该统计量的抽样分布.

12

$\bar{x}$的抽样分布

$\bar{x}$的抽样分布

$\bar{x}$的抽样分布是样本均值$\bar{x}$的所有可能值的概率分布

$\bar{x}$的数学期望

我们关心的是由大量简单随机样本产生的$\bar{x}$的所有可能的均值

$$E(\bar{x})=\mu$$

$E(\bar{x})$为$\bar{x}$的数学期望

μ为总体均值

$\bar{x}$的标准差

$$\sigma_{\bar{x}}=\frac{\sigma}{\sqrt{n}}$$

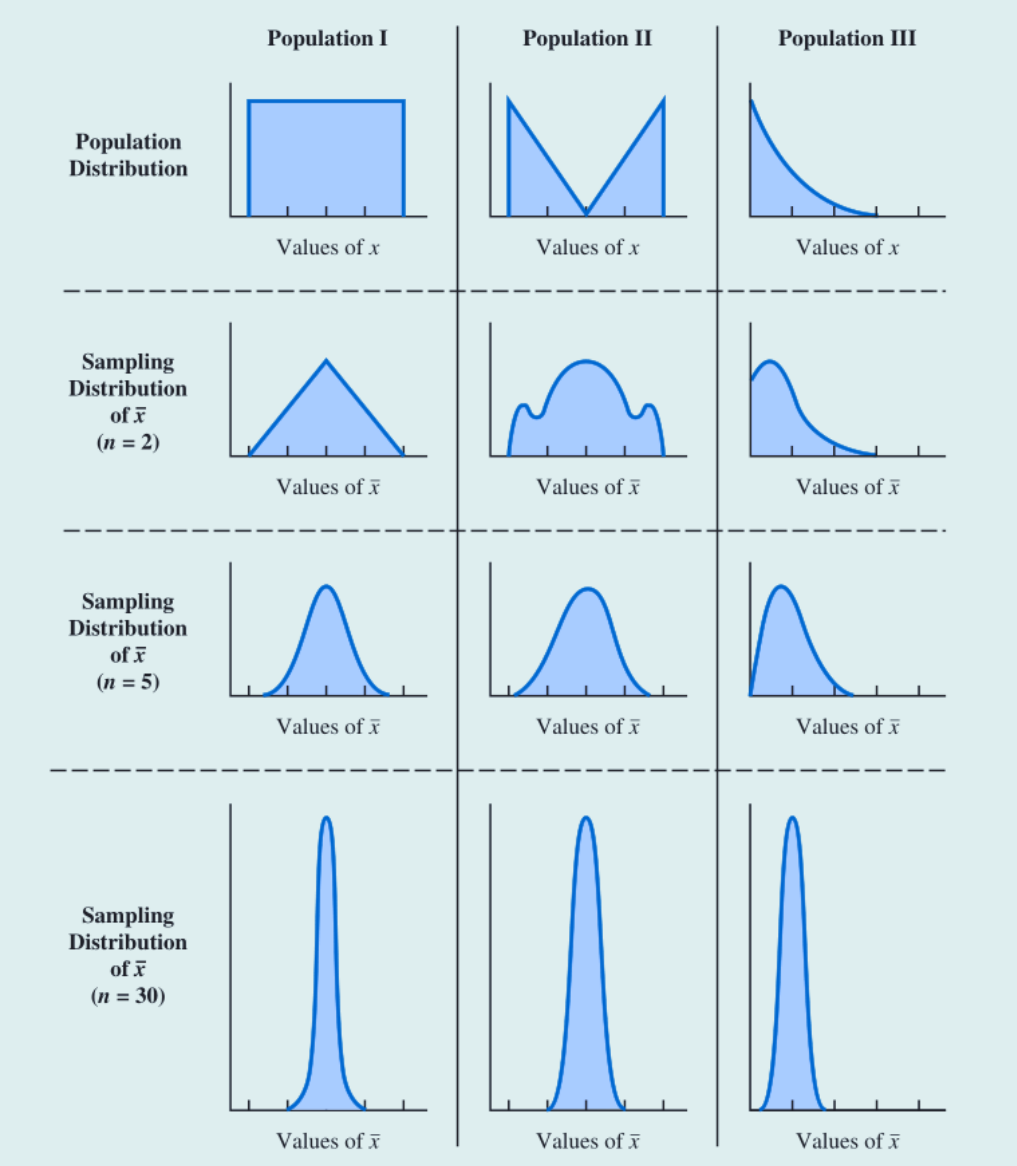
σ 总体标准差

n样本容量

标准误差

我们称$\bar{x}$的标准差$\sigma_{\bar{x}}$为均值的标准误差。一般地，标准误差指的是点估计量的标准差，它有助于确定样本均值与总体均值的偏离程度。

