企业抽样调查实践

Business Sampling Survey Practice

田秀华 编

中国统计出版社
China Statistics Press

企业抽样调查实践

田秀华　编

中国统计出版社
China Statistics Press

(京)新登字041号

图书在版编目(CIP)数据

企业抽样调查实践/田秀华编.
-北京:中国统计出版社,2005.3
ISBN 7-5037-4599-1

Ⅰ.企…
Ⅱ.田…
Ⅲ.工业企业-宏观管理-抽样调查
Ⅳ.F406

中国版本图书馆CIP数据核字(2005)第014787号

企业抽样调查实践

责任编辑/郭　栋
封面设计/艺编广告·杨燕超
出版发行/中国统计出版社
通信地址/北京市西城区月坛南街75号　邮政编码/100826
办公地址/北京市丰台区西三环南路甲6号
电　　话/(010)63459084、63266600-22500(发行部)
印　　刷/科伦克三莱印务(北京)有限公司
经　　销/新华书店
开　　本/880×1230mm　1/32
字　　数/300千字
印　　张/11.875
印　　数/1—2000册
版　　别/2005年3月第1版
版　　次/2005年3月北京第1次印刷
书　　号/ISBN 7-5037-4599-1/F·1760
定　　价/28.00元

序

抽样调查作为一种统计调查方法，已有百余年的历史。目前，抽样调查方法在世界各国的统计调查中广泛应用，也是各国政府统计部门采集数据的最常用、最主要的方法。我国的抽样调查工作，特别是企业抽样调查工作起步较晚，但经过多年的实践和探索，我国的企业抽样调查工作已取得重大突破，正逐步走向规范化的轨道。

田秀华同志在多年从事工业企业抽样调查实践的基础上，编写了这本《企业抽样调查实践》一书，全面、系统地介绍了工业企业抽样调查的操作过程，并在每个环节的操作中辅之案例说明，同时也介绍了一些相关的抽样理论和方法，并收录了一些国家的企业抽样调查方法和数据采集方法。因此，本书的最大特点就是在工业企业抽样调查的实践基础上，理论联系实际地将企业抽样调查工作从设计到发布的全过程进行详细介绍，这对于无论是从事理论研究还是实际工作的同志都具有一定的参考价值。

国家统计局局长李德水同志在全国统计工作会议上指出："随着我国经济主体多元化和经济活动复杂化程度的不断提高，大力推行抽样调查方法，确立抽样调查在统计工作中的主体地位已是刻不容缓的事情。"愿田秀华同志的这本书能对抽样调查技术与方法在我国统计工作中更加广泛应用起到积极推动作用，同时也希望更多的人关注并投身到抽样调查的实践和研究中来，以进一步推动我国抽样调查工作的开展，提高我国抽样调查工作水平。

宋汝红

前　言

我国企业抽样调查工作始于1996年，经过多年的实践探索和开展国际合作，全国范围内的企业抽样调查工作不论在技术层面，还是在应用范围层面，都已取得重大突破。特别是在工业统计领域的抽样调查技术与方法的应用获得了成功，形成了一套既符合中国国情，又与国际接轨、并得到国内外专家普遍认可的抽样调查方案，成功探索出了一条中国开展企业抽样工作的路子。

为使更多人了解国家企业抽样调查工作开展情况，使理论界更加关注我国的抽样调查工作，有针对性地开展研究，也为实际工作者提供一个规范的操作手册，本人在从事工业企业抽样调查实际工作基础上，尝试着编写了这本《企业抽样调查实践》一书，力求将我国工业企业抽样调查的操作过程做一系统介绍。本书的第一章至第六章是抽样设计阶段所需知识及企业抽样设计过程，包括抽样调查的基本知识，抽样设计的前期准备工作，法人企业、个体户的抽样设计和连续性调查中样本维护和样本轮换方法等内容；第七章至第十章是数据采集阶段的调查设计和组织实施过程，包括数据采集方法，调查问卷和控制表设计及调查的组织实施等内容；第十一章至第十四章是数据处理推算阶段所使用的方法，包括数据加工处理、估计和所使用的软件及数据结果及误差分析等内容；第十五章是数据发布；第十六章是对本次调查方案设计的评估方法。本书的最后一部分是附录，附录A和附录B是在D省试点时所使用的三个调查手册和抽样调查试点方案；附录C和附录D是在工业企业抽样调查中所使用的企业登记注册类型和国民经济行业分类

标准;附录E收集整理的是国外企业抽样调查方法实例。限于本人的学识和水平,难免有错误和不妥之处,敬请读者批评指正。

最后我要特别说明:这本书虽是我个人编写,但书中很多内容都是全国企业调查队系统从事企业抽样调查工作的领导和同志们集体智慧的结晶。在编写此书过程中,也得到了领导和许多同事的支持和鼓励。在此一并向他们表示诚挚的谢意!

