CDA 统计课程

==========

1.数据准备

1.1 数据描述

***何为变量?**

***分类变量与连续变量**

***测量级别问题**

1.2 数据分布与阶矩

***实验室的实验设计**

***问卷的理论框架**

***数据库的影响因素选择**

1.3 抽样技术准备

***如何确定样本量**

***抽样技术理解与实操**

1.1 数据描述

（1）何为变量?

（2）分类变量与连续变量

（3）测量级别问题

（1）何为变量?

基于以上对随机变量的描述，从变量取值多少、小数位意义、样本量大小、理论定义几个方面描述信息量，而信息量又可以进一步分成定量型和定性型，如下图随机变量的测量类型，定量变量又可以分成连续型和离散型变量；定性变量分成定序和名义；离散型变量包括计数型资料，并一同定序和名义统称为分类型变量，所以有时候我们在描述变量时，把连续和分类视为一对概念，将连续和离散放在一起视为一对、定量和定性放成一起。变量的测量级别是将名义、定序、定距和定比放在一起描述变量。

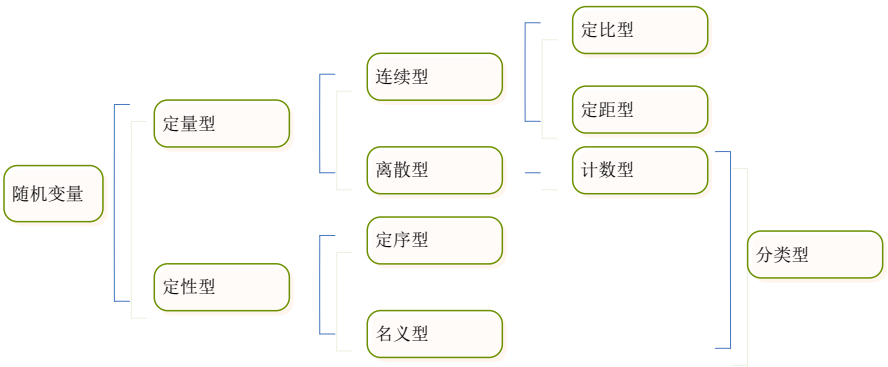


图 随机变量的测量类型[1]

