MBA CLASSICS

MBA精品系列

MBA

BUSINESS STATISTICAL ANALYSIS

商务数据分析与应用

王汉生／编著

中国人民大学出版社

MBA CLASSICS

MBA精品系列

MBA

商务数据分析与应用

王汉生／编著

中国人民大学出版社
·北 京·

图书在版编目（CIP）数据

商务数据分析与应用/王汉生编著. —北京：中国人民大学出版社，2011.9
（MBA精品系列）
ISBN 978-7-300-14346-0

Ⅰ.①商… Ⅱ.①王… Ⅲ.①经济统计-统计方法-研究生-教材 Ⅳ.①F222.1

中国版本图书馆CIP数据核字（2011）第186797号

MBA精品系列
商务数据分析与应用
王汉生 编著
Shangwu Shuju Fenxi yu Yingyong

出版发行	中国人民大学出版社		
社　　址	北京中关村大街31号	**邮政编码**	100080
电　　话	010－62511242（总编室）	010－62511398（质管部）	
	010－82501766（邮购部）	010－62514148（门市部）	
	010－62515195（发行公司）	010－62515275（盗版举报）	
网　　址	http://www.crup.com.cn		
	http://www.ttrnet.com（人大教研网）		
经　　销	新华书店		
印　　刷	涿州市星河印刷有限公司		
规　　格	185 mm×260 mm　16开本	**版　　次**	2011年9月第1版
印　　张	12.75插页2	**印　　次**	2011年9月第1次印刷
字　　数	203 000	**定　　价**	32.00元

前言
PREFACE

我在上课的时候常常和同学们开一个玩笑："上帝是靠什么来记录这个世界的?"作为一个统计学家，我会说："上帝是靠数据来记录这个世界的。"请原谅，这话也许有点夸张，但这是我的职业习惯，也是我的立场。仔细想想这并不是笑话。自然界的风光雨电早就被气象学家忠实记录，从中人们可以了解什么样的事件是"百年一遇"；社会经济活动中的各种商品价格，被几乎所有的国家政府系统采集，这才有了物价指数，从中我们可以判断房价到底是"涨"还是"跌"；医学研究中，科学家通过成百上千的生化指标，刻画一个生命的重要体征，并以此判断药物、医疗设备、治疗方案等是否有效。在过去的几十年里，随着信息技术的高速发展，以前这些传统且昂贵的数据采集方案逐渐被更加经济有效的信息技术代替。在生物信息技术中，以生物芯片为代表的新一代技术手段，使得生物科学家能够同时监控成千上万的基因表达水平，这为新药更加快速的研发提供了可能。

那么，在商务管理实践中，数据的故事如何呢？这有点像一首情诗（100%原创）："我就在你的身边，你却忽略了我的存在，直到有一天，别人开始重视我的时候，你才明白过来。"在和业界的接触过程中，我发现这样的故事比比皆是。例如，大多数超市都有会员卡，忠实记录了消费者的购买行为，但是这些数据发挥应有的作用了吗？很多银行信贷机构，通过尽职调查收集了很多企业的财务信息，这些信息被有效地整合为信用评级打分了吗？各大移动通信运营商拥有极其详细的关于手机使用者的消费信息，甚至可以通过通话记录了解消费者的社会关系网络，如此有用的信息，转化成了客户价值与忠诚度吗？在过去的几年里，SOLOMO（SOcial＋LOcal＋MObile）的创业理念在电子商务领域大受追捧，很多优秀的创业团队应运而生，充满梦想满怀激情，对不起请等一下，年轻的创业家们，你们懂得如何解读来自电子商务的非标准数据吗？例如，文本、网络日志、地理信息等。天，原来我真的就在你的身边，但是你从不重视我的存在，因为你太忙了。什么时候你才会注意到我？当一个强有力的同业竞争对手懂得我的

