第三讲

抽样

Sampling

抽样

- 抽样(sampling)是根据一定的规则和程序，从研究总体中抽取其中的一部分样本的过程。

- 普查（census）是对研究总体中的全部样本进行调查的过程。
- 为什么要抽样?

  无法全部进行调查

  调查具有破坏性

  调查的个体越多，非抽样误差越大

  有成本约束

  ……

抽样 vs. 普查

Type of Study	适用条件 Sample	Census
1. Budget 预算	Small	Large
2. Time available 可用时间	Short	Long
3. Population size 总体规模	Large	Small
4. 特征差异	Small	Large
5. 抽样误差的成本	Low	High
6. 非抽样误差的成本	High	Low
7. 测量的性质	Destructive	Nondestructive
8. 对个体案例的关注	Yes	No

抽样设计的过程

定义总体

↓

确定抽样框（ Sampling Frame）

↓

选择抽样技术（ Sampling Technique(s)）

↓

确定样本量（ Sample Size）

↓

执行抽样过程（ Sampling Process）

定义目标总体

目标总体（target population）是具有研究人员所要寻找的信息的个体或物体的集合。目标总体必须定义明确，目标总体根据如下四个方面来定义。

- An **element**（个体）是我们想要从中得到信息的物体，通常就是指调查对象。
- A **sampling unit** （抽样单位）有时指个体，有时指包含个体的单位，如家庭。
- **Extent**（范围） 指地理边界。
- **Time**（时间）指调研涉及的时点或时期。

百货商店顾客调查：目标总体

- 个体：家庭中负责在百货商店的大部分购物工作的男或女户主；
- 抽样单位：家庭
- 范围：上海市
- 时间：2006年

甄别问卷

- S1请问您是否是 上海 市居民吗？ 是，跳问S3，否，继续。
- S2请问您在 上海 市居住了多长时间了？2年以上，继续，2年不到， 终止访问。
- S3[出示卡片]请问您本人/家人/亲朋好友，有没有在以下所列的地方工作的呢？（单选）
- 市场调研公司或企业市场部
- 广告公司
- 公关公司
- 媒体领域
- 食品、饮料、啤酒生产商，零售商或分销商
- 酒吧/餐厅/饭店
- 大卖场/超市/百货商场
- 非以上机构或公司
- S4请问在过去六个月内，您有没有参加过任何类型的市场研究或访问呢？ 是，终止访问。

抽样设计的过程

定义总体

↓

确定抽样框（ Sampling Frame）

↓

选择抽样技术（ Sampling Technique(s)）

↓

确定样本量（ Sample Size）

↓

执行抽样过程（ Sampling Process）

确定抽样框架

- 抽样框架（sampling frame）指的是目标总体中的个体列表(list)，它由一份或一组用于识别目标总体的指示说明组成。
- 抽样框架和目标总体应最大限度吻合

- 理论上应该是目标总体决定抽样框架，现实中往往是抽样框架决定目标总体。

请定义目标总体和确定抽样框架

- 一家地方电视台想了解家庭的收看习惯和对节目的偏好
- 美国营销协会的地方分会想在亚特兰大测试新的吸收会员活动的有效性
- 针对上财2000年入学的学生的截面研究。