（2）分类变量与连续变量

分类体和连续体之间的界限是什么？下面提出四个疑问，也许有助于读者更好理解其区别。

第一，变量的取值多不多？

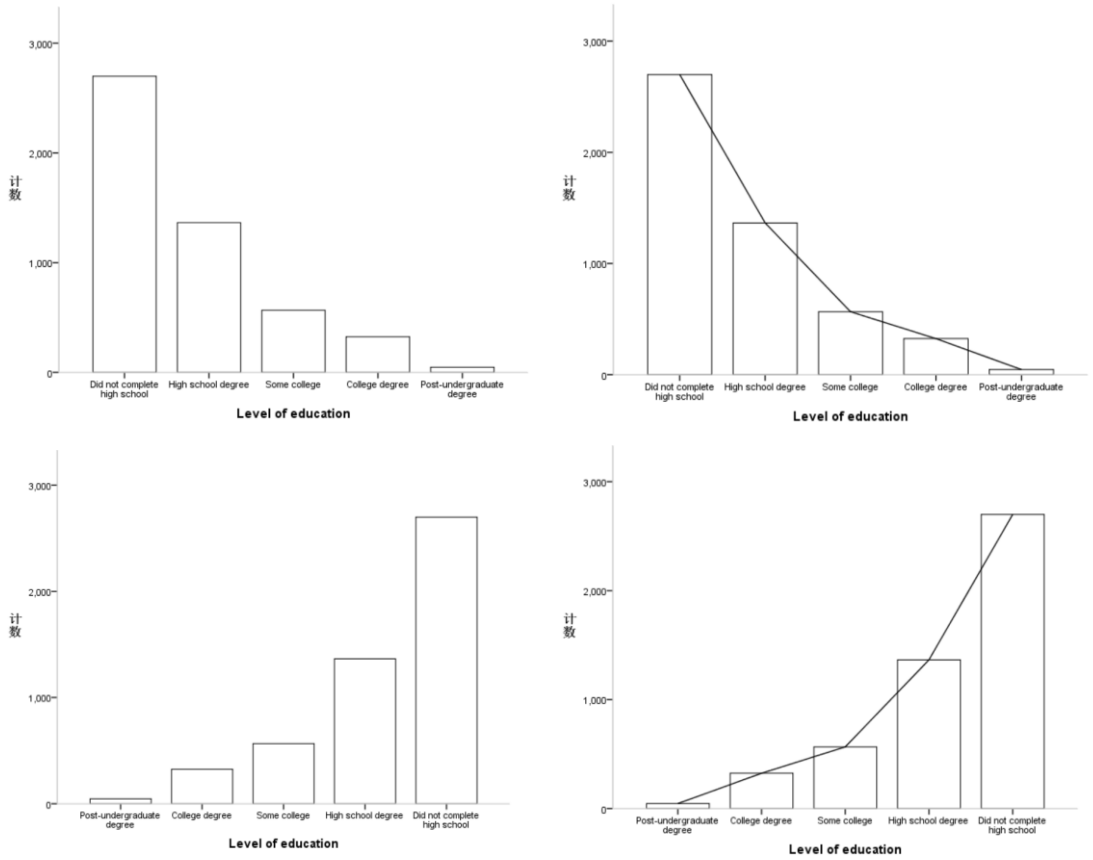
第二，带有小数位的取值有否具有实际意义？

第三，总样本量是多少？

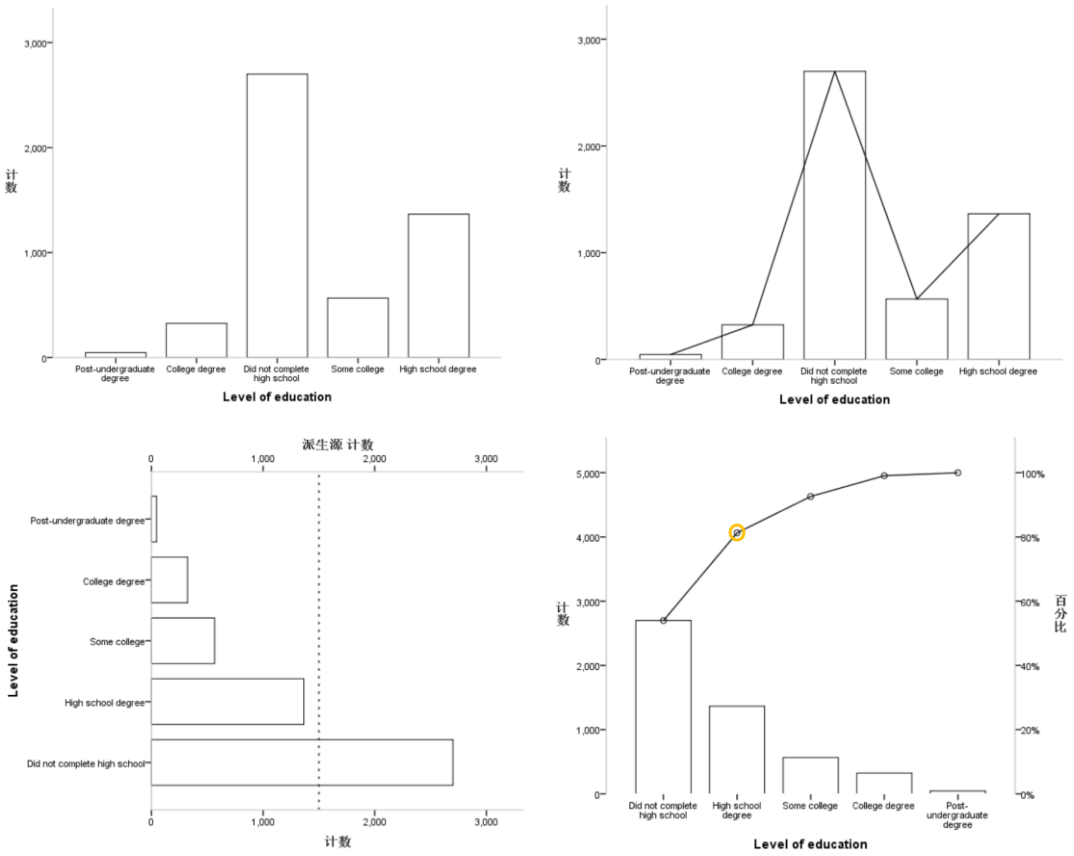
第四，理论或业务是如何看待变量的？

分类体的频数及比例问题。

条形图通常称为非专业化图形，主要是指没有统计专业基础的人也能够看得懂的图形，以下图所示，我们绘制了离散变量的几种不同的拓展图形，看一下每种图形所表达的重点在哪里？条形图其实本质上强调占比，但我们知道，占比通常使用饼图会更合适，所以条形图传达的衍生含义主要包括：趋势性、波动性和累计。

