Think Stats： 用Python数据分析

前言

本书介绍了跟数据分析相关的工具。参照平时我分析数据时的流程，组织本书内容：

* 数据导入与清理：无论数据格式怎样，通常都需要花时间和精力来读入数据，清洗和改变格式，check that everything made it through the translation process intact.
* 单变量探研：通常， 我会先集中研究单个变量，确定这些变量背后的意义，观察这些值的分布，最终确定合适的汇总性统计。
* 逐对探研：为了探测出变量之间可能的关联，我会观察相关表格和单点图，计算相关性和线性拟合。
* 多元分析： 如果发现明显的关联关系，我会使用多回归来添加控制变量，并审查更复杂的关联关系。
* 评估和假设性测试：在做统计结果报告时，回答这些问题是很重要的：效果有多大？如果再重新测量的话，会有多大的可复再性？那些明显的效果改变的可能性有多大？
* 数据可视化：在数据探研过程中，数据可视化可以很好地帮助我们找到可能的关联和效果。最终，结果为了经受住考验，数据可视化是一个有效的方式来跟人交流结果。

本书采用了程序模式（而非数学模式），相对于数学模式，程序模式有如下好处：

* 用Python代码，而不是数据公式来表达想法。Python代码更有可读性，同时，由于Python可直接执行，读者可下载执行并修改它。
* 每一章都包含了练习题，读者可下载下来体验并借以加固对学习的理解。一般来说，你写代码时，你就是在通过代码来表达理解；而调试代码时， 你又是在纠正理解。
* 一些练习题包含了试验来测试统计学特性。例如，你可以通过生成随机样本并计算他们和的方式来探研CLT。最终的可视化结果演示了为什么CLT是可行的，什么时候又失效。
* 用数学方式不容易理解一些想法时，使用模拟会更容易理解。例如，我们可以运行随机模拟来近似地估算出P值，这样可以加固对P值的理解。
* 由于本书使用了Python这个通用的编程语言，读者可以从几乎任何数据源导入数据。这些数据不再局限于那些已经清理且为特定统计工具而格式化好的数据。

本书以工程方式展开。在我的课上，学生们会参与到一个学期长度的工程中来，这个工程需要学生来提出统计方面的问题，寻找相关数据来解决这个问题，解决过程中使用学到的技术。

为了演示我解决统计分析问题的思路，贯穿本书所有章节的有一个案例分析。这个案例使用了下面两个来源的数据：

* 美国家庭增长调查（NSFG），这个调查是由美国疾病控制和预防中心实施的，旨在收集这些数据：美国家庭生活、婚姻与离婚、怀孕、不孕情况、避孕情况和男女性健康。
* BRFSS（）， 这个调查由执行， 旨在。

其它的例子使用的数据有IRS、美国统计局和Boston Marathon。

本书的第二版包含了第一版中的所有章节，其中不少章节是重写的，新写了些关于回归、时间序列分析、生存分析和分析方法。第一版本时没有使用Pandas,SciPy或StatsModels，这样所有的这些材料都是新的。

* 1. 怎么写本书

人们在写一新教科书时， 通常会先读一大堆旧的书。 这样的结果，大多数书都有差不多的内容，又是几乎一样的顺序。

我并没有这样做。实际上， 在写本书时，我几乎没有用到打印版的材料，有这样的几个原因：

* 我的目标是用新的方式来处理这些材料，所以也没多大的兴趣在已经存在的方式上。
* 由于我想通过一种免费的方式发布此书，我也想确保此书中任何部分都不受版权的限制。
* 本书的不少读者并不能方便地从图书馆借到书，所以我尽力使用网上可得到的材料作为引用。
* 一些旧媒体的支持者认为过多地使用电子材料是懒惰和不可靠的。在懒惰方面，他们是对的，不过在第二个问题上，我认为是错误的，所以我想测试下我的理论。

本书写作过程中， 我使用最多的是Wikipedia.一般来说，Wiki上读到的统计材料是很好的（尽管写作的过程中我还是做了些修改）。本书里也包含进了不少Wiki页面引用，我也鼓励大家来看这些Wiki材料.大多数情况下，Wiki页面的内容填补了我描述的空白。词汇和标记也保持着跟Wiki的一致，除非我有足够的理由来修改。其它一些有用的材料是Wolfram MatchWorld和Reddit统计论坛里的内容。

* 1. 使用代码

本书使用的代码和数据都可以在<https://github.com/AllenDowney/ThinkStats2>里找到，Git是一个版本控制软件，使用它可以跟踪项目中的文件。Git控制的文件集合称为仓库。GitHub是提供了Git仓库和方便的操作界面。

我仓库的GitHub主页里提供了多种方式来使用代码：

* 通过Fork方式来创建一个Copy。
* 直接Clone我的的仓库，不过，这样你不能再修改代码后上传。
* 不使用Git情况下，可直接从GitHub页面里下载Zip包。

所有的代码都是用Python2和Python3写的，不需要再做转换。

写本书的过程中，我使用了Anaconda这个工具（包含了pandas、NumPy、SciPy、StatsModels和matplotlib）。Anaconda可方便安装，默认情况下，安装时，是用户级的，而非系统级的，这样你不再需要管理员权限。它都支持Python2和Python3。

本书中使用的Python包，虽然是通用的，但并没包含在标准的Python安装文件里，在一些环境中也不好安装。如果你安装这些包有问题时， 我强烈推荐你使用Anaconda或其它一些已经包含这些工具的安装包。

在clone了Git仓库后（或解压了Zip包后），你将看到ThinkStats2这个文件夹下有一个叫nsfg.py的文件。运行此文件，它会读一个数据文件，执行些测试，打印些消息，如“All tests passed”。如果执行中有什么错，很有可能意味着还有些Python包需要安装。