价值的时候！请问：你要被动等待，还是主动学习？

如果你的决定是主动学习，那么本书是你最好的起点。本书是我在北京大学光华管理学院近十年教学经验的结晶，其结构组织、商务理解、案例收集是我过去近十年中不停思考的结果。相关课程先后给光华管理学院的商务统计学博士硕士研究生、管理学（营销、战略、会计、人力资源、组织行为等）博士硕士研究生，以及 MBA 学员授课总计几十次有余。

那么本书同其他类似教材的最大区别在哪里？这要先讲一个故事。在商学院的统计教学中，优秀的同学常常会问一个非常自然的问题：我为什么要学统计学，学了有什么用处？当第一次被问及该问题的时候，我无力回答。传统的统计学教材（不管中文还是英文教材）所给出的答案都苍白无力，那些玩具案例（toy example）不可能说服我那些优秀的学生，而我自己也从来没有真正地思考过。从本科接触统计学的第一天，到博士毕业，我没有任何实际经验，无力思考。于是，我对自己说：我，要把这个事情搞明白！因此，在过去近十年的教学生涯中，我努力地思考这个问题。最后不是我自己搞明白了，而是那些背景丰富的学生讲出了答案，教会了我。营销的学生（研究生或者 MBA）讲述了多元统计如何帮助他们做市场细分产品定位；战略的学生阐述了回归分析如何帮助他们解读企业的多元化战略以及对外直接投资的选择；组织行为学的学生熟知方差分析可以帮助他们判断组织结构与团队绩效之间的关系。这就是在商学院教统计学的乐趣，你可以从学生那里学到很多，还会得到众多优秀同事的指导，他们的背景有会计、金融、营销、战略、组织行为等，受益无穷！能和那么多优秀的学者共事是天大的幸事！

在这个学习的过程中，我逐渐形成了自己的教学理念。那就是：从课堂的第一分钟开始要告诉学生，统计学可以做什么。用绝对真实亲力亲为的案例告诉大家，统计学不高深，很朴素，但很有用。所以，本书的写作风格独树一帜。每一章的开头都不讲统计学，而是讲一个真实的案例背景，或是营销，或是会计金融，或是人力资源，这些问题都是管理实践中可能出现的典型问题。这些问题最开始的提出似乎都和统计学无关（如客户关系管理），但是最后到执行层面的时候，你会发现没有一套科学系统的数据分析方法不行，由此产生了学习相关统计学方法的原始动力。在此动力的推动下，我再抽丝剥茧一样慢慢地把相关统计学理论铺垫展开，同时伴以案例数据、程序演示（SAS＋R），告诉读者一个完整的数据分析过程。最后，再给出一个分析报告的样例，告诉大家优秀的分析结果应该如何陈述报告。没有最后这临门一脚，你的老板不会满意，你的客户不会满意，你的同事也不会满意。