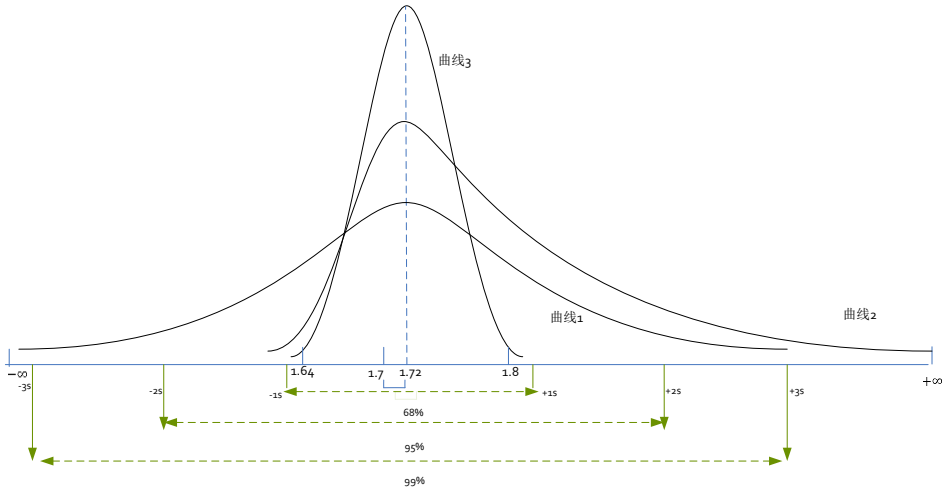


[1] 见《回归分析》的随机变量章节。



连续变量及描述

下面两幅示意图中，其中下图小型数据常见的三种分布和大型数据常见的三种分布。

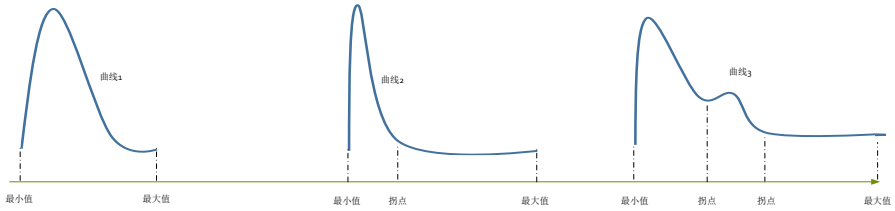


小型数据常见的三种分布

数据分析描述涉及到了阶矩问题 ，在数据体量与信息分布中，将不同统计量的指标统一用阶矩表达，如均值右上角可以看成是一次方、方差的右上角看成是二次方、偏度是三次方、峰度是四次方，这就是数据分布的阶矩特征。尽管我们上文一再强调数据分析需要区分不同领域的数据及统计方法，但特例是，有些模型具有跨界性，即同一模型在不同领域中都有使用，以方差分析为例，在实验室中方差分析是“主角”、在问卷中方差分析是“配角”、在数据库中方差分析最多是个“跑龙套”的。此意作何解？我们使用的是同一个方差分析，但是用的重点不同，一般而言，在精确的数据中，考虑的阶矩问题倾向于更高，例如在实验室中，

要考虑到 1 到 4 个阶矩带来的所有误差问题、在问卷中，通常是考虑 1 到 2 个阶矩，偶尔也会考虑 3 阶矩问题、在数据挖掘中，一般考虑一阶矩就足够了，只在很少的情况下会考虑到 2 和 3 的阶矩问题。

