



第 5 章 样本容量的确定

主要内容:

- 样本容量确定的方法
- 正态分布及其评价
- 理解总体、样本和抽样分布
- 点估计与区间估计
- 抽样平均数和抽样比例

5.1 确定概率抽样的样本量

在确定概率抽样样本容量的过程中会遇到涉及财务、统计和管理三个方面的问题。一般原则是，样本越大，抽样误差就越小。但样本大，耗费的成本也高，而且一个项目可支配资源毕竟是有限的。虽然抽样成本随着样本容量的增加呈直线递增（样本容量增加 1 倍，成本也增加 1 倍）。抽样误差却只是以样本量相对增长速度的平方根递减。即如果样本量增加了 3 倍，数据搜集成本也增加了 3 倍，而抽样误差只降低了 1/2。最后一点，样本容量计算还反映了管理方法的问题。要求多高的估计精确度？实际总体值在所选定的置信区间内的置信度是多少？正如你将在本章中学到的，有许多种可能性。有的情况要求精确度较高（抽样误差很小），并且要求总体值在较小误差范围以内的置信度较高。而有些情况则不要求这些。

5.2 确定样本容量的方法

5.2.1 可支配预算

某一研究对象的样本容量通常直接或间接地由可支配的预算额所决定。因此，顺序上，样本容量通常是稍后才确定的。一个品牌经理如果有 40000 美元预算可用于某项市场研究，那么除去其他项目成本（如调查方案和问卷的设计、数据的处理、分析等）后，余下的那部分预算才决定着被调查的样本容量的大小。如果可支配资金太少，可以确定的样本量太小，就必须做出决策，是补充更多的资金还是放弃这一项目。

虽然这种方法看来缺乏科学性和过于武断，但是在一个离不开财务资源预算编制的整体环境下它确实存在。财务上的限制要求调查人员的设计方案要利用有限的资源提供有利于决策的高品质的数据资料。“可支配预算”方法使调研人员不得不寻求多种选择的搜集方法并谨慎衡量信息的价值及其成本。

5.2.2 单凭经验的做法

一些客户人指定 REPs（对计划的具体要求），他武会要求样本容量为 200、400、500 或其他的特定量。这个数据的确定有时是出于对抽样误差的考虑，而有时则只是依据以往的经验 and 过去进行的相似调研中采用的样本量。对指定样本容量这种做法的全理解释归结起来只能说是“一种强烈的感觉”，认为某一特定的样本容量是必要的或适当的。

也许有人认为客户指定的样本容量有利于计划调研目标的实现。有些情况下，调研人员会认为指定的样本容量不符合要求。这时，调研人员有职责向客户提出扩大样本容量的建议并让客户做出最后的决定。如果扩大样本容量的建议遭到了否决，调研人员会拒绝提交计划，因为他（她）认为样本容量不合要求会严重影响调研成果。

5.2.3 要分析的子群数

在任何确定样本容量的问题中，都必须认真考虑所要分析并要据此做统计推断的总体样本的各个子群的数目的预期容量。例如，从整体上看样本容量为 400 很符合要求，但若分别分析男性和女性被调查者，并且要求男性与女性的样本各占一半，那么每个子群的容量仅



为 200。这个数字是否符合要求，能使分析人员对两组的特征做出预期的统计推断呢？再如，要按年龄和性别分析调研结果，问题就变得更复杂了。假设要按以下方式将总体样本划分为四组：

- 35 岁以下的男性
- 35 岁以上的男性
- 35 岁以下的女性
- 35 岁以上的女性

如果预计每组约占总样本的 25%，那么子群容量仅有 100。这个数字能否使我们按照调研目标的要求对各组分别做出统计推断呢？随着样本量的缩小，抽样误差的增加，会出现这样一个问题，那就是调查人员很难辨别依据现象所得的两组间的差别（如表明打算购买新产品的百分比）是真正意义上的差别还只是由抽样误差引起的差别。