本书的基本行文结构与模式有以下两个鲜明特点：第一，本书的案例，全部由本人

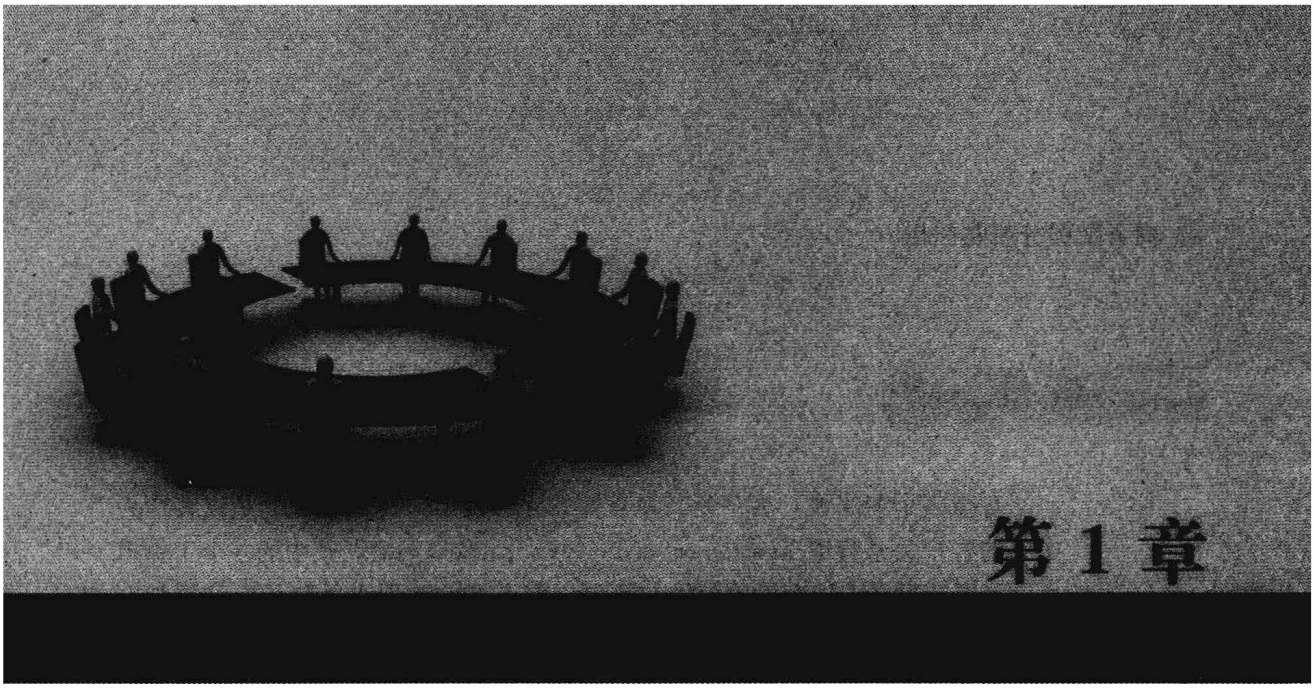
亲力亲为，并且都来自中国市场；第二，本书的统计软件演示以SAS为主，以R为辅。

坦白地说，写书是一件很苦的差事。在动笔之前，我很怀疑自己能否投入那么大的精力去做前期准备，很怀疑自己能否咬牙完成后期的文字。但我最终完成了，这要特别感谢中国人民大学出版社的编辑陈永凤老师，是她的理念与敬业让我理解教材对于教育的重要性，这对我是很大的激励，人大出版社工商管理分社于波社长和黄佳编辑的鼓励和支持也极其重要。在整个教材的准备期间，我得到了很多老师、同事、学生，以及出版社工作人员的大力帮助。需要感激的人太多，难以逐一罗列。愿珍惜最后一点空间，深深地感谢我的父母和岳父母，谢谢他们的生养之恩，谢谢他们在生活上给予我的巨大支持，退休后还要辛苦地照顾我们，我才有时间和精力追逐梦想。时常对比他们的辛苦和自己的回报，惭愧不已。我想深深地感谢我的太太，她是我们家庭最坚实的支柱，因为她的存在，我们这个4+2+1的大家庭紧紧地站在一起，谢谢她给予我的家，给予我的一切！我要深深感谢我的儿子，那个阳光自立快乐飞翔的精灵，带给我无穷的快乐，时常触动我心中最柔软的角落！

最后感谢北京大学光华管理学院！这个培养我十年并为我深爱的学院，教会我太多东西，给我无穷多的成长机会。愿你的明天更加美好！

王汉生
北京大学光华管理学院
hansheng@gsm. pku. edu. cn
http://hansheng. gsm. pku. edu. cn

目录 CONTENTS



第1章

线性回归

——以移动通信网络的客户价值分析为例

1.1 背景介绍

对很多企业来说，客户是其最重要的无形资产。与此相关的是以客户为中心的客户关系管理（customer relationship management），其中的一个工作重点就是要获得并维护有价值的客户资源。但市场的激烈竞争使获取和维护客户资源变得成本高昂。在一些竞争激烈的行业（如零售业），管理者常常感叹自己的利润已经薄如刀片！企业管理者常常面临这样的两难境地：一方面激烈的市场竞争使得企业能够用于客户关系管理的资源非常有限；另一方面，对手的步步紧逼又迫使企业不能减少，甚至增加相关投入。怎么办？