$\bar{x}$的抽样分布 形 式



总体服从正态分布

在任何样本容量下$\bar{x}$的抽样分布都是正态分布

总体不服从正态分布

中心极限定理

从总体中抽取容量为 n 的简单随机样本，当样本容量很大时，样本均值$\bar{x}$的抽样分布服从近似正态分布

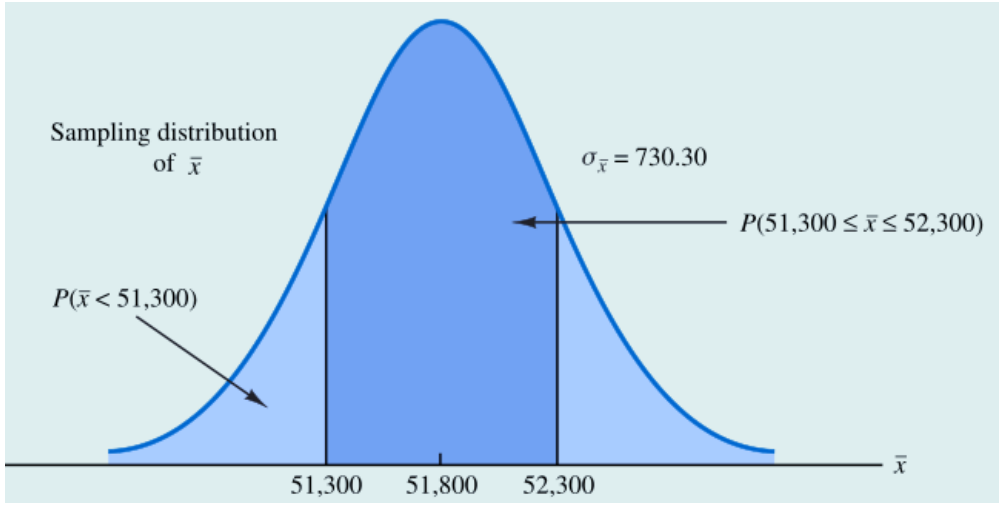
中 心 极 限 定 理

统计学家表明：在一般的统计实践中，假定当样本容量大于或等于30时， $\bar{x}$的抽样分布近似正态分布。当总体是严重偏态或者出现异常点时，可能需要样本容量达到50

$\bar{x}$的抽样分布 形 式

当抽取一个简单随机样本，用样本均值$\bar{x}$的值估计总体均值μ时，我们不能希望样本均值恰好与总体均值相等。我们对$\bar{x}$的抽样分布感兴趣的实际原因是，它可以用来提供样本均值与总体均值之间差异的概率信息