田秀华

2004年10月18日

目　录

第一章　抽样调查基本知识

§1.1 抽样调查的基本概念

一、概率抽样与非概率抽样

根据样本抽取方法的不同，抽样可分为概率抽样和非概率抽样两类。

1. 概率抽样（probability sampling）：也称随机抽样，与非概率抽样不同的是概率抽样是严格地按照给定的概率来抽取样本，即按随机原则抽取样本。概率抽样包括等概率和非等概率抽样，随机原则不等于等概率原则。估计量不仅与样本观测值有关，也与其入样概率有关。概率抽样能够计算抽样误差，并能从概率意义上控制误差并以此来保证推断的准确性。需要说明的是以下概念都是指概率抽样。

2. 非概率抽样（non - probability sampling）：是相对于概率抽样而言，并无严格的定义，如我国的典型调查和重点调查，在西方国家称为有目的的调查或判断抽样等都属于非概率抽样。特点是样本的抽选是根据主观判断有目的有意识地或根据方便的原则进行，而不是按随机原则来抽选。这种抽样效果的好坏在很大程度上依赖于抽样者的主观判断能力和经验，且不能计算抽样误差，不能从概率意义上控制误差并以此来保证推断的准确性。非概率抽样有三种类型：第一种类型的非概率抽样是任意抽样，也被称为方便抽样或偶然抽样；第二种类型的非概率抽样称为判断抽样或有目的抽样，也称为专家选择；第三种类型的

非概率抽样称为配额抽样。

二、总体与样本

1. 总体(population):总体就是进行调查对象的全体。如规模以下工业抽样调查,调查对象为规模以下的非国有工业企业和全部个体工业经营单位,这就构成了一个总体。

总体根据其包括总体单位的数目可分为有限总体和无限总体两种。有限总体是指总体单位能够明确确定,并且其单位的数目是有限的。在社会经济调查中其对象常常是有限总体,如一定时间和空间的企业数。反之,若总体中包括的单位为无限时则称为无限总体。

2. 样本(sample):是总体的一部分,从总体中按一定程序抽得的那部分个体或抽样单位组成。样本中包含的抽样单位数 n 称为样本量(sample size)。样本量 n 对总体总单位数 N 的比称为抽样比(sampling fraction): $f=\frac{n}{N}$。

三、抽样框

抽样框(sampling frame):抽样框是指包含所有抽样单位的名单或名册。如企业名单、整群抽样中的村名单等。但在实践中抽样框与所研究的总体常不一致,为此提出目标总体与抽样总体两个概念。

四、目标总体与抽样总体

1. 目标总体(target population):是真正作为研究对象的全体。

2. 抽样总体(sampling population):是用作抽样的总体,也就是抽样框。

五、调查单位与抽样单位

1. 调查单位(survey unit):抽样调查要通过对样本单位的观察或调查来取得有关数据或记录有关特征,这些单位称为调查单位。

2. 抽样单位(sampling unit):抽样单位是指将总体划分成不重叠的

有限多个部分的每个部分。每个抽样单位都由或多或少的个体组成，当然也可以只包含一个个体。如目录抽样中的每个企业和整群中的每个村都是一个抽样单位。

抽样单位有时与调查单位一致，有时不一致。例如目录抽样中的每个企业既是抽样单位，又是调查单位。而以村委会为群的整群抽样，则每个村委会为抽样单位，村委会中的个体户为调查单位。

六、误差

抽样调查中的误差可以分为两类：一类是非抽样误差，另一类是抽样误差。

1. 非抽样误差(non-sampling error)：是指除抽样误差以外的，由于各种原因而引起的误差，在各种方式的调查中都存在。产生于抽样调查的各个环节，即调查及抽样设计，调查实施与数据采集以及数据的汇总分析与处理等过程中。

2. 抽样误差(sampling error)：是由于用样本估计总体而产生的误差，也叫代表性误差。因为样本只是总体的一部分，用部分样本数据估计总体数据，不可能完全相同。与非抽样误差不同的是，抽样误差是能够计量的，而且可以得到控制。

抽样误差可以用均方误差(mean square error)表示。设 θ 为总体某个参数，$\hat{\theta}$ 是通过样本得出的 θ 的估计量，则均方误差可以表示为：

$$MSE(\hat{\theta})) = E(\hat{\theta}-\theta)^2 = E[\hat{\theta}-E(\hat{\theta})]^2 + [E(\hat{\theta})-\theta]^2$$

