

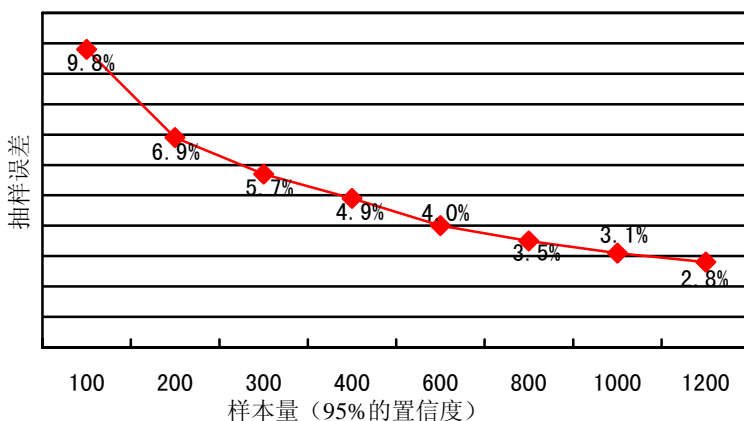
样本量缘何相同

(郑辉 奥得市场咨询公司高级统计分析师 复旦大学经济学院博士)

(李国强 奥得市场咨询公司高级研究经理)

[内容摘要] 在定量的市场研究中，尽管两个不同级别的城市目标消费群体的总人口不同，但抽取的样本量一般相等。直观的感觉是城市的总人口越多，要求抽样的样本量就越大，如何来解释直觉与抽样理论的矛盾。本文试图从现象上的统计推理和本质上的深入探讨这两个层面去剖析这一矛盾，以期解答市场研究客户和市场研究人员的困惑。

在定量的市场研究中，抽样必不可少，确定样本量是抽样的核心。然而，对于多数的市场研究公司来说，样本量的确定已不成问题，定量研究似乎已形成一条不成文的惯例：一线城市一般抽取 300 个样本，二线城市或三线城市最好也抽 300，但考虑成本及执行等问题，200 样本也可接受。



为了说明这一点，研究人员一般绘制上图形象地说明样本量与抽样误差的关系，当样本量为 300 时，最大的抽样误差是 5.7%，即在随后对数据进行统计分析的过程中，样本推断总体的精确度高达 94% 以上，200 样本的精确度也会在 93% 以上。实际的精确度要超过这两个值，这表明市场研究几乎具有完全的科学性，并带有少量的艺术性。

一、问题提出

众所周知，抽样的基本原则是样本要能代表总体，这是随后进行数据分析并推断总体特征的前提条件。样本能不能很好地代表总体，很大程度上取决于样本量的大小。显然，更大样本量的样本代表性更强，小的样本代表性必然较差，但是样本量越大，执行的成本也越高，实际的操作也越困难。样本的代表性和执行成本之间存在着不可调和的矛盾，追求较强的样本代表性就必然要承受较高的成本，而要降低成本就要损失样本的代表性。因而在抽样时，必须在样本代表性与成本核算之间进行权衡，所选取的样本量，既要满足样本对总体的代表性，又要使得成本不是很高。

在市场研究中，样本对总体的多大代表性算是比较合适，很难有一个确切的说法，不过一般认为，95% 的代表性，即最多 5% 的抽样误差是可以接受的（注意：市场研究用的置信度一

般也是 95%，置信度与代表性之间没有必然的联系)。所以，在市场研究中，5%的误差成了行业的定量设计参考线，所谓的 300 样本正是基于这个 5%的抽样误差计算出来的。但是细查上表，发现决定样本量大小的只有抽样误差，而与总体容量大小的没有关系，或者换句话说，不管总体容量多大，所抽取的样本量都是 300，这是令许多接受市场研究服务的客户感到困惑的地方。直觉告诉我们，要保证样本的代表性，越大的总体容量似乎要求更大的样本量。

比如，某次市场研究的对象是一线城市上海、三线城市无锡的城市总人口，上海市总人口为 1600 万，无锡总人口 160 万，如果在无锡，用 300 样本代表 160 万总体，那么在上海，用来代表 1600 万总体的似乎应该是 3000 样本。

在递交给客户的计划书里，除了上图之外，市场研究人员还会对样本量与抽样误差的关系做一番解释，“在 95% 的概率保证程度下，300 样本产生的抽样误差在 5% 左右”。对于这句话，研究人员早就习以为常，然而这句话并不能解释为什么总体容量对样本量的确定没有影响。对统计专业人士来说，他们更多地是去关注理论，意识不到上述问题，所以导致这一基本问题始终悬而未决，好像成了客户手中的“达摩克斯宝剑”，不时地祭出来，很是为难了不少的研究人员。不可否认的是，这一问题确实存在，而且始终没有令人满意的解答。因此，本文试图从两方面回答这一问题，一方面用统计理论，从表面层次上去解答；另一方面透过现象，从本质上深入解释。