翻开任何一本关于客户关系管理的教材，都会看到很多答案，其中之一就是要发现对自己最有价值客户，然后将有限的营销资源投入到这批客户中。传统的营销智慧告诉我们：往往是20%的最有价值客户，贡献了企业80%的利润。这就是著名的2/8定律。但是在实际操作中，找到那20%最有价值的客户绝非易事。这里涉及一个很具体的问题：客户价值如何测算？如何度量？

对于不同的行业、不同的目的，答案肯定是不一样的。对很多企业来说，客户价值在于他购买了多少产品或者服务，并因此带来了多少利润。这是一个合理的度量。但是，它全面吗？答案是否定的。很多处在创业初期的企业，由于还没有达到盈利规模，每个客户带来的现金利润可能为负，如在线视频、网上购物、社交网站等。那么能说这些客户没有价值吗？如果没有，企业为什么还要拼命保有这些客户呢？显然他们是有价值的，只是他们的价值在当期还没有通过现金利润表现出来。但是，在未来，他们可能很有价值。因此，该客户整个生命周期的价值（life time value），而不是当前价值，才是一个更好的标准。但实际工作中，说清楚一个客户的未来价值并不容易。例如，对于社交类型的网站（如开心网）或者软件（如MSN），客户并不直接从企业购买产品或者服务，因此并不直接产生现金利润。但是，一个客户的客观存在却是企业使用其他手段盈利的基础（如广告、直销）。对这类客户，他们在网站或者软件上的活跃程度也许是一个更好的价值测量标准。

如上所述，不同企业、不同时期、不同目的，都会影响到对客户价值的判断，进而影响相关指标的选取。一旦客户价值指标选定，我们就能够识别

哪些是最有价值的客户。到此结束了吗？不。我们看到的是企业已成过往的历史，我们更关心未来会怎样。过去有价值的客户，未来就一定还有价值吗？过去的低价值客户有可能培养成高价值客户吗？我们能否以历史数据为鉴，预测未来？所有这些问题都需要我们对数据进行详细的分析。

如何分析？统计学是这样一个特殊的学科，它不以数学上的烦琐为荣（所以很多影响深远的统计学论文不涉及过多的数学理论证明），也不炫耀匪夷所思的分析技巧（实践经验表明，越是质朴简单的统计方法，稳定性越好，适用性越高），它是人们在实践中对数字分析很多朴素直觉的规范以及汇总。因此，如何分析客户数据应该由实际目的决定，而不是由统计学教材（包括本书）决定，更不应该由统计学软件决定。那么我们检讨一下：实际应用到底需要什么？上文提到了著名的2/8定律。很多管理者想知道：到底哪 20%的客户最有价值？他们有什么特征？再说得通俗一点，他们“长相”如何？当我们形容一个自然人的长相时，会描述他的高矮胖瘦、肤色、五官等。那么如何刻画一个消费者呢？可以用人口统计特征（demographics），如性别、年龄、职业等；还可以用他的消费行为特征，如历史消费金额、频率等。我们很关心这样一些特征（即人口统计特征、消费行为特征等）如何影响该消费者的价值。在统计学中，这是一个标准的“回归”（regression）问题。什么是回归问题？即因变量（dependent variable）和自变量（independent variable）之间关系的问题。这里我们关注的因变量是客户价值，我们认为它会因为消费者特征的不同而不同。因此，自变量是消费者特征，其中包括人口统计特征、消费行为特征等。因为自变量在统计学意义下解释了因变量的部分行为，因此人们也把自变量叫做解释变量（explanatory variable）、协变量（covariate），或预测变量（predictor）。以上简要回顾的是相关背景知识。接下来，我们将以国内某区域的移动通信运营商为例做进一步展示。

