我最想要的EXCEL数据分析书

大数据时代，数据分析应该这样做

[日] 西内启（Hiromu Nishiuchi）◎著
马　惠◎译

大数据离我们太远？
是因为你没有找对**分析方法**！

sheet1　sheet2　sheet3　sheet4

不懂统计专业术语？没关系，本书绝对不会涉及！
不会专业统计分析软件？没关系，本书只用Excel一步步演示给你！

附赠最真实的原始数据，供你操练

化学工业出版社

我最想要的 EXCEL 数据分析书

[日] 西内启（Hiromu Nishiuchi）◎著
马 惠◎译

化学工业出版社
·北 京·

图书在版编目（CIP）数据

我最想要的Excel数据分析书 /［日］西内启著；
马惠译. —北京：化学工业出版社，2015.9
ISBN 978-7-122-24756-8

Ⅰ.①我… Ⅱ.①西… ②马… Ⅲ.①表处理软件
Ⅳ.① TP391.13

中国版本图书馆CIP数据核字（2015）第173495号

北京市版权局著作权合同登记号：01-2015-1064

责任编辑：张焕强　　封面设计：尚世视觉
责任校对：陈　静

出版发行：化学工业出版社（北京市东城区青年湖南街13号 邮政编码100011）
印　　装：三河市双峰印刷装订有限公司
710mm×1000mm　1/16　印张13　字数190千字　2015年10月北京第1版第1次印刷

购书咨询：010-64518888（传真：010-64519715）　售后服务：010-64518899
网　　址：http://www.cip.com.cn
凡购买本书，如有缺损质量问题，本社销售中心负责调换。

定　　价：32.00元

前言

数据分析其实是有规律可循的。而令人意外的是，这竟少有人知晓。为何会不为人所知？

我对此的解释是，大学里所教授的统计学或数据挖掘知识与“决定该如何分析实际数据”的技能分属不同维度。

在内行人士的指导下，随着基于现实的实际数据分析经验的丰富积累，自然而然地就能掌握上述数据分析的技能。我自己从事数据分析工作至今，已有十多个年头，这期间恩师的教导、共同研究者或客户所提供的数据与彼此的讨论让我受益匪浅。在这十年间，我每月都能接触到各种课题的各种形式的数据，这对一个年轻的统计工作者来说，可谓出人意料的幸运。

那么，想要掌握数据分析技能的人必须要积累足够的经验才能做到吗？

我的回答是：No。

这是因为，世上存在一种被称为“研究设计”的思考方式。所谓研究设计，是业界前辈所积累构架起的专业知识体系。这一体系中满是可以取为己用的技术窍门，如为了有效进行研究应该如何设定课题、使用何种手法搜集何种数据、应该如何推进科学的进步等。

职场人士没有必要像学者那样发现世界普遍存在的真理，不过，要发现职场这一有限范围内的真实情况，研究设计和数据分析也会成为强有力的武器。即便只是理解了其中很小的一

部分，也可能会提高商业的收益性。

近来，在职场中进行数据分析的人有所增加。然而，如果不能很好地理解研究设计的内涵，你只能漫无目的地——如“（从自己的感觉出发）进行假设”，再如“（参考相关人士的经验）听取假说”——确定分析方针。

数据分析是一种比利用经验或直觉更为准确、更为迅速地作出判断的技能。不过，如果说只能通过经验或直觉来掌握数据分析技能的话，那只能是一种讽刺。

本书的目的在于，让任何人都能够在尽可能短的时间内拥有部分数据分析的能力。为此，本书以大多数人都会用的 Excel 作为分析工具。除此之外，我们还准备了四组基于现实状况设定的样本数据。通过对这些数据的实际分析，相信大家会有意外的发现。在每一个事例中，我们也特意设定了一些难以处理的问题。这些问题很少会出现在统计学的教科书中，而现实中却很常见。

另外，本书注重内容的简单易懂，旨在让初学者能够轻松掌握，故尽可能省去了统计学上的专业介绍。虽然在各案例学习中也会对分析结果的解读方法进行说明，但是在此并不会对分析结果的计算过程多做赘述。