式中第一项 $V(\hat{\theta}) = E[\hat{\theta}-E(\hat{\theta})]^2$ 是估计量 $\hat{\theta}$ 的方差(variance)，而第二项 $B^2(\hat{\theta}) = [E(\hat{\theta})-\theta]^2$ 是估计量 $\hat{\theta}$ 的偏差(偏倚)的平方。

当偏差为零时，称 $\hat{\theta}$ 为 θ 的无偏估计量(unbiased estimator)。这时 $\hat{\theta}$ 的方差就等于它的均方差。

方差的平方根 $\sigma(\hat{\theta}) = \sqrt{V(\hat{\theta})}$，称为估计量的标准误(standard error)或标准差(standard deviation)。标准差与无偏估计量之比，称为估计量的变异系数(coefficient of variation)。

七、误差限与置信度

1. 误差限(error limit):估计量的精度通常用误差限来表示。误差限是在某种概率保证程度下的最大绝对误差或相对误差。

2. 可靠性或置信度(confidence level):在抽样调查中,用样本估计总体参数时,由于样本的随机性,其结论总是不确定的,因此采用一种概率的陈述方法,也就是数理统计中的区间估计方法,即估计值与总体参数在一定允许的误差范围以内,其相应的概率称为可靠性或置信度。误差限与置信度是成对出现的,只给出误差限,不给出置信度或只给出置信度不给出误差限都是没有意义的。

3. 置信区间(confidence interval):样本估计量 $\hat{\theta}$ 的分布叫作 $\hat{\theta}$ 的抽样分布(sampling distribution)。当样本量足够大时,这一分布通常近似于正态分布。如果 $\hat{\theta}$ 是无偏的且呈正态分布,则有 $P\%$ 的把握认为,θ 所在的区间为 $[\hat{\theta}-tS(\hat{\theta}),\hat{\theta}+tS(\hat{\theta})]$,$t$ 为概率度,反映极限误差的相对程度,它是确定概率保证程度大小的指标。给定 t 值,就可通过查概率分布表得到相应的概率。如:当概率度 $t=1$ 时,概率保证程度为 68.27%;当概率度 $t=2$ 时,概率保证程度为 95%;当概率度 $t=3$ 时,概率保证程度为 99.73%。$P\%$ 是概率保证程度。$[\hat{\theta}-tS(\hat{\theta}),\hat{\theta}+tS(\hat{\theta})]$ 称为置信区间。

八、抽样效率与设计效果

1. 抽样效率(sampling efficiency):是指两个抽样方案的抽样方差之比。当某个估计量的方差比另一估计量的方差小时,则称方差小的估计量效率比较高。由于方差的大小与样本的容量有直接关系,因此在比较时,通常以样本量相同时的方差进行比较。如果估计量是有偏估计时,那么也要考虑偏差的因素,因此就将两个均方误加以比较。均方误愈小则效率愈高。

在两种抽样设计的均方误相同的情况下,它们的费用比值称作费用效率。如果在均方误不等的情况下,费用效率为一种抽样方案的均

方误与其费用的乘积与另一种抽样方案的均方误与其费用的乘积的比值。

2. 设计效果(design effect,简写为 *deff*):就是把一个设计方案的方差与简单随机抽样的方差进行比较。*deff* 小于 1 时,表示设计方案的效率高于简单随机抽样,反之,效率低于简单随机抽样。设计效果除用来评估比较复杂的抽样方案的效率外,还可利用它来求复杂抽样设计的样本量,因此是一个很有用的指标。

九、三种性质的分布

1. 总体分布(population distribution):是指研究对象这一总体的各个单位标志值的分布状况。统计学上通常用次数分配来反映分布状况,也可用直观图形来表示。如某地区有工业企业 9533 个,其中年产品销售收入 500 万元以上的有 350 个,200 万元至 500 万元的企业 2110 个,100 万元至 200 万元的企业 1996 个,100 万元以下的企业 5077 个。各层中的企业个数就是 9533 个企业的分布状况。

2. 样本分布(sample distribution):若从总体中抽取一个容量为 n 的样本,那么这些样本单位的标志值也同样形成一个分布,由于样本是从总体中抽取的,其中就包括总体的一些信息,所以样本分布也称经验分布。仍以上述地区为例,从 9533 个工业企业中抽取 253 个样本,其中从 500 万元以上的 350 个企业中抽取 81 个,从 200 万元至 500 万元的 2110 个企业中抽取 79 个,从 100 万元至 200 万元的 1996 个企业中抽取 34 个,从 100 万元以下的 5077 个企业中抽取 59 个。这些样本单位也形成了一个分布,就称为样本分布。数理统计中的格列纹科定理证明了:当样本容量足够大时,样本分布将趋于总体分布。所以在抽样调查中常以样本的平均数来估计总体平均数,以样本的方差来估计总体方差。