基于以上特征我们常见于右偏分布，如果有偏性不是特别严重的话，如曲线 1，仍然可以使用常务性指标，如中位数、四分位距、异常、最大值与最小值；曲线 2 增加了拐点，拐点的意义在于区分左侧与右侧不同的用户群体特征，比如左侧的用户群体与普通的右偏分布大体相当，因此仍然可以使用传统的常务性指标描述大众与小众；但右侧的群体特征并不区分大众与小众，一般重点描述该群体的全距及等宽度区间对应的概率在业务上的意义。因此需要回答，此处为什么会有拐点？拐点前后群体特征是什么？群体上下限的意义？此外，曲线 3 增加了多处拐点，具体需要描述的问题与曲线 2 类似。



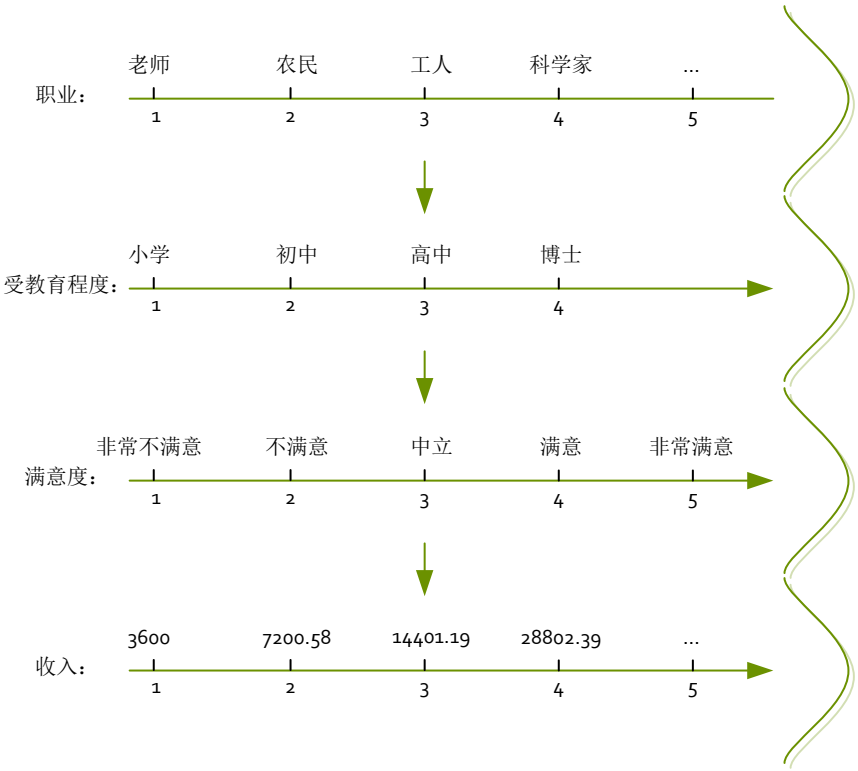
大型数据常见的三种分布

（3）测量级别问题

在测量学的尺度上，对于数值本身的关注，构成了测量级别的含义，关注的重点——类别、顺序、间距与幅度。类别强调分类体间的离散关系，即名义测量；顺序强调数值秩的等距关系，即有序测量；间距强调数值大小的线性关系，即间距测量；幅度强调数值大小的非线性关系，即比率测量。

名义测量的表述。

如果现在手中有一把尺子，尺子上的刻度分别表示变量取值，如下图所示，不同变量对应的刻度与分布，除职业外其他刻度都具有大小顺序之分。职业及其刻度仅仅用于区分离散体的不同取值，强调类别，无关乎大小和顺序，如我们不能说农民大于老师、科学家大于工人、也不能说农民小于工人，这在真实场景中没有意义。与此相似的指标还包括宗教信仰、民族、婚姻状态、手机颜色、购物渠道等变量。