当然，本书提到的只是一种方法，相较之下更为正确的分析方法，或能够进行更高精确度预测的方法比比皆是。不过，在此，笔者只希望读者能够在本书的帮助下实际体验数据分析，感受到发现带来的乐趣。你会发现，有了这种体验之后，那些看了几页就被扔到一边的统计学入门书也变得容易很多。最后，非常希望这本书能够成为众多数据分析工作者更上一层楼的好帮手。

西内启

目 录

第1章
用数据分析解决问题的基本思路

本章将会介绍“迅速”进行“有价值的分析”的思考过程。若盲目着手数据分析，很可能要面对费心费力却无甚收获的结果。所以，我们需要清晰掌握分析的重要体系，以便进行“有价值的发现”，进而拿出成果。

数据分析的正确方法：
要想分析有价值，需要注意这几点

提前做假设会遇到陷阱

某餐饮店会议室。面对堆满数字的工作表，数名职员胳膊抱在胸前面露难色。

“近三个月来，咱们店夜间的销售业绩总是上不去。大家有什么解决的对策？”

“还是广告的影响比较大吧。我觉得，不断在电视或杂志上做广告应该会取得不错的效果。通过广告提高店的知名度，客人应该会增加吧。”

“也许吧。不过，广告真的有效吗？”

“刚巧最近进行的顾客调查问卷结果出来了。这个问卷就咱们的广告形象以及光顾的频率也进行了调查。”

“是吗？那就研究一下广告形象与光顾的频率是否有关联性吧。”

……

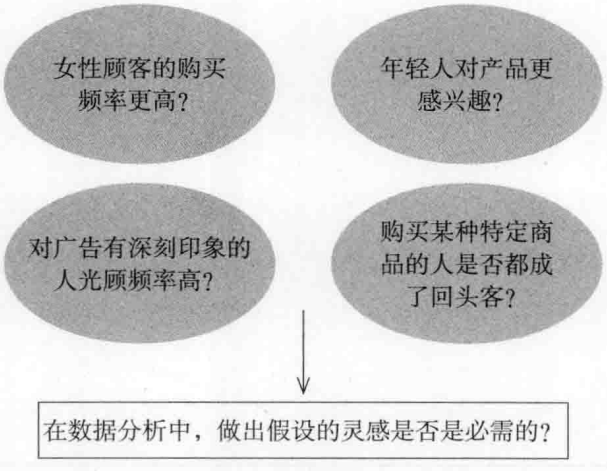

图 1–1　常见的假设

做出假设并对此进行验证。你所在的公司，是否也是如此呢？

公司里甚至可能还会有这样一种人，他们所做出的假设频频中的，被人看作极有分析感觉。众多统计学教科书中也写道：“做出假设，验证假设。”因此，应该有很多人相信那就是正确的做法。

如果回到十年前，这些或许是对的。在没有可用于分析的数据，调查成本也极大的情况下，不分主次胡乱收集数据并非良策。首先做出假设，在此基础上收

集必要数据进行分析是比较合理的。

然而，由于科学技术的进步，“依靠感觉做出假设，再予以验证”的做法已经无法适用于现代商业。通过互联网，我们甚至能够以数百、数千人为单位进行迅速、廉价的市场调查，而且，现在几乎所有的企业都使用电脑记录库存及销售情况。即便是廉价的电脑，也可以使用 Excel 处理数以万计的数据。

我们手中掌握着庞大的数据。只是不知道该如何使用罢了。在之前的时代，做出假设进而予以分析也可以说是“只看到了大数据的冰山一角”。

Point

不要一开始就做出假设，这会让你的视野变窄。

做开放式提问（open question）

做出假设进而验证的方式会让我们漏掉意想不到的重要发现。

比如，你的部下这段期间的业绩很差。你先入为主地认为“他业绩差是因为没有干劲”，要对此进行验证。

“你业绩差是因为没有干劲吗？”