抽样设计的过程

定义总体

↓

确定抽样框（ Sampling Frame）

↓

选择抽样技术（ Sampling Technique(s)）

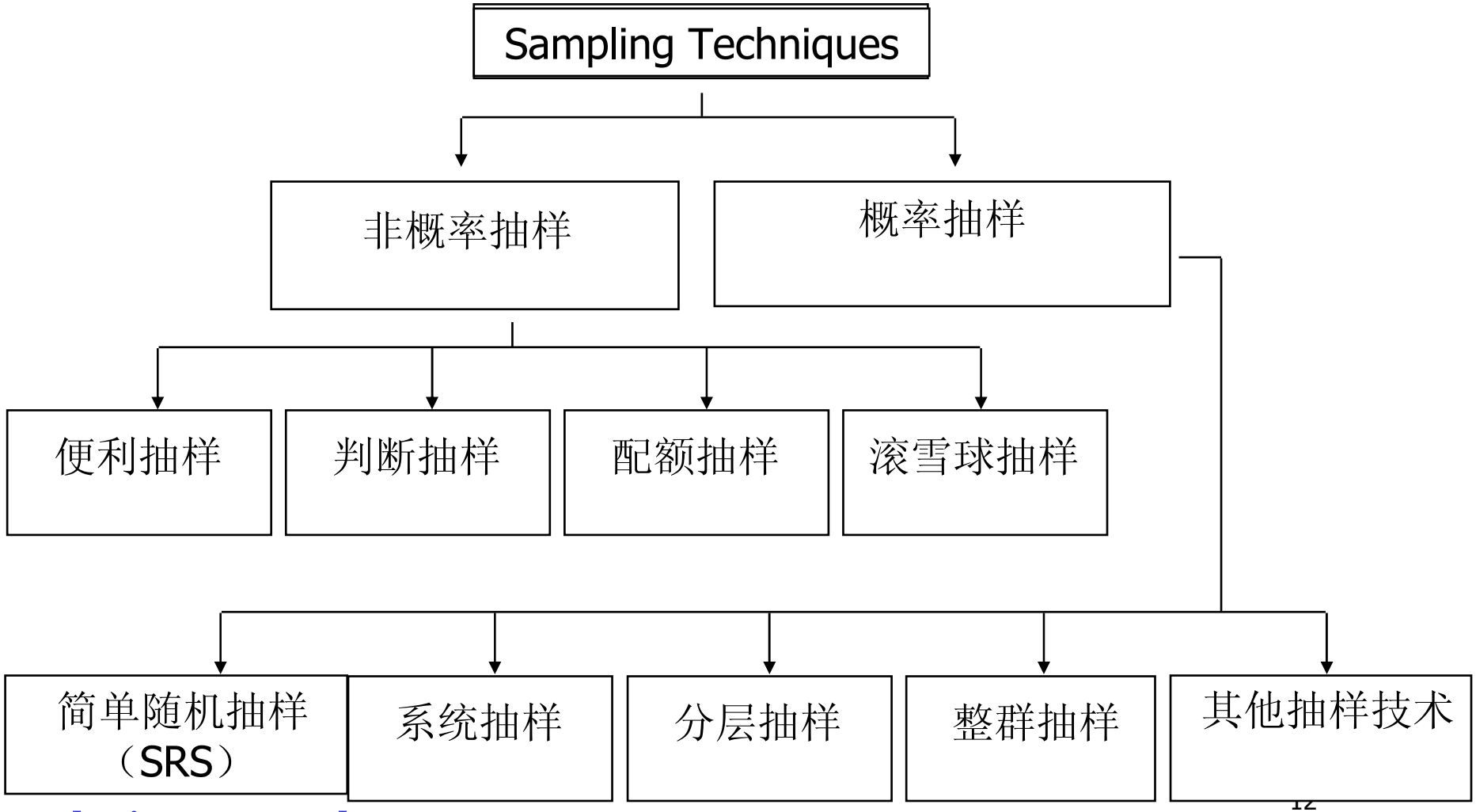
↓

确定样本量（ Sample Size）

↓

执行抽样过程（ Sampling Process）

抽样技术的分类


Sampling Techniques
非概率抽样
概率抽样
便利抽样
判断抽样
配额抽样
滚雪球抽样
简单随机抽样（SRS）
系统抽样
分层抽样
整群抽样
其他抽样技术

便利抽样（Convenience sampling）

便利抽样通过获取一些便利的个体得到样本，被访者由于碰巧在恰当的时间和恰当的地点出现而被选中。

- 使用学生、社会组织的成员等
- 商场拦截访谈
- 从杂志上撕下的问卷
- “街上的行人”访谈

便利抽样优缺点

Advantages:

- Least expensive
- Least time consuming
- Sampling units are accessible, cooperative, and easy to measure

Limitations:

- Many potential sources of bias are present
- Not representative of any population
- Not able to generalize findings

判断抽样（Judgmental sampling ）

判断抽样也是一种便利抽样，它根据研究人员的判断选出(典型）个体。

- test markets selected to determine the potential of a new product
- purchase engineers selected in industrial marketing research
- Selection of department stores to test a new merchandising display system