尺度与测量级别

有序测量的表述。

有序变量提供的信息大于名义变量，因为有序变量除了具有类别的功能外，又增加了类别间的顺序，例如我们可以说初中学习到的科学文化知识是多于小学的、也可以说博士学历大于高中、初中小于博士等，因此就序的问题上，分析数值具有极大的便利性，但问题也因此而产生，当我们关注数值差异时，却带来了识别上的“混乱”，如 1、2、3…分别表示不同学历，既然具有序的大小性，若 4（博士）减 3（高中）为 1、3（高中）减 2（初中）也为 1，这两个 1 是否可以划等号呢？显然，就数值而言，两者是一样的，应该划等号，但若考虑实际情况，博士和高中之间相差 10 年左右，但高中和初中则相差 3 年，所以两个 1 代表的年数不同，这样考虑就不能够划等号。我们说这两种思考都是有道理的，如果认为 1 等于 1，这其实就是强调类别间的等距性，这恰恰就是序的含义，若反之，不等距关注的点其实并不是类别问题，而是间距问题，比如如果强调具体 3 年与 10 年之别，数值具体的大小对应于刻度上的间距不同，即间距长短，如果使用尺子度量的话，往往就有小数位性质，所以我们说这是离散变量向连续变量过滤的标志。

间距测量与比率测量的表述

以变量幸福感和收入为例，见表 2-3，通过测评我们获得如下数据。

表 2-3 用户评估得分

用户	收入（0-10 万/月）	幸福感（0-10 分）
001	10	10
002	6.8	6.8
003	0.5	0.5
004	0	0

量化形式：是否形象化		?

变量测量的形式多样，而这种形式能不能准确传达真实环境中，人们所表达的信息内容，如一个苹果需要 1.2 元，如果对方说需要 1.2267890 元钱，显然过于啰嗦，但却很精确，不过我们说这不是测量技术追求的 “刚刚好”的状态，因为统计，并不是所有的模型都要求精确，必要的时候精确性是次要的，那这个“刚刚好”如何在不同性质的数据中准确表达呢？尤其是连续型变量对这一问题提出了挑战。

话说考虑有序型变量的间距长短，使离散体向连续体过渡，也许正因为关注间隔距离的长短，所以将其称之为间距变量，这是连续型变量的第一种形式，这种形式具备一系列的特征，刚好我们与此相似的比率变量作为对照——抽象性、零和倍的表达。

==========

1.2 数据分布与阶矩

***实验室的实验设计**

***问卷的理论框架**

***数据库的影响因素选择**

（1）实验室

实验研究中的实验室是一个广义的概念，不仅仅指的是有一个房子或场地，它主要用于表示实验中的控制思想，尤其是随机噪声和影响因素的控制。在这种环境下，很容易获取高精度的数据，但是实验室效应的一系列毛病，也会拉开与现实间的差距，从而限制解决更广泛的社会经济生活问题。那么直接放宽条件，增加更多的实验因素（这里特指自变量）可不可以？当然理论上说是可行的，但经验来看，实验室研究中的自变量通常不建议超出三个或四个以上，如果条件放宽增加更多的实验因素，就意味着，在保持高精度的数据质量前提下，大量实验因素的参与，实验误差大幅增加，高维复杂的实验如何设计？与此相应的统计模型应该长什么样？或许统计学家还没做好这个准备，或许统计学家不愿做好这个准备，或许这个准备的复杂度远远高于范式转移后的复杂度——问卷。

（2）问卷

广义上的问卷包括量表和狭义上的问卷两种形式，这里是指广义上的问卷。一般来说，从问卷结构上看，一个常见的问卷包括：一级概念、二级维度、三级指标。问卷结构主要特征，不在于概念维度或指标数量的多少，而在于这种结构是否源于理论。实践证明社会科学的理论维度通常可以在三到九个之间，这就大大拓展了现实当中使用维度的数量，也拓展了实验室研究的局限。那么这个领域中的数据又呈现出什么特点呢？

相较而言，实验室数据的核心信息分布在微观层面，问卷数据则分布在宏观层面。但是如果是站在今天，回顾一百多年来统计分析的数据质量的话，问卷的数据其实核心信息仍分布在微观层面，为什么这么说，因为在八九十年代随着数据量的突增，统计分析的重点又开始从问卷领域转移到数据库领域。

（3）数据库

数据量的增加，几乎必然会带来数据质量的下降。大型数据行的存储问题与数据列间的复杂性——对存储效率问题和列间结构性问题的探讨，进而产生数据库，但由于技术、人力以及资金的限制等原因，使得数据存储的质量，实际上是在下降的。所以大型数据被称之为数据挖掘，其言外有意——能够从大型的数据中抽取一丁点有价值的商业信息，即为挖掘。