在其他条件相同的情况下，所要分析的子群数目越大，所需的总样本容量也就越大。一般认为样本量要足够大，以便每个主子群的容量至少为 100，而每个次子群的容量至少也有 20-50。

5.2.4 传统的统计方法

你可能在其他书上见过确定简单随机样本的传统方法。回顾一下这些方法。在利用抽样结果做重要推断时需要三条信息：

- 总体标准差的估计值
- 抽样的允许误差范围
- 抽样结果在实际总体值的特定范围（抽样结果±抽样误差）内的预期置信度。

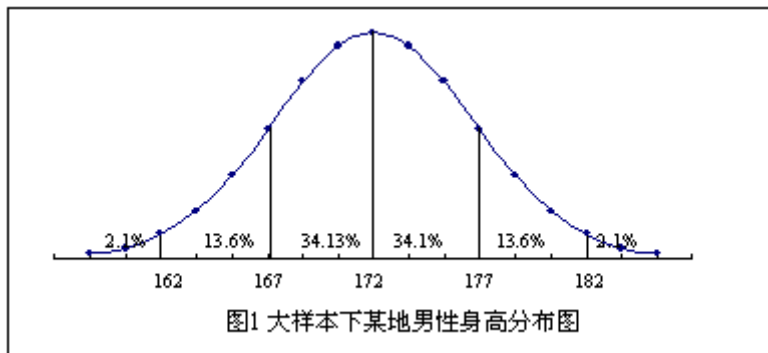
有了以上三条，就可以计算出简单随机抽样所需的样本容量了。

5.3 正态分布

5.3.1 总体特征

在古典统计推断中，正态分布居于特别重要的地位。这有以下几方面原因：首先，市场人员遇到的许多变量其概率分布都趋于正态分布。如，软饮料包装的数量；爱吃快餐的人平均每月吃快餐的次数；每星期看电视的平均小时数。其次，有理论上的原因。比较重要的一条是大数定理、中心极限定理。根据该定理，对于任何总体，不论其分布如何，随着样本容量的增加，其抽样平均数的分布趋于正态分布。这种趋向的重要性将在后面做详细说明。再次，许多离散型概率的分布也近似于正态分布。例如，将大量的某地男性身高值标在一张图表上，就会得到如图 5-1 的分布图，这种分布就是正态分布，它有以下几个重要的特征：

- (1) 正态分布呈现钟形且只有一个众数。众数代表着集中的趋势，是发生频率最高的那个特殊值。两峰的（两个众数）分布有两个峰值；
- (2) 正态分布关于其平均对称。也就是说它是对称的。它集中趋势的三个衡量标准（平均数、中位数和众数）是相等的。
- (3) 一个正态分布的特殊性由其平均数和标准差决定。
- (4) 正态曲线下方面积等于 1，表明它包括了所有的调查结果。
- (5) 正态曲线下方在任意两个变量值之间的面积，等于在这一范围内随机抽取一个观察对象的概率。以图 5.1 为例，一次抽取到一名男性，其身高在 172cm-177cm 之间的概率为 34.13%。
- (6) 正态分布还有一个特点，就是所有的正态分布在平均数±1 个标准差之间的面积相同，都占曲线下方面积的 68.26%或者说是占全部调查总体结果的 68.26%。这叫做正态分布的比例性，这一特点为本章将要讨论的统计推断提供了基础。



