统计学原理(Statistic)

胡华平

西北农林科技大学

经济管理学院数量经济教研室

huhuaping01@hotmail.com

2022-03-26

第二章 数据收集、整理和清洗

2.1 数据目标

2.2 数据收集

2.3 资料整理和数据清洗

2.4 数据的数据库化

2.5 数据质量

2.6 抽样设计

2.7 抽样分布和抽样误差

2.8 问卷设计技术

2.2 数据收集

数据的来源

数据的载体

数据的状态

收集二手数据

收集调查数据

收集实验数据

数据的来源

不同研究方法会产生不同类型数据：

- 观察数据
- 调查数据
- 实验数据

数据的来源

从产生数据的方式方法上又可以有：

- 问卷数据
- 访谈数据
- 文献数据
- 痕迹数据：大数据。（注意不是**痕迹证据**！）

在获得数据的同时，应该还有一份数据，是记录数据获得过程的，通常称之为日志，它要记录数据是从哪里来的、什么情况下得到的、数据的基本特征又是什么，比如文字数据有多少页、图片数据有多少张，这就是日志数据

数据的载体

从是否数字化来看：

- 数字化的数据
- 非数字化的数据

从是否数值化来看：

- 数值数据
- 非数值数据

数据的载体

从具体形态来看：

- 文本数据：
  - 访问、观察中的文字记录
  - 数字化的字符形态的数据
  - 任何文字加载体的数据，比如文字加载于纸张、羊皮卷等
- 图片数据：
  - 访谈时拍的照片、搜集到的图片、照片的底片等等
  - 数字化为像素点形态的图片数据
  - 任何图形加载体的数据，比如图形加载于纸张、胶片、计算机存储等

数据的载体

- 音频数据：
  - 访问录音、观察中的语音日志、搜集到的音频记录等。
  - 数字化为波形形态的音频数据。
  - 任何音频加上载体，比如音频加载于钢丝、胶片、磁带、光碟、磁碟、闪存盘、硬盘等
- 视频数据：
  - 访谈时的全程录像、搜集到的各种各样视频。
  - 数字化为像素点加上波形形态的视频数据
  - 视频加上载体，比如比如视频加载于胶片、光碟、闪存盘、硬盘等
- 实物数据
  - 任何有实物才可以完整保存信息的实物载体数据
  - 访谈中搜集到的实物、观察中观察到的实物，比如出土文物、建筑等

（提问）课堂思考

以上关于数据来源与形式的分类是完全是互斥的吗？

以调查问卷为例：

- 传统纸版问卷，主要是文字、图片形态的数据。
- 新媒体电子问卷，不管是哪一个类型的电子问卷，主要是数据形态的数据，当然也会有图片的、音频的、视频的数据。

以上的分类并不完全是互斥的，只是根据显性的特征来做一些划分，其实我们很难找到一个标准把数据的形态类型区分得非常清楚。

数字与数值是一个意思吗？

图片、音频、视频看起来的确是数字的，但数字不等于数值！

- 传统照片不是数字的。
- 数码照片的数字指的是像素点的数字
- 音频、视频是同样的道理。

（提问）课堂思考

> “老师，不管什么时候我都要用计算机做笔记的。”

信息化时代，传统手写记录的文本数据是不是越来越没有价值？

- 用计算机或各类终端设备来做电子化记录。
- 用笔和本子做传统记录。

数据的状态

> “老师，我要做一个研究”
> “你的数据从哪里来？”

根据数据是否能够直接用于研究分析，数据的状态可以分为：

原始数据：一般不能直接用于研究。

研究数据：是处理为结构化的、有变量、数值、变量、属性标签的数据。

数据的状态

根据研究数据的持续性，数据的状态有：

1.已经存在的数据。公开数据、正式出版数据、发布的数据，都可以直接使用。

- 政府各类统计数据。包括经济、就业、人口、健康、教育、产业等等数据。
- 上市公司公开数据。根据相关法律，公司的财务数据、生产数据应该公开。
- 研究机构或者研究者个人公开的数据。

2.将要产生的数据。是系统采集的、不断在推进补充的数据。

数据的状态

根据研究数据是否由研究者本人产生，数据的状态可分为：

一手数据：是指自己调查获取的数据。自己调查数据是一个不得已的选择，对任何研究者而言，都应该是第二选择而不是第一选择。

二手数据：是指已经被使用过的数据，拿来再做分析。如果你的研究能够使用已经存在的数据，尤其是很多人用过的数据，那么最好用这样的数据(为什么呢？)。

- 数据的可靠性已经被检验过
- 研究的成果具有可比性
- 通过调查来获取数据，需要专门的能力，包括组织能力、获取数据的能力、评估数据质量的能力、有效运用数据的能力，还需要一定要有资源。