3. 抽样分布:是指样本估计量的分布。如果按照一定的样本容量,一定的抽样方式反复抽取样本,每一套样本可以计算一个估计值。由于样本是随机抽取的,因此估计值是一个随机变量,它也遵从一定的分布,这种分布叫做抽样分布。例如总体单位数为 N,采用不重复抽样,

样本量为 n，则共有 $\binom{N}{n}$ 个可能的样本，每个样本可得到一个样本的均值 $\bar{y}$，这 $\binom{N}{n}$ 个样本均值也构成一个分布，这就是样本均值的抽样分布。抽样调查主要是根据估计值的抽样分布来对参数进行区间估计。根据中心极限定理，随着样本容量的增大，样本均值等一些统计量均趋于正态分布，因此在大样本的情况下，可用正态分布来作区间估计。

§1.2 几种基本的概率分布

一、正态分布

正态随机变量 X 的概率密度函数的形式如下：

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad -\infty<x<\infty$$

其中，μ 为随机变量 X 的均值，σ^2 为随机变量 X 的方差。

通常对具有均值为 μ，方差为 σ^2 的正态概率分布，记为 $N(\mu,\sigma^2)$。于是有正态随机变量 $X\sim N(\mu,\sigma^2)$。一般来说，正态分布的密度曲线是以 μ 为中心，在 μ 的两侧呈对称的形状，无论参数 μ 和 σ 取何值，密度曲线下所覆盖的面积均等于 1。正态分布曲线下，位于 $\mu\pm\sigma$，$\mu\pm2\sigma$，$\mu\pm3\sigma$ 之间的面积分别约占总面积的 68.26%，95.45% 和 99.73%，如图 1-1 所示。

在正态分布的概率密度函数中，当 $\mu=0,\sigma=1$ 时，我们称随机变量 X 遵从标准正态分布，记为 $X\sim N(0,1)$。关于正态分布的理论已很完善，数学上也易于处理。当总体概率分布为正态分布时，作为从中抽出的样本，其统计量的样本概率分布有 χ^2 分布、t 分布、F 分布。正态分布是一个非常重要的概念。

二、χ^2 分布

如果从标准正态分布 $N(0,1)$ 的总体中得到 n 个随机变量，分别为

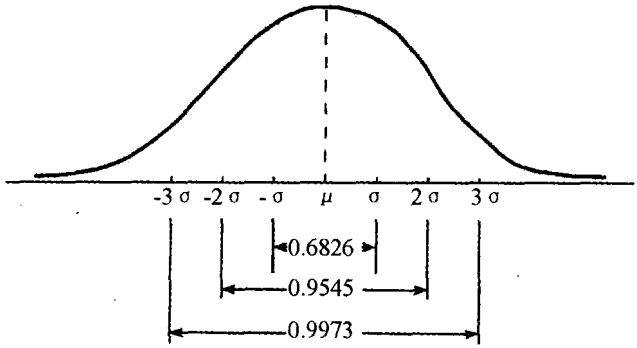

图 1-1

$X_1, X_2, \cdots, X_n$ 时，则由 $\sum X_i^2$ 得到的分布叫做自由度为 n 的 χ^2 分布，记为 $X \sim \chi^2(n)$。

χ^2 分布的数学期望和方差分别为：

$$E(X) = n, D(X) = 2n$$

图 1-2 给出了 n 取不同值的概率密度曲线。

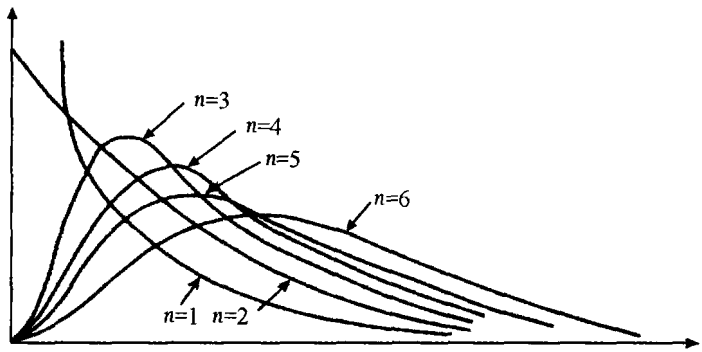

图 1-2

χ^2 分布与 $N(0,1)$ 分布之间有如下关系：

设 $X_1, X_2, \cdots, X_n$ 是相互独立的随机变量，并且 $X_i \sim N(0,1)$，$i = 1, 2, \cdots, n$，则