二、理论解释

统计学的中心极限定理保证了大样本近似服从正态分布，一般来说，30 以上的样本就可以称为大样本，在定量的市场研究中，这一基本前提都是可以满足的。此时，抽样误差 δ 满足：

$$P\{\delta/\sqrt{\sigma^2(1/n-1/N)} < Z_{\alpha/2}\} = 1 - \alpha$$

其中， $1 - \alpha$ 为置信度， σ 为样本标准差， Z 为正态分布分位数， n 为样本量， N 为总体容量。所以，

$$\delta = Z_{\alpha/2} \times \sigma \times \sqrt{(1/n-1/N)} \quad \text{公式 (1)}$$

在市场研究中，给定的置信度一般为 95%，查正态分布分位数表可得 $Z_{\alpha/2} = 1.96$ ，最大的样

本标准差为 0.5（事实上 $\sigma = \sqrt{P \times (1-P)}$ ， P 为概率，数学知识告诉我们，当 $P = 0.5$ 时， σ 取得最大值 0.5）。因此采用公式 (1) 计算得出的抽样误差 δ 为最大误差，在调查问卷的诸多题目中，几乎没有一个题目能满足 $P = 0.5$ 这一条件，因此实际的抽样误差要比公式 (1) 计算出的小很多。

查看公式 (1)，当总体容量 N 很大时，如上海 1600 万人口，则 $1/N$ 就很小，可以忽略不计，一般来说，当 N 超过 10 万， $1/N$ 的值就可以忽略不计，因而抽样误差 δ 就只与样本量 n 有关，而与城市的人口总量没有关系，这就可以回答为什么上海和无锡的城市人口总量差距很大，但在市场调查时抽取的样本量却相同这一问题。

将 $n=300$ 代入公式 (1)，并忽略不计 $1/N$ ，计算出的最大抽样误差为 5.66%，而实际的抽样误差并没有这么大。因此，选择 300 的样本量来估计总体，其精确度是相当高的，同样的方法计算可得，200 样本的最大抽样误差为 6.9%。

事实上，即便真的考虑了总体容量，在保证同样抽样误差的前提下，一线城市和三线城市

的样本量也不会有很大的差异。如果在无锡抽取了 300 样本，要在一线城市的上海抽取多少样本才能保证与该三线城市的抽样误差相同呢？根据公式（1），计算的结果为 300.06，一线城市和三线城市的样本量差异微乎其微。

三、进一步的解释

诚然，从直觉上看，要满足样本对总体的代表性，似乎是总体容量越大，要求的样本量也越大，即便样本量之比不等于总体容量之比（这里指的是大总体的容量与小总体的容量之比），至少也应该介于 1 和该比值之间。然而这只是直觉，直觉经常是靠不住的，因为若按这一逻辑，假如某次研究的总体容量无穷大（ ∞ ），样本量也应该是无穷大，但无穷大的样本是没法进行统计估计的。因此，从这个角度讲，样本量的大小不可能是与总体容量的大小成一定比例的。

事实上，总体内部的个体之间具有很强的同质性，如果按照某一指标（属性）进行聚类分析，会形成不同的类别，尽管类与类之间具有较大的差异性，但各类内部的个体则是同质的。在这些同类的个体中，只要选择其中的一个进行研究，研究结果就可以很好地代表该类的状况。因此，所谓的抽样，实际上就是从这些类别中各抽取一个个体组成样本，用该样本的信息去推断总体状况。因而，决定样本量大小的并不是总体容量，而是更多地取决于总体可以划分为几个类别。

在市场研究中，我们很少只是对总体的一个特征进行研究，而通常是研究总体的诸多属性，针对一个属性可以进行一种聚类，形成不同类，而针对另外一个属性又可以进行另外一种聚类，形成另外不同类，样本中的每个个体是这些不同类别之间交叉形成的最小类中的一分子，这就是为什么我们每次抽样的样本容量都较大（300 或 200），而不是只抽取十几或几十个。

尽管上海的总人口是无锡的十倍之多，但因为总体中的个体之间具有同质性，按三线城市总人口划分的类别与按一线城市总人口划分的类别是一样的，使得上海和无锡抽取的样本量相等。上海的人口总量多，但并不多在类别上，而仅仅是多在各类别中具有同质性的个体数量上。这就是总体容量不等但样本量相等的本质原因。

作者简介

郑辉，男，1978 年生，奥得市场咨询公司高级统计分析师，就读复旦大学经济学院世界经济系博士，2003 年毕业于厦门大学，获统计学硕士学位。

李国强，男，1977 年生，奥得市场咨询公司高级研究经理，2000 年毕业于复旦大学。

联系地址：上海市静安区青海路 118 号云海办公楼 18 楼奥得市场咨询公司

邮政编码：200041

联系电话：021-52288318*224,13564347268