1.2　案例介绍

本案例的数据由国内某移动通信运营商提供。该运营商的主要业务是提供无线通信服务，并收取相应的费用。就像零售行业一样，无线通信行业的竞争也日趋激烈，各大运营商都通过各种手段努力扩大市场份额，提高盈利

能力。在整个市场渐趋饱和的情况下，拉拢竞争对手的客户资源成了最常见的一种手段，而这就造成了客户离网率居高不下的现象。很多低端客户，今天用A公司的服务，明天只要B公司给一点点好处，就用B公司的，再过几天，一看A公司有新的打折促销计划，又改用A公司。在这样一个周而复始的拉锯战中，企业耗尽了有限的营销资源，甚至该客户本身也没有得到什么实质性的好处，因为对客户而言，更换服务商何尝不是一种消耗？因此，本案例提供者开始思考有无其他更好的方式为客户带来真正的价值，并以此提高客户忠诚度。这就有了下面要详细介绍的校园网计划。

具体来说，校园网计划就是一个客户忠诚度的培养计划。该运营商的客户大部分是高校在校生，他们本身就是一批非常优良的客户，有稳定的消费（如话费、短信、彩铃等）；此外，他们对各种新鲜的增值业务也乐于尝试。该运营商的很多新业务都是以高校在校生为切入点的。特别值得关注的是，高校在校生毕业后往往能找到较稳定且收入良好的工作，有潜质成长成为高价值客户。因此，如何深度“套牢”这样一批优质客户是该运营商一直都很关注的问题。按照校园网运营规则，如果一名高校在校生希望加入校园网，他首先必须已经是该运营商的客户，此外，还得由已有校园网用户进行短信邀请。接受邀请，则成为校园网的一员。作为回报，所有的校园网内通话资费都会非常便宜。当然，与网外朋友通话，资费照旧。所以，为了进一步降低自身的资费水平，网内成员有很大的动力邀请朋友加入校园网。而已经加入校园网的成员则发现很难离开，因为大部分朋友及主要的社交网络都还滞留在校园网中，一旦离开，同他们的沟通交流将变得非常昂贵。

那么，运营商的付出与回报又如何呢？首先，天下没有免费的午餐。为了深度“套牢”学生客户，运营商有重大付出，即降低资费。此外，为了迅速扩张，鼓励大家推荐新客户，运营商对推荐者有一定的现金奖励。那么，运营商希望的回报是什么呢？第一，高忠诚度，低离网率。这样可以间接地降低客户的获取以及维护成本。第二，通过资费的下调，刺激消费量的上升，使得总利润不降反升。然而真实情况怎样呢？好消息是离网率确实下降不少，但坏消息是总利润上升并不明显。一线业务员反馈的信息表明，由于缺乏可靠的手段识别被邀请的用户是否真的是高校在校生，很多非高校在校生的低端客户被加进来。这部分客户的总消费量（如通话时长）并没有因为入网而有任何上升；相反，由于资费的下降，他们对公司利润的贡献大幅下降。当然，也有正面案例。有的客户自己入网后，能够进一步吸引一大批优

质客户入网。相比入网前，他们的沟通交流更加密切，因此，尽管单位时长的资费水平下降很多，但是他们对企业的总利润贡献上升不少。

这说明不同的客户影响是不一样的。此外，一线业务员还反馈了一个重要信息，即一名客户作为一个“被推荐者”对校园网贡献大小，除了依赖于自身的消费者特征以外，还极大地依赖于推荐者，即向他发送短信邀请的那个人。这个发现很重要。前面提到，为了在最短的时间内以最快的速度扩张网络，运营商对推荐者有一定的现金奖励。但是，现在看来并非每个人都能推荐有价值的客户，甚至有的推荐者带来的客户对企业的贡献是负的。因此，有必要研究一下，带来低价值客户的推荐者同带来高价值客户的推荐者之间有没有系统性的差异？如果能够在一定程度上认识把握该规律，就可以把有限的现金奖励资源，有针对性地投放到那些能够为企业带来高价值客户的推荐者身上。这样，既能节省企业有限而宝贵的营销资源，还能改善客户结构。因此，我们将详细研究：什么样的推荐者能够带来高（或者低）价值客户？