如果你这样去询问部下，是不可能获得有用信息的：若对方做出肯定的回答，你会觉得“果不其然”；即便是对方否认，你也会觉得他在说谎吧。

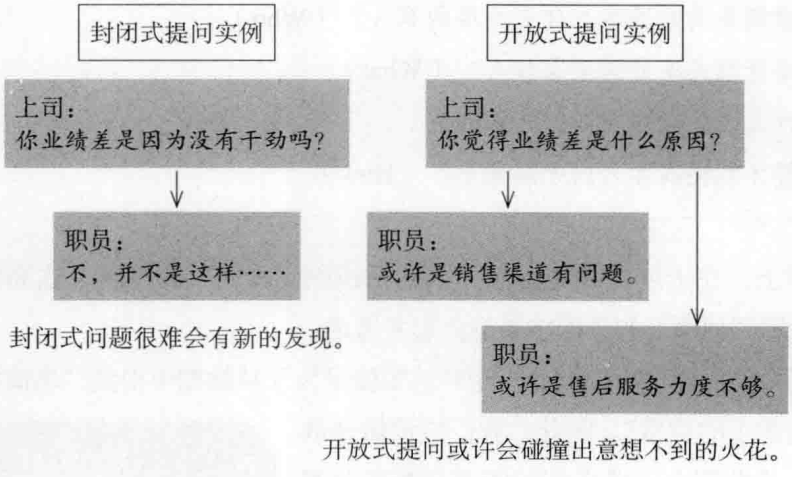

图 1–2　开放式问题

能够用“是”或“不是”来回答的问题称为“封闭式问题”，这经常会让提问者不由自主地将自己的想法强加给对方。

那么，怎样才能在不把自己的想法强加于人的前提下，从对方身上得到想要的信息呢?

只要做开放式提问即可。要让对方主动思考 5W1H，即“何时”(When)、“何地”(Where)、“何人”(Who)、“何事”(What)、“为何”(Why) 和“如何”(How)。

Point

能够用“是”“否”来回答的封闭式问题不可能引出有价值的发现。

在上述事例中，提出的问题——“你觉得业绩差是什么原因”就是一个开放式问题。或许可以借此发现能够解决的某些原因，而非干劲、能力等谁都可以想到的抽象原因。

Point

在提问时，要时刻将 5W1H 的开放式问题放在心上。

数据分析也几乎如此。

“对广告有深刻印象的人光顾频率高 ?”这是个能够以“是”或“不是”回答的封闭式问题，不过，我们可以将它改为开放式问题。留下后半部分的“光顾频率高”，将前半部分转换为 5W1H，就得出了下面的开放式问题。

- 光顾频率变高是哪个时期？（When）
- 光顾频率高的店铺位于哪里？（Where）
- 光顾频率高的顾客是什么类型的客人？（Who）
- 提高光顾频率的因素是什么？（What）
- 为什么光顾频率会高？（Why）
- 怎样才能提高客人的光顾频率？（How）

实际上，在上述问题中，最困难的是找出“为何”的答案，这需要将“何人”“何处”“何事”的分析结果综合起来考虑。

我们为什么要进行数据分析呢？其实就是为了从数据中得到“事前根本想象不到的商业上的启发”。因此，我们应该做开放式提问，将数据中能够得到的所有信息作为答案的候选。这样一来，就很有可能获得有价值的发现。

Point

数据分析的意义在于获得意想不到的、与直觉相悖的发现。

反言之，真正有价值的发现，多是那些结

果出来前谁也料想不到或者是跟直觉相悖的内容。而如果在会议室里提出跟大家直觉相悖的假设，则恐怕很难被认为是“有分析感觉”。

关联性并非答案

随着电脑的迅速升级换代，进行数据分析所耗费的精力及演算时间基本上不再成为问题。

只要利用恰当的工具，就算要从 10 万份包含 100 个项目的顾客数据中调查得出“光顾频率高的顾客是什么类型的客人”的结果，也不过是一瞬间的事。

不过，并不是说不顾章法地总体评估数据就够了。这种做法并不一定能够找出我们真正想要知道的东西。

20 世纪 90 年代，数据挖掘（data mining）这一名词盛行，有人提出了“考虑假设的并非人类，计算机自会为人类发现一切”的主张。这无疑是主张事无巨细对各项目进行调查的方法论。数据挖掘将有悖于人类直觉的各种关联性明确表现了出来。例如，通过数据挖掘，我们也许会发现“周末超市的纸尿裤购买率与啤酒的购买率有很高的关联性”。