回到EIA问题，人事部经理关心的问题如下：根据30名管理人员组成的简单随机样本，得到的样本均值在总体均值附近±500美元以内的概率有多大？



概率计算

总体均值为51800美元，标准差为730.30美元

当$\bar{x}$ = 52300时：

$$z = \frac{52300 - 51800}{730.3} = 0.68$$

当$\bar{x}$ = 51300时：

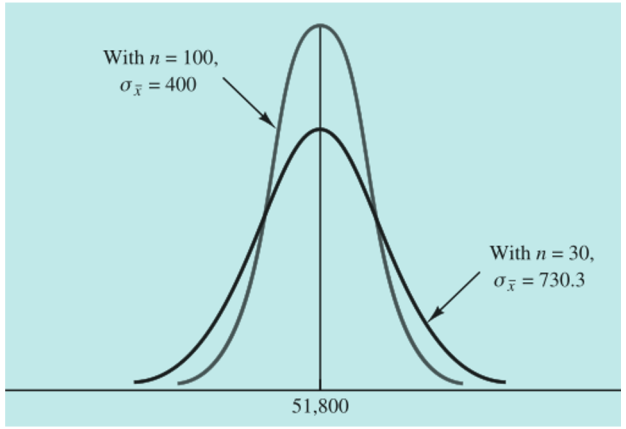
$$z = \frac{51300 - 51800}{730.3} = -0.68$$

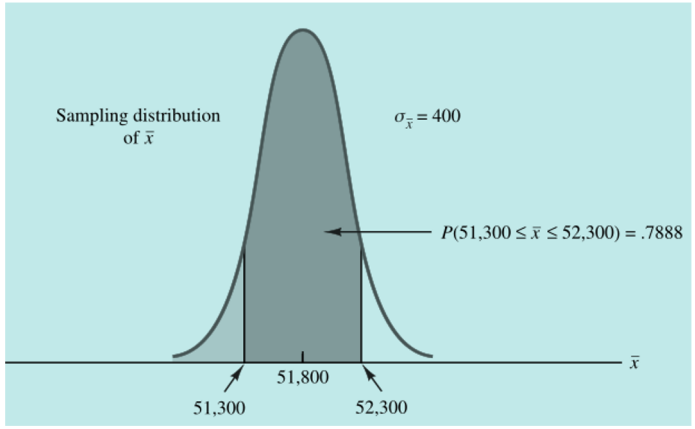
$$P(51300 \le \bar{x} \le 52300) = P(z \le 0.68) - P(z \le -0.68) = 0.7517\text{-}0.2483 = 0.5034$$

上述结果表明，由30名EAI管理人员组成的一个简单随机样本中，以0.5034的可靠性保证均值$\bar{x}$在总体均值$\pm$ 500美元以内。

样本容量与$\bar{x}$抽样分布的关系

我们增加样本容量到达100名，总体均值不发生变化，它不受到样本容量的影响。然而均值的标准误差$\sigma_{\bar{x}} = \sigma/\sqrt{n}$与样本容量的平方根有关。





概率计算

总体均值为51800美元，标准差为400美元

当$\bar{x}$ = 52300时：

$$z = \frac{52300 - 51800}{400} = 125$$

当$\bar{x}$ = 51300时：

$$z = \frac{51300 - 51800}{400} = -1.25$$

$$P(51300 \le \bar{x} \le 52300)$$

$$= P(z \le 1.25) - P(z \le -1.25) = 0.7888$$

上述结果表明，由30名EAI管理人员增加到100名，概率值从0.5034增加到0.7888

结论：样本容量越大，样本均值落在总体均值附近某一特定范围内的概率越大

04

抽样分布的性质

17

点估计的性质

在一个样本统计量作为点估计量之前，需要检查该统计量是否具有好的点估计量应具备的性质。好的点估计量应该具备以下三个性质：

无偏性

如果$E(\hat{\theta}) = \theta$,我们称样本统计量$\hat{\theta}$是总体参数θ的无偏估计量。$E(\hat{\theta})$代表样本统计量$\hat{\theta}$的数学期望

有效性

两个无偏点估计量$\hat{\theta}_1$和$\hat{\theta}_2$，我们倾向于采用标准误差较小的点估计量，称有较小标准误差的点估计量比其他点估计量更相对有效

一致性

随着样本容量的增大，点估计量的值与总体参数越来越近，则称该点估计量是一致的