5.3.2 标准正态分布

任何正态分布都可以转换为标准正态分布。标准正态分布的特点与正态分布相同。只有标准正态分布的平均值等于 0，标准差等于 1。正态分布的任何一变量值 X 通过一个简单的转化公式就能变换成相应标准正态分布中的 Z 值。这种转换是由正态分布的比例性决定的。用符号表示：

$$Z = \frac{\text{变量值} - \text{变量平均值}}{\text{变量标准差}}$$

用符号表示：

$$Z = \frac{X - \mu}{\sigma}$$

- 式中 X——变量值；
- μ——变量平均值；
- σ——变量标准差。

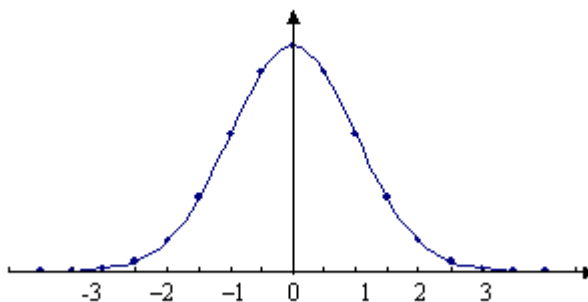


图 5.2 标准正态分布图

变量 Z 的标准正态分布曲线下各块面积（全部百分比）都列在表 5.1 中。

表 5.1 Z 值为 1, 2, 3 时标准正态曲线下方的面积

Z 值（标准差）	标准正态分布曲线下方的面积（%）
1	68.26
2	95.44
3	99.74



5.4 总体分布、样本分布和抽样分布

进行抽样调查，目的是要对总体做出推断，而不是为了描述样本的特征。总体，就像前面定义的，包括可以从中获取信息达到调研目标的全部可能的人体或物体。样本是总体的子集。

总体分布是总体中所有单位的频率分布。这一频率分布的平均数，通常用希腊字母 μ 表示，标准差用希腊字母 σ 表示。样本分布是单个样本中所有单位的频率分布。样本分布的平均数常用 \bar{x} 表示，标准差用 S 表示。

在这里，有必要介绍一下三种分布，样本平均数的抽样分布。理解这一分布对于充分认识估计简单随机抽样误差的依据十分重要。样本平均数的抽样分布是指从一个总体中抽取一定数量的样本，由样本平均数构成的概率分布。虽然人们对很少计算这种分布，但它的特性具有很大的实际意义。要获得样本平均数的分布，首先要从特定总体中抽取一定量的样本（如 25000），接着，计算各样本的平均数，并排列出频率分布。因为每个样本由样本单位的不同子集构成，因此样本平均数不会完全相同。

当样本的单位数和随机性足够大，样本平均数的分布近似于正态分布。这一论断的基础是中心极限定理。该定理说明，随着样本容量的增加，从任一总体中抽取的大量随机样本的平均数的分布接近正态分布且平均数等于 μ ，标准差（也称之为标准误差）等于：

$$S_x = \frac{\sigma}{\sqrt{n}}$$

式中 n ——样本单位数。

值得注意的是，中心极限定理的成立不考虑样本总体的分布形状，也就是说忽略了总体的分布类型，样本平均数的分布会趋于正态分布。常用来表示总体分布、样本分布和抽样分布的平均数及标准差的符号都列在表 5-2 中。

表 5.2 参数、统计量符号

分布	平均数	标准差
总体分布	μ	σ
样本分布	\bar{x}	S
抽样分布	$\mu_x = \mu$	S_x

平均数的标准误差（ S_x ）之所以按前面所示的方法计算是因为，一个特定的样本平均数分布的方差或是离差会随着样本数量的增加而减少。由常识可知，样本数越大，单个样本的平均数就越接近总体平均数。图 13-3 表明了平均数的总体分布、样本分布和抽样分布之间的关系。我们将深入讨论平均数的抽样分布，而另一个比例抽样分布，将在以后介绍。



5.5 平均数的抽样分布

5.5.1 基本概念

考虑一个抽样案例：一位调查人员以“在最近 30 天内至少吃过一次快餐的所有顾客”为总体，从中抽取了 1000 组容量为 200 的简单随机样本。调查目的是要估计平均一个月内这些人吃快餐的平均次数。计算出每一组的平均数，按相关值确定区间，整理后便得到表中 5-3 的频率分布图。而图 5-4 以直方图的形式表示这些频率，直主图上方还可见到一条正态曲线。正如你所看见的，直方图十分接近正态曲线的形状。如果我们选取足够的容量为 200 的样本，计算每组的平均数，整理排列后所得的分布就是正态分布。图 5-4 的正态曲线就是这项调查中平均数的抽样分布。大样本平均数的抽样分布有以下特征：