判断抽样优缺点

Advantages:

- Low cost
- Less time consuming than most techniques
- Convenient to use

Limitations:

- No direct generalizations to a specified population
- Results entirely dependent on the judgment of the researcher

配额抽样（Quota sampling ）

配额抽样可看作两阶段有约束的判断抽样。

- 第一阶段确定总体中个体的控制类别或者配额组成。
- 第二阶段，个体在便利或判断的基础上被选出来。

Control Characteristic	Population composition Percentage	Sample composition Percentage	Number
Sex			
Male	48	48	480
Female	52	52	520
	100	100	1000

配额抽样（例子）

省份	IT/视听	灯具	工具	家电	零部件	总计
江苏	17	20	39	20	27	123
山东	2	1	17	6	4	30
上海	9	13	16	7	13	58
四川	1				1	2
厦门	8	9	1	2	3	23
浙江	20	49	51	88	57	265
总计	57	92	124	123	105	501

配额抽样优缺点

Advantages:

- Lower cost in selecting elements for each quota
- Greater convenience in selecting elements for each quota
- With tight controls on interviewers and interviewing, results can be compared to probability techniques

Limitations:

- No assurance that a sample is representative
- Increasing controls decreases the ease of conducting the interview
- Many sources of bias can be present due to the selection process

滚雪球抽样（snowball sampling）

在滚雪球抽样中，先随机选出一组最初的调查对象，

- 在访谈之后，要求这些被访者推荐一些属于目标总体的其他人。
- 根据这些推荐选出后面的被访者。

- 适合访问符合特征的个体少且分散的总体。

滚雪球抽样优缺点

Advantages:

- Good at locating and estimating various characteristics that are rare in a population
- Low sample variance and costs when used

Limitations:

- Difficulty in obtaining the initial sample
- Limited to studies of a particular sample

Procedures for Drawing Probability Samples

Simple Random Sampling

1. 选择一个适当的抽样框架
2. 给每个个体分配一个1到 N(pop. size)之间的数字
3. 生成 n (sample size) 个不同的介于1到 N的随机数字
4. 所生成的数字代表了应该被包括进样本中的个体

简单随机抽样（SRS）

- 总体中每个个体都有一个已知且相等的抽中概率。

  若总体为N，样本量为n，则每一抽样单位的抽中概率：

  $$p = n/N$$

- 每个规定了规模(n)的样本都有一个已知且相等的被选为实际样本的概率。

简单随机抽样的优缺点

Advantages:

- 易于理解
- 样本结果可映射到目标总体上

Limitations:

- 很难构建一个抽样框架
- 样本大或跨越地理区域广，增加了数据收集的成本
- 比其他概率抽样技术的精确度低
- 若样本量很小，很可能没什么代表性

Procedures for D
Probability Sar

Systematic Sampling

1. 选择一个适当的抽样框架

2. 给每个个体分配一个1到 N(pop. size)之间的数字

3. 确定抽样间距 i:i=N/n. 如果 i 是个分数，将其四舍五入到最接近的整数

4.在1到 i之间选择一个随机数字r

5. 带下面数字的个体将组成系统随机样本: r, r+i,r+2i,r+3i,r+4i,...,r+(n-1)i

系统抽样(systematic sampling)

- 把总体中的个体进行排序，通过选择一个随机的起点，然后从抽样框架中连续地每隔i个个体挑出一个个体构成样本。
- 抽样间距（sampling interval）i＝N/n，并四舍五入。
- 如果个体排列的顺序与所研究的特征直接相关，系统抽样可增加样本的代表性。
- 如果个体排列的顺序呈现循环形式，则系统抽样会降低样本的代表性。

- 与SRS不同之处：只有可以被抽出的样本量为n的可容许样本具有已知且相等的选择概率，其余样本量为n的样本被选择的概率为0。

系统抽样的优缺点

Advantages:

- 当个体排序与研究特征有关时，系统抽样比SRS更有代表性
- 只需做一次随机抽样，成本更低
- 在不了解抽样框架的情况下也可以使用

Limitations:

- 个体不适当的排序会降低样本的代表性

Procedures for D
Probability Sar

Stratified Sampling

1. 选择一个适当的抽样框架
2. 选择分层变量和层数 H
3. 将总体分为 H 层，根据分类变量，总体中的每个个体被分配到H层中的一层
4. 在每层中，将个体从 1 到 N_h (第h层的总体规模）编号
5. 根据按比例分层抽样或非比例分层抽样确定每层的样本量 n_h,

$$\sum_{h=1}^{H} n_h = n$$

6. 在每一层中，选择一个样本为n_h的简单随机样本。

分层抽样(stratified sampling)

- 分层抽样是一个两阶段过程，总体被分割为子总体（subpopulations）或称"层（strata）"。
- 每个个体都必须分配且只能分配到某一层，不应有遗漏。
- 用随机方法，通常是SRS，从每一层中选出个体。
- 分层抽样的主要目的是在不增加成本的同时增加精确度。

分层抽样(stratified sampling)

- 同一层内的个体应尽可能同质（homogeneous），不同层的个体应尽可能异质（heterogeneous）。
- 分层变量（stratification variables）应该与需要研究的特征紧密相关。
- 分层变量应该易于测量和应用，从而降低分层过程的成本。
- 在按比例分层抽样（proportionate stratified sampling）中，从每层中抽出的样本量与该层在整个总体中的相对量成比例。
- 在非比例分层抽样（disproportionate stratified sampling）中，从每层中抽出的样本量与该层在整个总体中的相对量成比例，且与该层中所有个体特征分布的标准差成比例。

成正比例分层抽样法

Consumer type	Group size	10 Percent directly proportional stratified sample size
Brand-loyal	400	40
Variety-seeking	200	20
Total	600	60

成反比例分层抽样法

Inversely Proportional Stratified Sampling

分母	→	600/200 + 600/400 = 3 + 1.5 = 4.5
Heavy Drinkers proportional and sample size	→	3/ 4.5 = 0.667; 0.667 * 60 = 40
Light drinkers proportional and sample size	→	1.5 / 4.5 = 0.333; 0.333 * 60 = 20

成反比例分层抽样法

Consumer type	Group size	10 Percent inversely proportional stratified sample size
Heavy drinker	200	40
Light drinker	400	20
Total	600	60

分层抽样的优缺点

Advantages:

- 比SRS的精确度高
- 确保所有重要的子总体在样本中得到体现

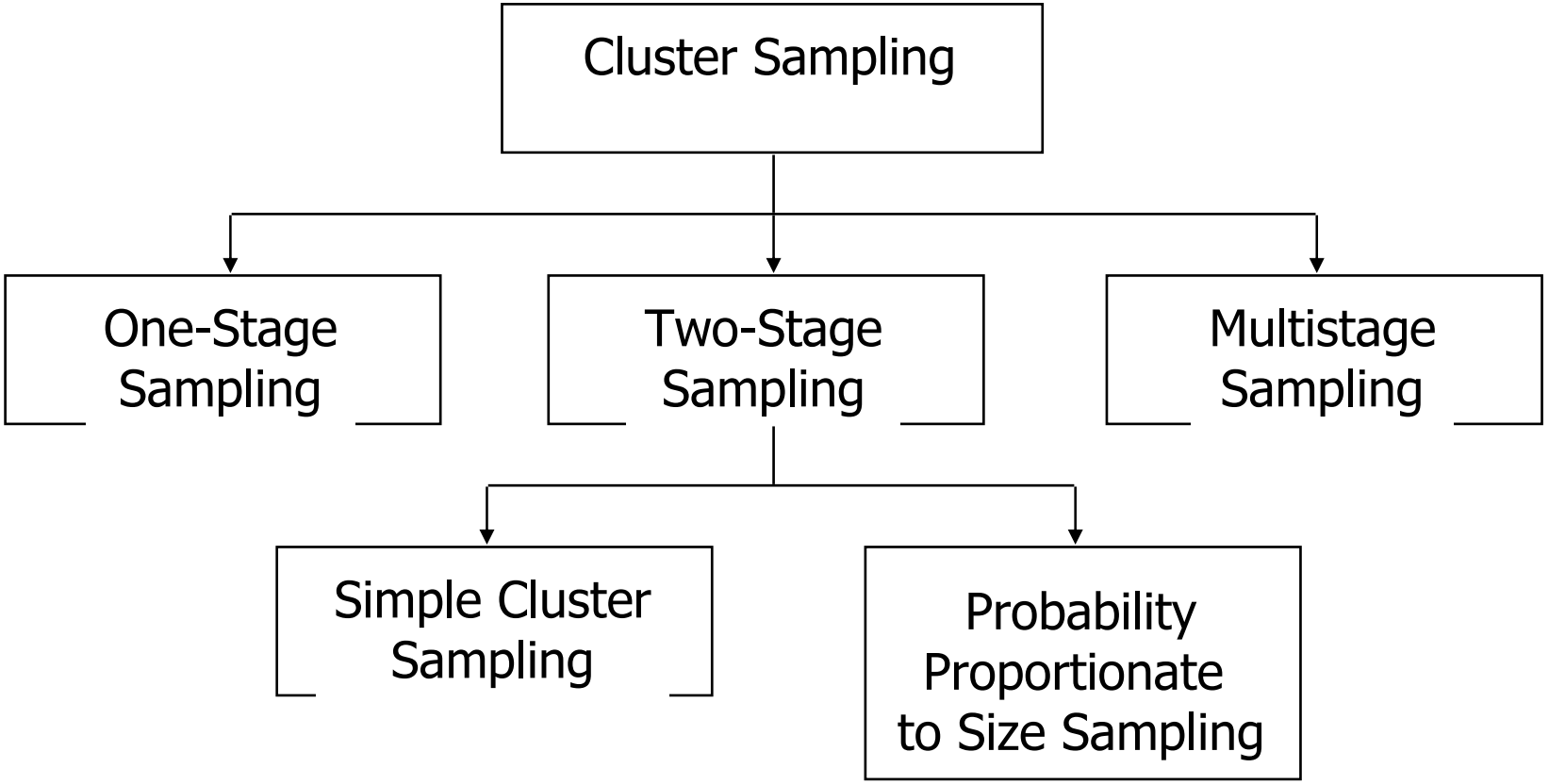
Limitations:

- 采用非比例分层抽样需要估计不同层的标准差

整群抽样（cluster sampling)

- 首先将目标总体分为相互排斥且没有遗漏的子总体或称群（**clusters**）。
- 然后用概率抽样技术，如 SRS选出由某些群构成的一个随机样本。
- 对于每个被选中的群，如果所有个体被包括进样本，称为单阶段整群抽样， 如果按概率抽样技术抽出一个个体的样本，称为两阶段整群抽样。
- 一个群内的个体应该尽可能地异质，但是各群本身应该尽可能地同质。
- 在与规模成比例的概率抽样（ **probability proportionate to size sampling**）中，大的群比小的群更有可能被抽中，在第二阶段，一个抽中群中个体被选中的概率与该群的大小呈反向变化。

整群抽样的类型


Cluster Sampling
One-Stage Sampling
Two-Stage Sampling
Multistage Sampling
Simple Cluster Sampling
Probability Proportionate to Size Sampling

整群抽样的优缺点

Advantages:

- 成本低，因为只有被抽中的群才需要抽样框架
-

Limitations:

- 样本不够精确
- 一般很难构造异质的群
- 对基于群的统计数字进行计算和解释 也可能很困难。

比 较

分层抽样	整群抽样
群内同质	群间同质
群间异质	群内异质
包括所有群	只包含随机选出的群
通过使准确度提高的速度快于成本提高的速度来改善抽样效率	通过使成本降低的速度快于准确度降低的速度来改善抽样效率

Choosing Nonprobability vs. Probability Sampling

考虑因素	Conditions Favoring the Use of Nonprobability sampling	Conditions Favoring the Use of Probability sampling
研究的性质	探索性	归纳性
总体的变异程度	低（同质）	高（异质）
统计上的考虑	不利	有利
操作上的考虑	有利	不利