那么，这一分析结果是否有价值呢？

当然，纸尿裤与啤酒，这两样商品的销售量增加，多少会带来一些收益。不过，所得收益是否与用于分析的工具或人员成本投入相称，就是另一个问题了。数学层面上的关联性高，并不能保证其在商业层面上有价值，或者分析结果为人所乐见。

Point

即便数据的关联性很高，也不一定是有价值的发现。

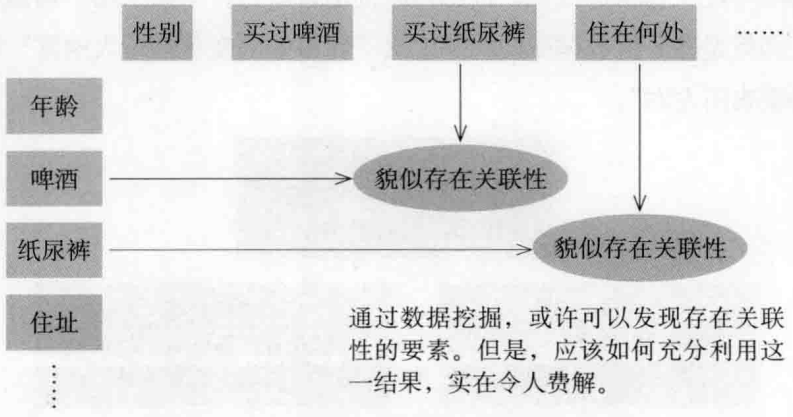

图 1–3　即便存在关联性，也不能保证有价值

确定分析方针的三要点

有人也许觉得纠结：不能事先做出假设，而总体评估数据也得不出有意义的结果，那到底该怎么做才好呢？

没有必要为此烦恼。若计算机没办法思考分析结果的价值，那么你可以从这个角度整理数据及确定分析方针："何事如何变化会是我们所期望的？"

本书中，将采用非常简单的体系来确定分析方针。其中的要点即下述三个：

- 输出结果；
- 分析单位；
- 解释变量。

确定好这三个要点，接下来只要遵循常规操作就能够做有价值的数据分析了。

虽然这里突然出现了大家不甚熟悉的用词，但是理解它们并不需要具备专业知识。

"输出结果"，是指通过数据分析得到的最让人高兴的变量。何事如何变化是众望所归的——也可以看作是成果指标。

"分析单位"是构成上述输出结果的单位。如"合乎期望"的顾客、商品、员工等，分析单位是进行数据比较的基础单位。

"解释变量"可以让分析单位出现不同的特征。假设分析单位是"顾客"，那么左右年龄、性别、之前是否来过等"期望值"的要素是什么？这个"什么"就是解释变量。

按照顺序逐一考虑这三点，谁都可以提出分析方针。如果用一句话来概括这一过程，那就是在分析数据前要明确定义"何事如何变化会皆大欢喜"和"这到底被何种要素所左右"。

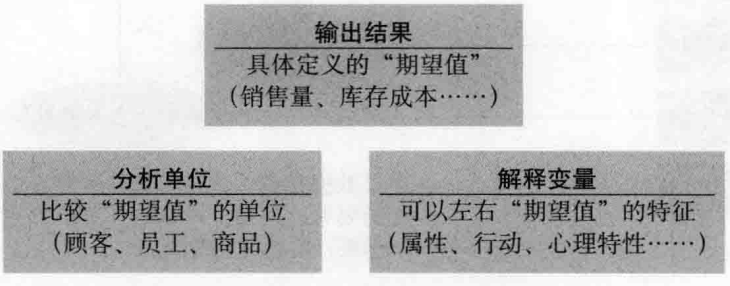

图 1-4　确定分析方针的要点

小结

- 如果在起初就做出假设，会与有悖于直觉的发现、无法想象的发现失之交臂。
- 对庞大的数据进行总体评估，就算发现某些关联性，也很难充分利用。
- 确定了“输出结果”“分析单位”“解释变量”这三要点，自然而然会定下分析方针。