- (1) 是正态分布
- (2) 分布的平均数等于总体平均数。
- (3) 分布有标准差，称为平均数的标准误差，它等于总体标准差除以样本容量的平方根：

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

将标准差称为平均数的标准误差表明它更适用于样本平均数的分布，而不是总体或样本的标准差分布。记住这种计算只适合简单随机样本，其他类型的样本（如分层样本和整群样本）要用非常复杂的分式计算标准误差。

表 5.3 1000 个样本平均数的频数分布

序号	组别	频数	序号	组别	频数
1	2.6-3.5	8	10	11.6-12.5	110
2	3.6-4.5	15	11	12.6-13.5	90
3	4.6-5.5	29	12	13.6-14.5	81
4	5.6-6.5	44	13	14.6-15.5	66
5	6.6-7.5	64	14	15.6-16.5	45
6	7.6-8.5	79	15	16.6-17.5	21
7	8.6-9.5	89	16	17.6-18.5	16
8	9.6-10.5	109	17	18.6-19.5	9
9	10.6-11.5	125	合计		1000

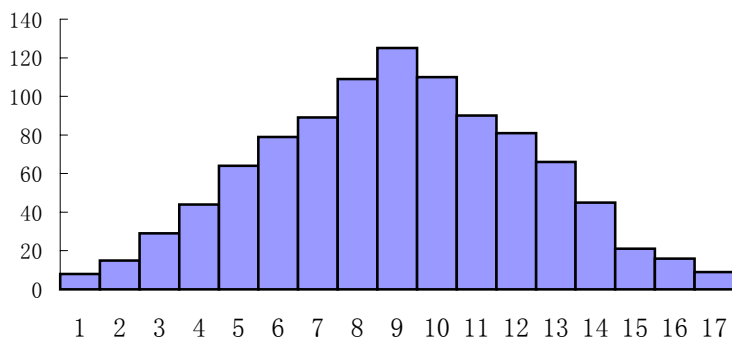


图 5.3 平均数的实际抽样分布



5.5.2 根据单个样本做出推断

在实际操作中，人们往往不愿从总体中抽出所有可能的随机样本，画出像表 5.3 和图 5.4 那样的频率分布表和直方图来。人们希望进行简单的随机抽样，并据此对总体进行统计推断。问题出现了，通过任一简单的随机样本对总体均值进行的估计，其估计值在总体平均值±1 个标准误差内的概率究竟为多大？根据表 5.2 可知概率为 68%，因为所有样本平均数有 68% 都在此范围内。而通过简单随机样本对总体做的估计为实际总体平均值 2 倍标准误差范围内的概率为 95%，在实际总体平均值 3 倍标准误差范围内的概率为 99.7%。

5.5.3 点估计和区间估计

当利用抽样要对总体平均值进行估计时，有两种估计方法：点估计和区间估计。点估计是指把样本平均值作为总体平均数的估计值。观察图 5.3 的平均数抽样分布可知某一特定的抽样结果，其平均数很可能相对更接近总体平均数。但是，样本平均数分布中的任一个值都可能是这一特定样本的平均值。有一小部分的样本平均值与实际总体平均值有相当的差距，这种差距就叫抽样误差。

抽样结果的点估计在很少的情况下完全准确，因此人们更偏于区间估计。区间估计就是对变量值如总体平均值的区间或范围进行估计。除了要说明区间大小外，习惯上还要说明实际总体平均值在区间范围以内的概率。这一概率通常被称为置信系数或者置信度，区间则被称为置信区间。