数据的状态

研究数据的**获取权限**一般有如下情形：

- 无需授权就可以使用的数据。正式出版物提供的数据只需要在使用说明中正式说明出处，就不需要授权。
- 需要申请授权的、公开的数据。大多数的学术研究数据，如果你要使用，是需要申请并且被授权。
- 需要通过授权的、未公开的数据。行为痕迹管理机构的数据，包括政府数据、赢利和非赢利服务机构的数据，都属于这类数据。
  - 政府数据：几乎任何一笔收入，都是经过机构管理的，都有痕迹数据。
  - 银行数据：每个人都有银行账号，只要是经过银行卡的，都会留下数据。
  - 电信数据：只要是通过网络通信的数据，都会留下数据记录。

> “老师，他们保存多久呀？”

二手数据收集：搜索引擎工具集

搜索引擎：

- 谷歌搜索（需VPN）
- 谷歌学术（需VPN）
- 谷歌图书（需VPN）
- 必应搜索（可直接访问）

electric cars statistic da

https://www.google.co.jp/webhp?sourceid=chrome-instant&rlz=1C1AVNE_enCN657&ion=1&espv=2&ie=UTF-8#q=electric%20cars%20statistic%20data%20%2B%20US

Bookmarks 业余学习库 农业专业库 社科专业库 经济思想史 课题网站 每日文献 计量经济学 搜索引擎库 大学 西北农林科技大学 经济管理学院 从 Firefox 导入 Google 火狐官方站点 其他书签

Google

electric cars statistic data + US

Gmail 图片

全部 图片 新闻 视频 购物 更多 搜索工具

已启用安全搜索

找到约 5,930,000 条结果 （用时 0.49 秒）

Electric Cars and Electric Mobility - Statistics & Facts | Statista

https://www.statista.com › ... › Vehicles & Road Traffic ▾ 翻译此页

2016年11月4日 - Statistics and facts about electric mobility ... Best-selling all-electric cars in the U.S., based on sales in units ... Miscellaneous, Values, Statistic.

Worldwide number of electric vehicles 2016 | Statistic

https://www.statista.com › ... › Vehicles & Road Traffic ▾ 翻译此页

This statistic shows the number of electric vehicles in the world 2012-2016. ... With Statista you get straight to the point: analyzing data, rather than searching for it. ... Electric vehicles: range of selected vehicles on the U.S. market 2015.

大宝！谷歌搜索

Electric vehicle market statistics 2016 - How many electric cars in UK ?

www.nextgreencar.com/electric-cars/statistics/ ▾ 翻译此页

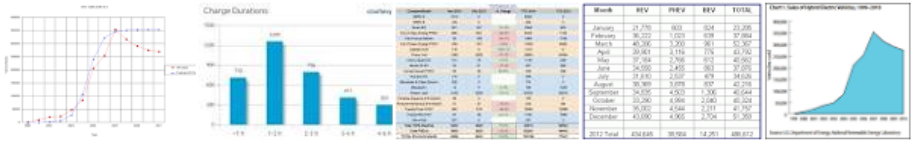
2016年1月7日 - Looking for the latest statistics on the electric car market? ... Third party use: this data can be used by third parties as long as the Next Green Car logo is displayed, the source .... US approves $14.7 bn VW emissions settlement.

Alternative Fuels Data Center: Maps and Data - U.S. HEV Sales by ...

www.afdc.energy.gov › AFDC ▾ 翻译此页

This chart shows the number of hybrid electric vehicles (HEVs), broken down by model, sold in the United States between 1999 and 2015. HEV sales surged in ...

electric cars statistic data + US的图片搜索结果 举报图片



Google 学术搜索

不限语言 中文网页 简体中文网页

关注以下作者所著文章

当这些作者写了新文章时，订阅提示通过 wsql2008@gmail.com 收到电子邮件通知

Viral Acharya
Reserve Bank of India, (On leave) CV Starr Professor of Economics, Department of Finance …

Jun Yu
Lee Kong Chian Professor of Economics and Finance, Singapore Management University

Subal C Kumbhakar
Distinguished Professor of Econimics, SUNY Binghamton, NY

Siem Jan Koopman
Professor of Econometrics, Vrije Universiteit Amsterdam

J. Isaac Miller
Department of Economics, University of Missouri

Ian Scoones
Professorial Fellow, STEPS Centre, the Institute of Development Studies, University of …