大多数练习是使用Python脚本的，不过有些也使用IPython notebook。如果以前没有使用过IPython的话，可参考<http://ipython.org/ipython-doc/stable/notebook/notebook.html>文档。

写本书时一个假定：读者对Python和OOP已经熟悉，对pandas、NumPy和SciPy没有要求。如果你对这三个已经很熟悉了，可以跳过其中的一些章节。

数学方面，我也假定读者知道些基本的数学常识，如对数和求和。期间我会提到一些微积分的概念，不过你不需要微积分有了解。

如果你没有学习过统计学，我认为此书是一个不错的开始。

1. 探研数据分析

本书的命题：武装了可用方法的数据可以解答问题并帮助在不确定情况下做决定。

作为一个例子，列举一个案例分析。妻子和我在等待第一个孩子时，听说一般第一个孩子会延迟一段时间出生，从统计上来说，对否？这个案例就是这个问题引起的。

如果搜索下的话，会发现不少这方面的讨论。一些人认为这个说法是对的，一些人认为不靠谱，也有一些有完全相反的说法：第一个孩子一般会早生。

一些讨论中，人们会提供些数据来证明他们的观点。我发现下面这些比较典型：

“我两个朋友最近刚生了第一个孩子，他们都是有两周的延迟。”

“我第一个孩子比预产期晚了两周，现在我认为第二个孩子要早出生两周了！！”

“我不认为这个说法是对的，因为我姐姐是妈妈的第一个孩子，她还早出生了，不少表兄妹都这样。”

这样的报道称为轶事证明，由于他们都是基于些未经公开发布和个人的数据。一般来说， 这样的轶事也无可厚非，所以我也不想指责有这样说法的人。

不过， 我们还是想更有说服力和可靠的答案。这些轶事证明大多是失败的，因为：

* 观察的数据量太小:对第一个孩子来说，如果怀孕周期长的话，the difference is probably small compared to natural variation。
* 选择性偏见： 参与讨论的人们可能会因为他们的第一个孩子出生而感兴趣。这样情况下，选择的数据会把结果带偏。
* 确认的偏见：相信这个说法的人会贡献证据来证明，反对的人会提供反对的证据。
* 不准确：轶事通常是个人讲述的， 经常会记错、表述错误并不准确地重复等。

这样，我们怎么才能做的更好呢？

* 1. 统计学方式

为了纠正轶事的不限制，我们使用统计工具，包括：

* 数据收集： 我们将使用全美国范围收集的数据，这些数据当时就是为了有效地统计美国人口的。
* 描述性统计：我们将生成统计结果， 它将简明地给出数据摘要，使用（evaluate）不同的方法来可视化数据。
* 探研式数据分析:我们将找寻模式，区别和其它的特性（针对我们感兴趣的问题）。同时，我们将校验数据的一致性并找到不足。
* 估算：我们将使用样式数据来评估出些整体特征。
* 假设测试：在发现有明显效果的地方（如两组有不同时）,我们会评估这个效果是不是偶然造成的。

经过上面步骤后，我们可以得出一个可信且极有可能是正确的结论。

* 1. NSFG

从1973年开始，美国的CDC开始执行NSFG调查，旨在收集“这些方面的信息：家庭生活、结婚和离婚、怀孕及不孕情况、避孕使用、男女性健康”。收到的数据为了制定健康服务部署和健康教育计划。

我们将使用这些数据来调查是否第一个孩子出生会晚些和其它问题。为了有效地理解这些数据，我们需要理解此调查的构思。

NSFG是一个跨Sectional的调研，也即在时间的某个点上本调研取一个组织的快照。跟跨Sectional相对的是纵向的调研，这种调研方式会在一定时间内重复地观察一个组织。

NSFG已经进行了7次了，它每一次执行称为一个Cycle。我们将使用Cycle6的数据，这个数据是从2002年1月到2003年3月。

调研的目标是针对人口情况得出统计结论，NSFG调研的目标人口是美国15到44年龄段的人。理想情况下，调查应该收集人口中每一个成员的数据，不过这个很不现实。相反，我们从一个称为样本的人口子集里收集数据。参与调查的人口称为调查对象。

通常情况下，跨Sectional的调研应该是representative，意味着：每一个人口对象都有平等的机会参与。这个目标也很达成，不过， 还是要尽可能地接近。

NSFG并不是representative，相反它估计设计成oversampled（采样过密）。调研的设计者募集了三组（西班牙裔、非洲裔和十几岁的青年人），比例要其比整个美国人口比例高， 以便能推导出足够的统计结论。

当然，采样过于密集的一个问题是，用这个统计结果可能不容易得到更为普遍的结论。后续我们会再讨论这个不足。

在处理这些数据时， 很有必要熟悉下codebook，这个文档描述了NSFG调研的设计，调研的问题和受访的反馈。相当材料请见http://www.cdc.gov/nchs/nsfg/nsfg\_cycle6.htm

* 1. 导入数据

代码和使用的数据可以从<https://github.com/AllenDowney/ThinkStats2>得到。

下载后， 可以找到这个文件ThinkStats2/code/nsfg.py。运行一下，它会读一个数据文件，执行一些测试，并打印下这样的Message：“All tests passed”。

我们看下，都做了什么。NSFG Cysle6的怀孕情况数据在一个叫2002FemPreg.dat.gz的文件中。它是一个gzip压缩的文本文件，有固定宽度的列。文件中每行是一个记录，它含有怀孕情况的数据。

文件的格式记录在2002FemPreg.dct中， 这是一个Stata字典文件。Stata是一个统计软件系统，本语义环境中的字典文件是一个列表，每一项都有变量名字、变量类型和索引（这个索引会告诉我们从哪一行中找到每一个变量）

1. 数据分布

2.1 直方图显示