平均数的区间估计按以下步骤推导。从总体上抽出一定量的随机样本，计算出样本平均数，可知这个样本平均值存在于所有样本平均数的抽样分布中，但确切位置不清楚。此外还知道，这个样本平均数在实际总体平均值±1 个标准误差范围内的概率为 68%，由此可知，实际值等于样本值加上或减去 1 个标准误差的置信度为 68%。用符号表示如下：

$$\bar{x} - 1\sigma \leq \mu \leq \bar{x} + 1\sigma$$

同理可知，实际值等于样本估计值加上或减去 2 倍标准误差（严格上是 1.96，但为了计算简便通常用 2）的置信度为 95%，实际值等于样本值加上或减去 3 倍标准误差的置信度为 99.7%。

以上都假设总体标准差已知，大多数时候，情况不是这样。如果总体标准差已知，根据定义可以知道总体平均值，那就没有必要事先抽取样本了。而如果不知道总体标准差，那就必须通过样本差去估计。

5.6 比例的抽样分布

市场研究中经常会偏于进行比例或百分比方面的估计。下面是一些常见例子：

- 知道某一广告的总体百分比；
- 平均一周上网 1 次以上的总体的百分比；
- 最近 30 天内吃过快餐和吃过 4 次以上快餐的总体百分比；
- 观看某一电视节目的观众的总体百分比；

在上述情况下，总体比例或百分比是重要的因素，因此有必要介绍比例抽样分布。

从特定总体中抽出大量随机样本，这些样本的抽样比例的相对频率分布就是比例抽样分布，它有以下特征：

- 近似于正态分布
- 所有样本比例的平均值等于总体比例。
- 比例抽样分布的标准误差可以按下面的公式计算：

$$S_p = \sqrt{\frac{P(P-1)}{n}}$$



式中 S_p —比例抽样分票误差;

P —总体比例的估计值;

n —样本单位数。

考虑一下, 如果需要估计一下最近 90 天内曾在网上购物的所有成年人的百分比, 那么就像要得到平均数的抽样分布一样, 要从成年人总体中选取 1000 组容量为 200 的随机样本, 计算出 1000 组样本中所有在最近 90 天内曾在网上购物的人数比例。这些值排列将形成一个趋于正态分布的频率分布。这一分布的估计比例标准误差可以用在前面计算比例标准误差的公式来计算。

读完下一节, 你就会明白, 市场人员对于样本容量问题, 更趋于进行比例估计而不是平均值估计, 是有其原因的。

5.7 样本容量的确定

5.7.1 平均值问题

考虑前面那个估计平均一个月快餐族吃快餐次数的案例, 如果管理层需要对顾客的平均光顾次数做出估计, 从而决定是否实行正在拟定的新促销计划。为了得到这个估计值, 市场调研经理打算在总体中考察某个简单随机样本。问题是, 确定本次调查样本容量的要素是什么? 首先, 对于估计平均值问题, 计算所需的样本容量的公式是:

$$n = \frac{Z^2 \sigma^2}{E^2}$$

式中 Z —标准误差的置信水平;

σ —总体标准差;

E —可接受的抽样误差范围 (允许误差)

计算所需的样本容量要有三种资料:

- (1) 抽样误差的可接受的或允许的详细范围 (E)。
- (2) 标准误差置信水平的允许确切值, 也就是 Z 值。换一种说法, 即总体平均值包括在指定置信区间内的置信度是多少?
- (3) 最后需要估计一下总体标准差 (σ)。

计算中要用到的置信水平 (Z) 和误差 (E) 必须由调查人员与他 (她) 的客户进行磋商后才能确定。如前所述, 置信水平与误差范围的确定不仅要根据统计原则, 同时要顾及财务与管理方面的要求。理想的情况下, 我们总是希望置信度很高, 误差很小。但要知道, 这是经营决策, 必须考虑成本问题。因此, 要在精确度、置信度与成本之间进行权衡。有的时候, 不要求很高的精确度与置信度。例如, 你也许只想通过调查基本了解一下消费者对产品的普遍态度是正面有还是负面的。这里精确度就显得不太重要了。但如果是一项产品创意测试, 就需要精确度较高的销售估计值, 以便做出是否向市场推荐某种新产品的高成本、高风险的决策。