Google

三宝！谷歌图书

图书

我的书架

搜索我的图书馆 搜索

新书架

我的书架

我在 Google Play 上的图书 (7)

我的收藏 (10)

正在阅读的图书 (7)

计划阅读的图书 (16)

已读图书 (0)

您可能喜欢的图书

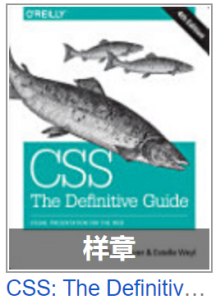
我的历史记录

在 Google Play 上选购图书

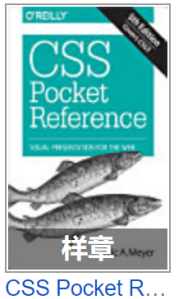
浏览世界上最大的电子书店，今天就开始网络、平板电脑、手机或电子阅读器上的阅读之旅吧！

立即转到 Google Play »

我在 Google Play 上的图书 - 私有图书



CSS: The Definitiv…
Eric A. Meyer, Este…



CSS Pocket R…
Eric A. Meyer

中国农业政策的…
张红宇, 赵长保

The Economics o…
Steven C. Blank

Wuthering Heights
Emily Bronte

The Three Musk…
Alexandre Dumas

Treasure Island, …
Robert Louis Ste…

已购图书 - 私有图书

此书架还没有书。了解详情。

评价过的图书 - 公开

收集二手数据：综合型数据平台(国内文献和数据)

国内文献和统计数据：

- 中国知网（内含统计年鉴资源）——学校图书馆网站
  - CNKI中国知网—CNKI中国期刊全文数据库
  - 中国知网-统计年鉴数据库
- 搜数网——学校购买暂时无访问权限
  - 新版搜数网- 中国资讯行
- 人大经济论坛：论坛币下载

收集二手数据：综合型数据平台(国外文献和数据)

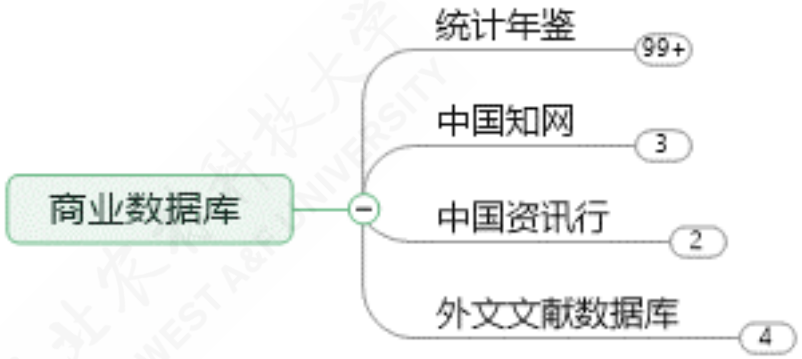
国外文献和统计数据：

- 电子期刊：SpringerLink电子期刊及电子图书
- 电子期刊：Wiley Online Library
- 电子期刊：ScienceDirect
- 电子期刊：Emerald
- 学位论文：ProQuest 学位论文全文库

收集二手数据：一个项目示例1/2

- 中国旱区农业科技资源配置研究

收集二手数据：一个项目示例2/2



收集二手数据：专项型数据平台（国外）

几个主要的数据来源：

- 美国大学联盟数据集成中心(ICPSR)。机构在密歇根，是世界上最大的学术数据源。
- 美国芝加哥大学-广泛社会调查(GSS)
- 美国芝加哥大学-收入动态调查面板数据(PSID)
- 美国密歇根大学-健康和退休调查数据(HRS)，公开自1990年
- 英国艾塞克斯大学-认识社会调查数据库(Understanding Society)。

收集二手数据：专项型数据平台（国内）

- 北京大学中国社会科学调查中心(ISSS)。主要的中国家庭追踪调查（CFPS）、中国健康与养老追踪调查（CHARLS）
- 中国人民大学中国调查与数据中心(NSRC)。主要的数据源有中国综合社会调查（CGSS）、中国教育追踪调查（CEPS）、中国老年社会追踪调查（CLASS）
- 中国疾病控制中心(CDC)。主要的数据源包括了慢病、流行病、艾滋病等多种涉及健康与疾病的调查。

示例1：中国家庭追踪调查（CFPS）的数据申请

示例1：中国家庭追踪调查（CFPS）的数据申请

CFPS | 中文 | English

CFPS China Family Panel Studies 中国家庭追踪调查

为了确保对CFPS数据本身及其知识产权的尊重和对受访者的保护，我同意遵守以下协议:

- ☐ 【同意】1.我有义务保护和尊重数据库中所牵涉受访者的隐私，不会探究、公开、或散布可能确认受访者身份的信息。
- ☐ 【同意】2.我承诺在本协议约定范围内任何使用CFPS数据的地方进行数据来源标示。具体格式如下：“本论文（书）使用数据全部（部分）来自北京大学和国家自然科学基金资助、北京大学中国社会科学调查中心执行的中国家庭追踪调查。
- ☐ 【同意】3.我不会将CFPS数据库的全部或一部分、或将其内容稍加修改以原名称或更换名称发表或转让给他人。
- ☐ 【同意】4.我不会以任何形式公布、发表所获之全部或部分CFPS数据库。
- ☐ 【同意】5.我只会将CFPS数据用于个人学术或政策研究活动，而不会用于任何盈利活动。

下一步　不同意

　地址：北京大学理科5号楼四层 总机：(010)62767908　电子邮件：isss.cfps@pku.edu.cn

示例1：中国家庭追踪调查（CFPS）的数据申请

示例1：中国家庭追踪调查（CFPS）的数据申请

示例2：中国健康与养老追踪调查（CHARLS）的数据申请

示例2：中国健康与养老追踪调查（CHARLS）的数据申请

二手数据收集：数据使用的几个问题

- 二手数据可以进行的反复多次的再分析。
  - 同样的数据集，使用不同的方法，可以进行检验或者商榷；
  - 同样的数据集，用于不同的研究主题和研究目的，则可以用于不同的研究目的。
  - 不同的数据集，不同的方法，可以以达成特定的研究目的。
- 使用二手数据，应按照学术规范说明数据来源。（千万别忘记！）
- 使用二手数据，往往面临数据处理、转换、加工等技术性的问题。
  - 参考哈佛大学和MIT联合建立的定量社会科学研究中心IQSS。
- 使用**综合性数据库**还是**专门性数据库**，这是个问题！
  - 综合性数据不一定能够满足专业兴趣的要求和需求。
  - 专业性数据库可能比较专业，难以与你的研究目标一致。

示例：农产品市场价格数据的爬虫自动抓取I

西安市农业农村局（网址）每日会发布不同农产品、多个市场（批发市场、零售市场如超市等）的农产品价格数据。

网站**页面分析**：

- 不需要登陆权限；
- 但是不具备table元素。

编程爬虫自动化数据**抓取方案**：

- 按页数（选择最大页），循环抓取页面
- 从最后一页开始抓取页面，然后依次往前一页抓取页面
- 每次抓取页面，把数据表**增量式**写入数据库`data/market -xian/market.db`的相应table中去

示例：农产品市场价格数据的爬虫自动抓取2

示例：农产品市场价格数据的爬虫自动抓取3（动图）

收集调查数据：自填式问卷调查

自填式问卷调查：没有调查员协助的情况下由被调查者自己完成调查问卷 问卷递送方法：调查员分发、邮寄、网络、媒体

优点：要求调查问卷结构严谨，有清楚的说明

缺点：

- 问卷的返回率比较低
- 不适合结构复杂的问卷
- 调查周期比较长
- 数据搜集过程中出现的问题难于及时采取调改措施

收集调查数据：面访式问卷调查

面访式问卷调查：调查员与被调查者面对面提问、被调查者回答的一种调查方式。

优点：

- 可提高调查的回答率
- 可提高调查数据的质量
- 能调节数据搜集所花费的时间

缺点：

- 调查的成本较高
- 调查过程的质量控制有一定难度

收集调查数据：电话式问卷调查

电话式问卷调查：通过电话向被调查者实施调查。

特点：

- 速度快，能在短时间内完成调查
- 适合于样本单位十分分散的情况

局限性：

- 如果被调查者没有电话，调查将无法实施
- 访问的时间不能太长
- 使用的问卷需要简单
- 被访者不愿意接受调查时，难以说服

收集调查数据：一个小结

特征	自填式	面访式	电话式
调查时间	慢	中等	快捷
调查费用	低	高	低
问卷难度	要求容易	可以复杂	要求容易
有形辅助物的使用	中等利用	充分利用	无法利用
调查过程控制	简单	复杂	容易
调查员作用的发挥	无法发挥	充分发挥	一般发挥
回答率	最低	较高	一般

> 有时间大家可以先看看“调查问卷设计”和“市场调研”相关图书！

收集实验数据

此处略！

本节结束