抽样:确定最终与最初样本量

定义与概念

- **Parameter**: 参数是指对目标总体的一个固定特征或测量值的概括性描述。

- **Statistic**:统计量是指对样本的一个特征或测量值的概括性描述。样本统计量一般用做总体参数的估计值。

- **Finite Population Correction**: 有限总体校正（fpc）指当样本量占总体量的10%以上时，对均值或比例等总体参数方差的过高估计值进行的校正。

定义与概念

- **Precision level**: **精确度**是样本统计量与总体参数之间的最大允许差异。

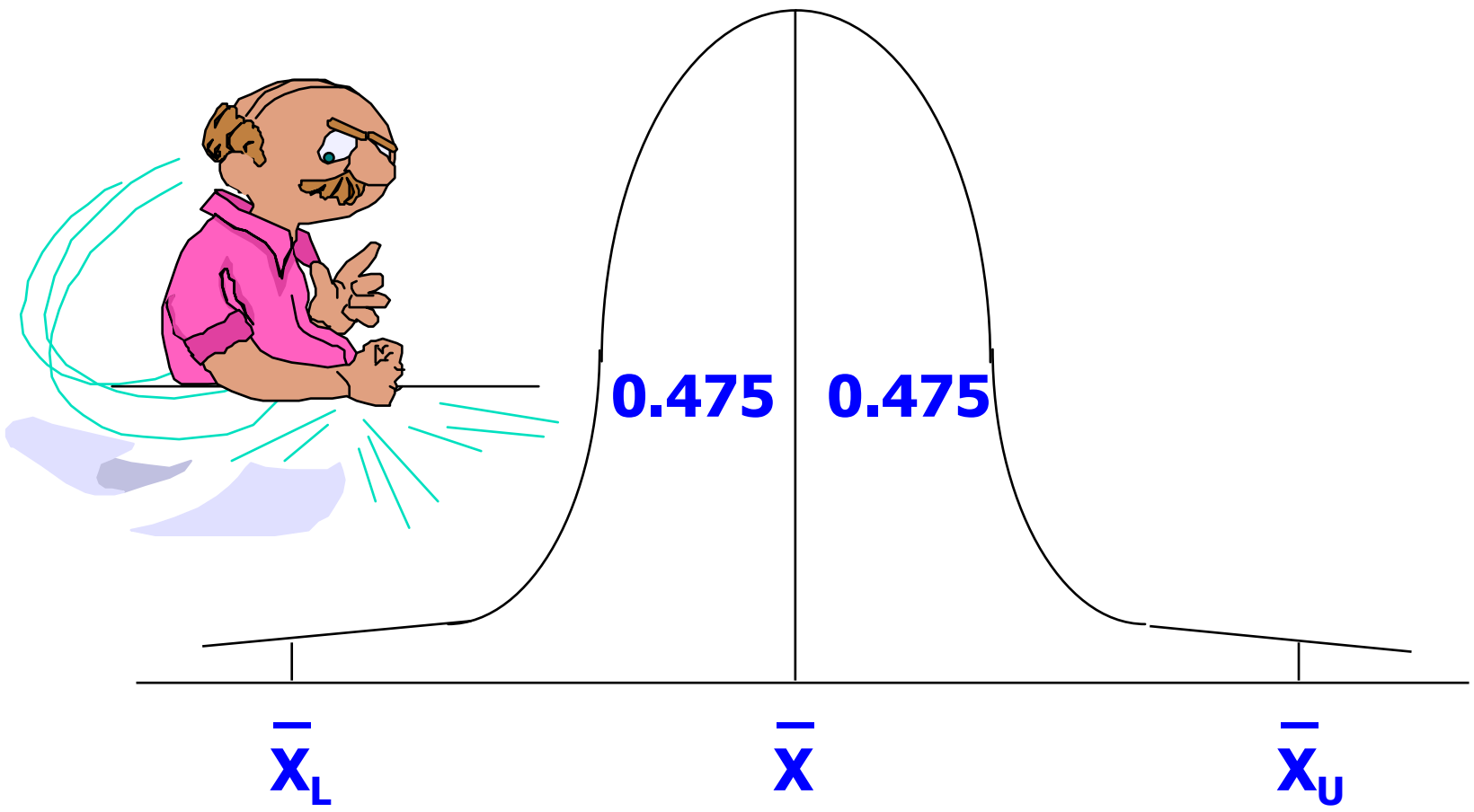
- **Confidence interval**: **置信区间**是指假定一个给定的置信水平，真实的总体参数将会落入的范围。