第三项是总体标准差的估计值, 这是一个更麻烦的问题。我们在前面说过, 如果总体标准差已知, 那么也就能知道总体平均数 (总体平均数是用来计算总体标准差的)。这样的话就没必要抽取样本了。但调查人员如何不抽取样本就估计出总体标准呢? 结合使用以下四种方法可以解决这个问题:

- (1) 利用以前的观察结果。许多情况下, 公司以前曾经进行过类似的调查, 这时, 可以利用以前的调查结果作为本次总体标准差的估计值。
- (2) 进行试点调查。如果调查对象规模太大, 可以投入一定的时间和资源对总体进行小规模的小规模试验调查。根据调查结果估计总体标准差确定样本容量。
- (3) 利用二手数据。有时候通过二手数据也可以对总体标准差做出估计。



(4) 通过判断。如果其他方法都失败了，还可以判断总体标准差。即把许多管理人员的判断集中起来进行分析，而这些管理人员都有能力对有关的总体参数做出有所根据的猜测。

当完成了调查，计算出样本平均值和样本标准差后，调查人员就可以正确估计出总体标准差，并确定所需的样本容量了。这时如果需要，可以对以前的抽样误差估计做出调查。

再来考虑估计快餐族平均每月吃快餐的平均次数。以下这些值将代入下面的公式。

- 与公司的管理者进行磋商后，市场调研经理认为有必要估计一下吃快餐的平均次数。考虑到管理者对精确度的要求，她规定估计值不得超过实际的 0.10 (1/10)。这个值 (0.01) 将作为 E 值代入公式。
- 此外，市场调研经理还认为，考虑全局，需要把实际总体平均值在 (样本平均值 ± E) 区间以内的置信度定为 95%。而若要置信度为 95%，就必须在 2 倍标准误差范围内 (严格是 1.96)。因此，2 作为 Z 值代入公式。
- 最后，确定公式中的值。幸好公司一年前曾做过类似的调查。调查对象是最近 30 天内吃快餐的平均次数。其标准差是 1.39，以此作为 σ 值最好不过。因此把 1.3 代入公式。然后进行计算，通过计算，可知样本容量为 772 时可以满足提出的要求。

$$n = \frac{Z^2 \sigma^2}{E^2} = \frac{2^2 (1.39)^2}{(0.1)^2} = 772$$

5.7.2 计算比例的问题

考虑估计最近 90 天内曾在网上购物的所有成年人的比例或百分比的案例。其目标是从成年人总体中抽取一个简单随机样本，估计其比例是多少。下面讨论一下如何确定代入公式的那几个值：

- 像前面说的，要根据抽样结果估计总体平均值，首先要确定 E 的值。例如，假设可接受的误差范围为 ±2%，那么将 0.02 作为 E 的值代入公式。
- 其次，假设调查人员要求抽样估计在实际总体比例的 2% 范围以内的置信度为 95%，那么按前面讲的，把 2 作为值代入公式。
- 最后一点，在一年前的一次类似调查中，有 5% 的被调查者表示在最近 90 天内曾在网上购物。我们可以用 0.05 作为 P 值代入公式。

计算过程如下：

$$n = \frac{Z^2 [P(1-P)]}{E^2} = \frac{2^2 [0.05(1-0.05)]}{0.02^2} = 475$$

根据要求，需要一个 475 个人的随机样本。要注意的是，与确定估计平均值所需的样本容量的过程相比，调查人员在确定估计比例所需的样本容量时有一个优势：如果缺乏估计 P 的依据，可以对 P 值做最悲观或最糟糕的假设。给定 Z 值和 E 值，P 值为多大时要求的样本量最大呢？当 P=0.5 时，“P (1-P)” 有极大值 0.25 存在，如此设定 P 值样本是最大。而给定 Z 值和 E 值，对于与平均估计所需样本量有关的值就没有最悲观的假设。