1.3 指标设计

在实际工作中，问题的定义永远都是模糊笼统的，如本章所关心的问题：什么样的推荐者能够带来高（或者低）价值客户？但是，问题的研究却是具体的。怎样把一个抽象的问题具体化？谁来起到桥梁的作用？那就是指标设计。好的指标设计能够把抽象概念具体化，而且具有直接的管理实践含义。例如，推荐者应该如何描述？客户价值应该如何评估？在本案例中，我们考虑如下指标。

什么样的指标能够刻画推荐者价值？这是我们的因变量。首先，推荐者本身也是校园网的普通消费者。因此，毋庸置疑，他对企业的直接利润贡献是其价值的一个重要组成。为方便起见，简称这部分价值为该客户的“直接价值”。但本案例更关注推荐者通过推荐其他客户所带来的“间接价值”。需要说明的是，我们从不否认研究推荐者直接价值的意义所在，仅仅是出于节省篇幅的考虑，将集中分析间接价值。在现有的营销文献中，研究客户直接价值的数不胜数，但由于关于间接价值的数据资源奇缺，相关研究很少。但这部分价值的重要性广为人知，同客户的口碑（word of mouth）价值紧密相关。

归根到底如何量化一个推荐者的间接价值呢？假如一名推荐者为企业推荐了三名客户，那么，他为企业带来了多少利润呢？这依赖于这三名客户在被推荐前后的行为变化。如果在被推荐加入校园网之前，他们每个月总共贡献利润 100 元，加入校园网后变成了 80 元，那么这就是一个失败的推荐者，他的推荐行为为企业带来的利润相对变化为：(80－100)/100＝－20%。如果在加入校园网后，这三名客户的利润贡献是 120 元，那么推荐者对企业的间接利润贡献为：(120－100)/100＝20%。因此，我们的因变量就是某推荐者所有推荐客户，在加入校园网前后的相对利润变化。

就像所有的研究一样，指标设计永远没有最好，好的实际指标往往都带有明显缺陷，但这不妨碍它的价值所在。我们前面定义的推荐者间接价值度量，也不可避免地带有很多局限性。例如，目前只考虑了入网当月同前一个月的对比。这种对比刻画的是入网当月这个特定的时刻，对未来的借鉴意义尚不清楚。理论上我们不排除这种可能，被推荐的客户入网前每月消费 100 元，入网当月由于新奇消费 120 元，此后好奇感渐趋减少，最终变成每月消费 80 元。在这种情况下，我们就不能只考虑一个变量，而应考虑多个，这样会更加全面一些。从统计学方法论的角度来看，研究一个因变量同分别研究多个因变量没有本质差异，因此，我们在这里只集中讨论一个因变量。

确定了因变量以后，再考虑解释变量。前面提到，对一个自然人的描述要靠高矮胖瘦等指标。但是，对一个推荐者的刻画就得靠具有营销实践意义的消费者特征。在实际工作中，研究者已经积累了大量的有用指标，能够极其详细地刻画一个推荐者的方方面面。例如我们可以考虑消费者的消费行为，主要包括该用户在各项通信及增值业务（如通话、短信、彩铃、上网）上的花费；我们还可以考虑消费者的通话特征，包括该用户的通话时长、通话频率、通话时间（早上、中午、晚上），我们还可以进一步地将通话时长拆分成主叫、被叫、本地、长途、漫游等。总而言之，实际工作中可以考虑的解释变量可以很多，为简单起见，我们只考虑下面几种。同前，解释变量的多少，一般不引起统计学方法论的本质改变。

1. 通话总量（X_1）

我们第一个考虑的解释变量是通话总量，以分钟计。毫无疑问，这是一个很重要的变量，它直接刻画了用户的活跃程度。由于校园网提供非常优惠的通话资费，因此对那些高通话总量的用户有很强的吸引力。因此，具有高