- **Confidence level**: **置信水平**是指一个置信区间将包括总体参数的概率。

总体和样本变量的符号

Variable	Population	Sample
均值	μ	$\bar{X}$
比例	Π	p
方差	σ^2	s^2
标准差	σ	s
规模	N	n
均值的标准误	$\sigma_{\bar{x}}$	$S_{\bar{x}}$
比例的标准误	σ_p	S_p
标准化变量 (z)	$(X-\mu)/\sigma$	$(X-\bar{X})/S$
变异系数 (C)	σ/μ	$S/\bar{X}$

95% Confidence Interval



用统计学方法确定样本量

步骤	均值	比例
1. 指定精确度	D = ±$5.00	$D = p - \Pi = \pm 0.05$
2. 指定置信水平 (CL)	CL = 95%	CL = 95%
3. 确定与 CL 相对应的 Z 值	z 值为 1.96	z 值为 1.96
4. 确定总体的标准差	估计 σ: σ = 55	估计 Π: Π = 0.64
5. 用标准误公式来确定样本量	$n = \sigma^2 z^2/D^2 = 465$	$n = \Pi(1-\Pi)\, z^2/D^2 = 355$
6.如果样本量相当于总体的 10%，应用有限总体校正（FPC）	$n_c = nN/(N+n-1)$	$n_c = nN/(N+n-1)$
7. 如果必要，通过用 s 对 σ做估计来重新估计置信区间	$= \bar{X} \pm z s_{\bar{X}}$	$= p \pm z s_p$
8.如果指定的精确度是相对值而非绝对值，则使用这些公式，式中 C 是变异系数.	$D = R\mu$ $n = C^2 z^2/R^2$	$D = R\Pi$ $n = z^2(1-\Pi)/(R^2\Pi)$

M

例题

- 一家当地餐馆的管理人员想确定家庭在价格高昂的餐馆的平均月消费数目。有些家庭根本不消费，而另一些家庭的花销高达每月300美元。管理人员想要95%置信度的结果，并且不希望误差超过正负5美元。
- （1）应该用多少样本量来确定平均月家庭支出？
- （2）执行调查之后，发现平均支出为90.30美元，标准差为45美元。构建一个95%的置信区间。

- （1）估计σ = 300/6 = 50.（6 Sigma姆指法则）
- z = 1.96
- D = 5
- 所以, n = 384.16，约等于 385

- （2） $90.30 + 1.96(45/\sqrt{385}) = 94.81$
- $90.30 - 1.96(45/\sqrt{385}) = 85.79$

调整统计学方法确定的样本量

发生率（Incidence rate） 指符合参加研究条件的人的百分率。

In general, if there are c qualifying factors with an incidence of Q_1, Q_2, Q_3, ...Q_C,each expressed as a proportion,

Incidence rate $= Q_1 \times Q_2 \times Q_3 \ldots \times Q_C$

完成率（completion rate) 指完成访谈的合格对象的百分率。

$$\text{Initial sample size} = \frac{\text{Final sample size}}{\text{Incidence rate} \times \text{Completion rate}}.$$

确定样本容量的定性因素

- 决策的重要性（the importance of the decision）
- 研究的性质（the nature of the research）
- 变量的数目（the number of variables）
- 分析的性质（the nature of the analysis）
- 类似研究的样本量（sample sizes used in similar studies）
- 发生率（incidence rates）
- 完成率（completion rates）
- 资源约束（resource constraints）

市场调研中通常的样本量

研究类型	Minimum Size	Typical Range
问题识别研究（如市场潜力）	500	1,000-2,500
问题解决研究（如定价）	200	300-500
产品测试	200	300-500
试销研究	200	300-500
电视/广播/印刷品广告 （每个广告）	150	200-300
试销市场审计	10 stores	10-20 stores
专题组座谈	6 groups	10-15 groups