==========

1.3 抽样技术准备

（1）如何确定样本量

单样本情况

一般在调查研究中，调查通常有两种情况：预调查和正式调查。预调查通常可以使用20—40 人，而在正式调查中，样本量确定会显得有些专业。一般需要考虑的影响结果的因素包括：抽样样本的抽样误差、是否需要估计总体、问卷信效度、受访者作答过程的质量控制、后期需要使用的方法、关键变量的方差变异、时间进度、项目经费等。因此样本量的多少不能一言一概之。

Cochran 于 1977 年提出变量百分比的样本量的近似公式（参见《抽样调查设计导论》）：

$$n = \left(1 - \frac{n}{N}\right) \cdot \frac{t^2(p \cdot q)}{d^2}$$ = 有限总体纠正因子*概率水平*方差/置信区间

注：需要考虑应答率，将 $n/应答率$=实际调查人数。

其中，

有限总体纠正因子：一般当样本量小于总体的 5%时，有限总体纠正因子的影响非常微弱，可能也只有几个样本的差异（其他既定的情况下），对应抽样调查而言，往往可以忽略不计，如总体为 2000 个，样本量为 100 个，1-0.05=0.95，0.95 对后面公式的修正很有限。

概率水平 t：是标准差的得分，正态分布情况下，95%对应的 2 倍标准差得分是 1.96。用于表示 100 次重复抽样或调查中，希望有 95%的把握，样本的置信区间包含了总体值。

方差 p*q：分类变量（编码成 2 个类别）的方差，为这 2 个概率的乘积。如果多分类，则需要将其编码成 2 分类。如果是连续型变量可以使用方差、标准差和标准误。

方差需要我们预估：如果有理由认为当前样本与总体一致，可以使用已有的总体信息，或根据预调查的样本估计。

置信区间：希望的估计值落入该区间的大小。

例如，有研究需要调查该地区的拥有住房率情况，通过预调查发现，某几个小区的总和

拥有率为 20%，如果概率水平确定在 1.96，置信区间确定在 0.05（统计学家的一般建议），则：$n=\frac{1.96^2\cdot 0.2\cdot 0.8}{0.05^2}$=246 个样本。这里没有考虑有限总体修正。

双样本情况

假设检验中 2 种类型的错误

α：原假设为真，但拒绝了原假设；

β：原假设为假，但接受了原假设；

样本量的确定中，1- β被定义为检验力，其目的是确定合适的样本量，以在最大限度地增加原假设被拒绝时，拒绝它的概率。所以它与α是从不同的角度来避免样本量的大小带来的决策误差。

表 1：α与β的单位正态偏差

α与β	单侧检验	双侧检验
0.01	2.33	2.57
0.025	1.96	2.24
0.05	1.65	1.96
0.1	1.28	1.65
0.2	0.84	1.28

一般经验法则建议将β设置为 4 倍的α。

假如希望检验本地区的拥有住房率 20%与总体 30%的拥有住房率相比是否存在差异。需要确定的样本量为：

$$n=(\frac{Z_{\alpha/2}\sqrt{p_0\,q_0}+Z_{\beta}\sqrt{p_1\,q_1}}{p_1-p_0})^2$$=((1.96*sqrt(0.2*0.8)+ 0.84*sqrt(0.3*0.7))/(0.3-0.2))**2=137

相同的，在本问题中，方差需要我们预估：如果有理由认为当前样本与总体一致，可以使用已有的总体信息，或根据预调查的样本估计。分类变量编码成 2 分类，连续变量有时也会根据感兴趣的区间转化成分类概率。例如年龄问题中，如果我特别感兴趣班级中大龄儿童在友谊互动中的表现，可以分成两个年龄段计算概率。

诸多样本量确定技术

a) 调查分析中的样本量不建议小于 30 个；
b) 关键变量的测量误差，尤其是因变量，如果有理由认为误差增加，样本量需相应增加；
c) 变量个数的 10 倍-30 倍之间为宜；
d) 模型的复杂度，如果涉及到多组模型的联立，建议 200 样本量以上，模型越复杂，所需样本量越多；
e) 一个常用的折中方案：普通模型的样本量在 60 到 200 之间；结构方程、广义方程、混合模型等类的方法建议（稍微复杂一些的方法），300 到 500 之间。
f) 变量的测量上，如果级别较低，但需要将其视为较高级别的变量，如将次序型变量视为间距型变量，则需要的样本量需要增加，如 200 以上。