5.7.3 总体容量样本容量

你也许会注意到计算样本容量的公式中没有用到总体容量。学生们 (和经理们) 经常会注意到这个问题。表面上看来好像是要抽取的样本量越大，其总体容量也应该增大。其实不然。通常，总体容量与在一定误差和可靠度范围内估计总体参数所需的样本容量之间没



有直接的关系。实际上，总体容量只有当样本容量相对它而言过大时才会起作用。根据经验，当样本容量超过总体的 5% 时，就需要调整样本容量了。一般都假设样本的抽取是相互独立的（独立假设），这一假设在样本相对于总体很小时成立。当样本量占总体比例相对较大（5% 以上）时就不成立了。因此，我们必须调整一下标准公式。譬如，前面的计算平均数标准误差的公式是：

$$\sigma_x = \frac{\sigma}{\sqrt{n}}$$

当样本量占总体 5% 以上，就要推翻独立假设。调整后的正确公式是：

$$\sigma_x = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

其中， $(N-n)/(N-1)$ 被称为有限总体修正系数（FPC）。当样本占总体的 5% 以上，调查人员可以通过 FPC 来减少所需的样本容量。计算公式如下：

$$n' = \frac{nN}{N+n-1}$$

式中， n' 为修改后的样本量； n 为原样本量； N 为总量。

如果总量为 2000，原样本量为 400，则：

$$n' = \frac{400(2000)}{2000+400-1} = 333$$

经过 FPC 的调整，需要的样本量由原先的 400 变成了 333。

问题关键不是样本量大小与总量大小的关系，而是选取的样本是否能真实代表总体的特性。经验表明，经过仔细挑选的样本，尽管容量不大，却也能十分准确地反映总体特征。许多著名的全国性调查和民意测验的样本数都不超过 2000。盖洛普民意测验、哈里斯民意测验和尼尔森电视节目受欢迎程度调查都是很好的例子。这些例子都表明，即使调查对象是数千万人的行为，也可以通过对于总体相当小的一部分样本进行十分准确的预测。

5.7.4 确定分层样本和整体样本的容量

本章列出的计算样本容量的公式只适用于简单随机样本。当然也有适用于其他如分层样本、整群样本确定样本容量和抽样误差范围的公式。虽然本章提到的许多概念对这些样本都适用，但它们的计算公式却要复杂很多。而且，公式中要用到的数据往往很难得到。因此，这些样本的容量确定问题超过了本书的介绍范围。有兴趣的读者可以参考高级教材。

5.8 统计权

尽管在市场调研中用本章节公式计算样本量是十分标准的作法，但这些公式都只承认第一类误差（不存在差值时推断差值存在而产生的误差）。它们显然不考虑第二类误差，即实际存在差值时认为没有差值而产生的误差。不发生第二类误差的概率叫统计权。计算样本容量的标准公式默认统计权为 50%。举个例子，如果要确定两种产品中哪一个对目标顾客群更有吸引力，并且可能进行购买的目标顾客的百分比之间可以有 5% 的差值，这时标准样本容量公式要求每项产品需要的样本容量大约为 400。通过这一计算，我们默认了一个事实，即有 50% 的可能我们会错误地推断出两种产品具有相等的吸引力。

参考文献：

- 1 《当代市场调研》 Carl McDaniel, Jr and Roger Gates 著，范秀成等译 机械工业出版社出版 2001
- 2 《实用统计分析方法》 蒋庆琅著，方积乾等译 北京大学、中国协和医科大学联合出版社出版 1998
- 3 《社会统计分析方法》 郭志刚主编 中国人民大学出版社出版 1999