实战演练

- 在你所负责的业务领域，何事如何变化会是你所期望看到的？试着写几条期望达到的状态。

确定直接与利润挂钩的要素：输出结果

要分析数据，首先应该确定的就是“输出结果”。从分析输出结果得到什么结论会让你高兴？销售额的提高？还是成本的削减？若一开始就确定好输出结果，分析的整体方针也会随之确定。

“输出结果”是便于做出快速判断的成果指标

通过数据分析得到的最让人高兴的变量——仿照医学和决策科学领域的表述，我称之为“输出结果”。这里的意思是“成果”。

所谓输出结果，就是一种指标，即治疗方法或决策所应追求的“终极目标”，你达成了多少。在医学领域，定义为输出结果的多是“死亡率”“发病率”等。例如，“怎样做才能降低某疾病引起的死亡率”会成为研究目的。

在商业领域，目的是提高利润。因此，直接与利润挂钩的指标即输出结果。

> **Point**
>
> 直接与利润挂钩的指标即输出结果。

相较于一开始就做出假设或总体评估数据，先定下输出结果是很有益处的。

假设有这样一组数据，包含 100 个项目。针对所有项目，要讨论其全部关联性，就必须对 $100 \times 99 = 9900$ 个指标逐一确认。若是超市的数据，或许会得出“纸尿裤的购买率与啤酒的购买率”或“性别与光顾频率”的关联性高的结论，但列举出这些数量庞大的指标，到最后也不过是一片混乱，不知该如何是好。

如若确定一个假设——例如“广告形象与光顾频率”，那就会与或许掩藏在剩下的 9899 个指标中的有用想法失之交臂。

这时候，就需要输出结果。

设定一个输出结果后，需要关注的关联性指标会骤减。如果有 100 个项目，必须关注的指标数量最多也只有 99 个。如果仅评估关联性强的指标，那最多不过十几个。一旦清楚了什么影响着输出结果，那自然会明白应该做些什

> **Point**
>
> 第一步设定输出结果，可以减少必须要考虑其关联性的要素。

么以增加利润。

恰到好处地找到增加利润的线索，迅速做出判断。为此，我们需要设定输出结果。

应该将什么设定为输出结果

那么，我们应该将什么设定为输出结果呢?

当然，如果数据中包含“利润”这一项，只须直接将其设定为输出结果即可。不过，大多数情况下，数据中不会包含此项。就算在顾客问卷调查中列出“您为敝公司带来了多少利润”，估计也不会有人做出回答。

那什么是利润? 就是“销售额”与“成本”的差。只须将体现销售额或成本的项目设定为输出结果即可。

如果数据中不存在体现销售额的项目，那么可以将购买频率或光顾频率等“与销售额直接相关的项目”设为输出结果。另外，将员工数或滞销库存数等“与成本直接相关的项目”设为输出结果也是一种选择。

> **Point**
>
> 与利润直接挂钩的销售额或成本是输出结果的候选。

面对一组数据，将其中哪一项设定为输出结果并非只有一个正确答案，不过至少要遵循两个基准。

基准一：在可能的范围内，其最大化或最小化的确与利润直接相关。

基准二：使其发生变化并不困难。

顾客满意度能定义为输出结果吗

首先，我们来论证下第一个基准所提到的“在可能的范围内，其最大化或最小化的确与利润直接相关”。

在顾客问卷调查中，多数情况下会含有“顾客满意度”或“广告好感度”等项目。那么，这类项目是否能够设定为输出结果？

让顾客满意、希望广告可以给顾客留下好印象，这些本身都并非商业行为的终点。通过提高顾客满意度或广告好感度，吸引顾客在更长的一段时间内购买更多的商品或服务才是重点。

当然，入手的数据中，“顾客满意度”或“广告好感度”跟利润最为息息相