（2）抽样技术理解与实操

- **概率抽样**
- **非概率抽样**

抽样技术包括概率抽样和非概率抽样

*抽样总体与目标总体*应该是完全一致的，但往往很难完全保证。如目标总体是某地区所有持有驾照的人的驾驶行为研究，以持有驾照为抽样框，但有很多人有驾照，但从不开车。

抽样框：抽样总体的具体表现是抽样框，是一份能包括所有抽样单元的名单。常见的抽样框有——学校的学生名单、工商局的企业名册、电信的电话簿等。

- **概率抽样**

简单随机抽样（simple random sampling）的定义是：任何样本数为 n 的样本组合中选的机率都是相等的。

特点：

a) 需要抽样框，当 N 很大时这样的抽样框很难获得。

b) 这种调查获得的样本比较分散，调查效率变低。所以大型抽样时很少使用。

c) 但精度较高。

系统抽样（systematic sampling）是先把全体总数N除以样本数 n，得到 k，也就是每间隔 k 个抽一个，再用随机数表自 1 到 k 选一个随机数 r，则 r，r + k，r + 2 k，……，r +（n − 1）k 等号码中选。

特点：

a) 操作简单，如果总体内是有组织排列的结构，可以有效的提高精度。

b) 总体内是有组织结构不易判断时，估计量方差很难估计。

分群抽样（cluster sampling）是先把总体分割成许多小集群，把这些小集群编上号码，然后随机抽取这些号码，凡是被抽中的，则整个小集群的所有成员全部调查。

特点：

a) 抽样框简化：相比获得全市学生名单，而言获取全市的学校（群）名单要容易的多。

b) 调查更加便利：只需要对抽中的群进行调查就可以了，调查对象相对集中。

c) 但精度低于简单随机抽样。

PPS 抽样（与规模大小成比例的有放回的不等概率抽样）

分层随机抽样（stratified random sampling）是先把总体的所有个体依某些特征分类，也就是分层，然后在各层之内再进行独立的随机抽样（参见《抽样技术》）。

特点：

a) 抽样样本与总体结构更相近。

b) 既可以对总体参数估计，也可以对各层目标量进行估计

分群与分层的区别：分群目的是使群内差异最大化，而分层目的是使层内差异最小化。

分层样本量的确定：一个可行方案

现希望获得某地区 32 所小学中，所有学生的数学成绩状况。总人数为 6800 名，但项目费用有限，前提限定了抽样样本不能超过 600 个，通过预调查发现，全市小学需要分成 3 个地区（根据师资、人口等因素），学生人数分别为 1700、2266、2834，并且平均调查费用比为 5：2：1，预调查样本标准差为 12、8、14（分别对应地区一二三）。

h	N_h	W_h	c_h	S_h	W_hS_h
1	1700	0.25	5	12	3
2	2266	0.33	2	8	2.64
3	2834	0.42	1	14	5.88

总计	6800				11.52

根据内曼最优分配法(假设各层调查费用相同)：

$$n_h = n\frac{W_h S_h}{\sum_{h=1}^{L} W_h S_h}$$

n_1 =600$*\frac{3}{11.52} = 156.25; n_1$ =600$*\frac{2.64}{11.52} = 137.5; n_1$ =600$*\frac{5.88}{11.52} = 306.25$

根据一般最优分配法(假设各层调查费用不同)：

$$n_h = n\frac{W_h S_h/\sqrt{c_h}}{\sum_{h=1}^{L} W_h S_h/\sqrt{c_h}}$$

多段集群抽样（multi-stage cluster sampling）是指在第一个阶段先抽出一部分集群，然后在下一阶段选中的集群中，再抽出一部分集群，到最后阶段在抽取若干基本单元（最小单位）。

两段集群抽样——首先随机抽取某些群（初级单元），并且在选中的群内再随机抽取若干基本单元（最小单元）。

三段集群抽样——首先随机抽取某些群（初级单元），第二阶段随机抽取二级单元，并且在选中的群内再随机抽取若干基本单元（最小单元）。

……

特点：具有分群抽样的特点。但需要注意，抽样阶段每增加一级，就多出一份抽样误差，最终对总体的估计会更加复杂。