通话总量特征的推荐者也更有可能带来优质的客户。为了分析理解方便，我们对通话总量做了对数变换。

2. 大网占比（X_2）

该变量衡量了在用户的所有通话时长中，有多少发生在该运营商的网内。如果我们将一个人的通话总量看作他的社会关系网络，那么大网占比测算了该推荐者的社会关系网络被现运营商所覆盖的程度。

3. 小网占比（X_3）

该变量是大网占比的有力补充，它衡量的是用户所有发生在大网（即本运营商的网络）内的通话时长中，有多少发生在校园网（即一个更小的网络）内。直观地想，如果一名用户小网占比很高，那么他的主要可被推荐社会关系网络（请注意，大网以外的客户是不能加入小网的）中的绝大部分已经加入了校园网，因此，该用户没有充足的被推荐对象，所以，他能为企业带来的价值应该不大。

以上就是本案例所考虑的三个解释变量。它们显然不全面。由于实际经验的积累，一线工作者往往能够构造出更好的解释变量。在这里我们仅仅以此为例。大家还可以注意到，对指标的设计我们应该是有所预期的。由于现代信息技术的发达，人们很容易就能够获得几十个甚至几百个解释变量。因此，很多分析师也乐得不假思索地把所有的解释变量放入模型中一起分析。但这种做法反映的是不深刻理解管理问题本身，完全依赖统计学软件的一种盲目的统计分析过程，并不值得提倡。因此，解释变量的设计不需要太多，但是要精，要深思熟虑，要对管理实践有指导意义。

1.4 描述分析

在正式的统计分析之前，先做一个详细的描述分析对后面的建模非常有帮助。什么是描述分析？简单地说就是对数据的基本描述，不涉及任何统计推断（inference）。例如，样本均值、方差、最大值、最小值、中位数等都是描述统计量，而各种各样的统计检验以及回归模型就不是描述分析的范畴。描述分析的优点在于它能够被最广泛的受众接受。由于不涉及任何假设

检验，因此只要有良好数学基础的人都能够读懂描述分析（如样本均值）。但是，描述分析也有缺点，即无法给出一个统计推断，同时无法综合考虑众多因素，这是假设检验和回归模型的任务。

本案例的数据文件已经提前准备好，假定存在目录“D:\商务数据分析与应用\案例数据”下的CSV（逗号分隔）文件“第1章.csv”中。可以用Excel打开，简单浏览如下：

	A	B	C	D
1	因变量	通话总量	大网占比	小网占比
2	0.21262	2.822822	0.903759	0.219549
3	0.275616	2.628389	0.971765	0.028235
4	0.168753	2.537819	0.991304	0.223188
5	0.154443	3.201124	0.898678	0.112649
6	0.333799	3.13258	0.846721	0.153279

下面我们想办法将其读入SAS软件中，SAS程序如下：

```
data A0;
    infile "D:\商务数据分析与应用\案例数据\第1章.csv"
        firstobs=2 delimiter=",";
    input Y X1 X2 X3;
run;
```

对比该SAS程序以及前面的Excel表格不难发现，Y代表了因变量，X_1代表通话总量，X_2代表大网占比，X_3代表小网占比。值得一提的是，本书的重点不是讲解SAS的编程细节，因此，我们假设读者对SAS编程已经有一定的基础。如果没有，可以参考市面上任意一本SAS的入门教材。上面的SAS程序将“第1章.csv”读入到SAS的工作环境中，并且命名为A0。在SAS的资源管理器中，可以将该数据集打开，界面如下：

	Y	X1	X2	X3
1	0.21261967	2.822821645	0.903759398	0.219548872
2	0.275615636	2.62838893	0.971764706	0.028235294
3	0.16875259	2.537819095	0.991304348	0.223188406
4	0.154442549	3.201123897	0.898678414	0.112649465
5	0.333799014	3.132579848	0.846720707	0.153279293

我们再通过下面的程序对每个变量的各种描述统计量做统计计算。