社会统计学及SPSS软件应用
STATISTICS WITH SPSS

Instructor:王荣欣

Email: rxwang@qq.com

周一3-4节、单周周四3-4节，3A106-2

September 13, 2020

Example 1

某社会工作者在主管一个低收入人口就业能力提升项目时，该项目在运行3年后需要评估其实际效果，用实际证据证明该项目是有效的，以获取后期持续的资助。

Example 2

当你进行一个社区资源整合的项目时，你最需要做的事情就是展开一项科学调查，以评估社区潜在的社会工作服务需求。

社工为什么要学研究方法和统计学？

1. 循证实践（evidence-based practice，简称EBP）或基于证据或证据为本的实践
   - 1966年，Donald Campbell和Julian Stanley出版Experimental and Quasi-experimental Designs for Research。（参考教材第40页）
   - 20世纪80年代来源于医学。
   - 证据来自观察或实验。证据可以是测量的数字，也可以是访谈的文本资料、图像等。
   - EBP结合最优科学实践、实践者经验和服务对象本身的独特性。
2. 知识的生产者
   - 证据为本的干预
   - 项目评估（e.g.留守儿童的项目）

文科生为什么要学习编程？

1 “新文科”概念最早由美国希拉姆学院在2017年提出，要求将新技术融入到哲学、文学和语言等课程之中，实现跨学科的融合和交流。

2 2009 年，哈佛大学David Lazer等15 位美国学者在《Science》上联合发表了一篇具有里程碑意义的文章“Computational Social Science”。

SOCIAL SCIENCE

Computational Social Science

A field is emerging that leverages the capacity to collect and analyze data at a scale that may reveal patterns of individual and group behaviors.

David Lazer,[1] Alex Pentland,[2] Lada Adamic,[3] Sinan Aral,[2,4] Albert-László Barabási,[5] Devon Brewer,[6] Nicholas Christakis,[1] Noshir Contractor,[7] James Fowler,[8] Myron Gutmann,[3] Tony Jebara,[9] Gary King,[1] Michael Macy,[10] Deb Roy,[2] Marshall Van Alstyne[2,11]

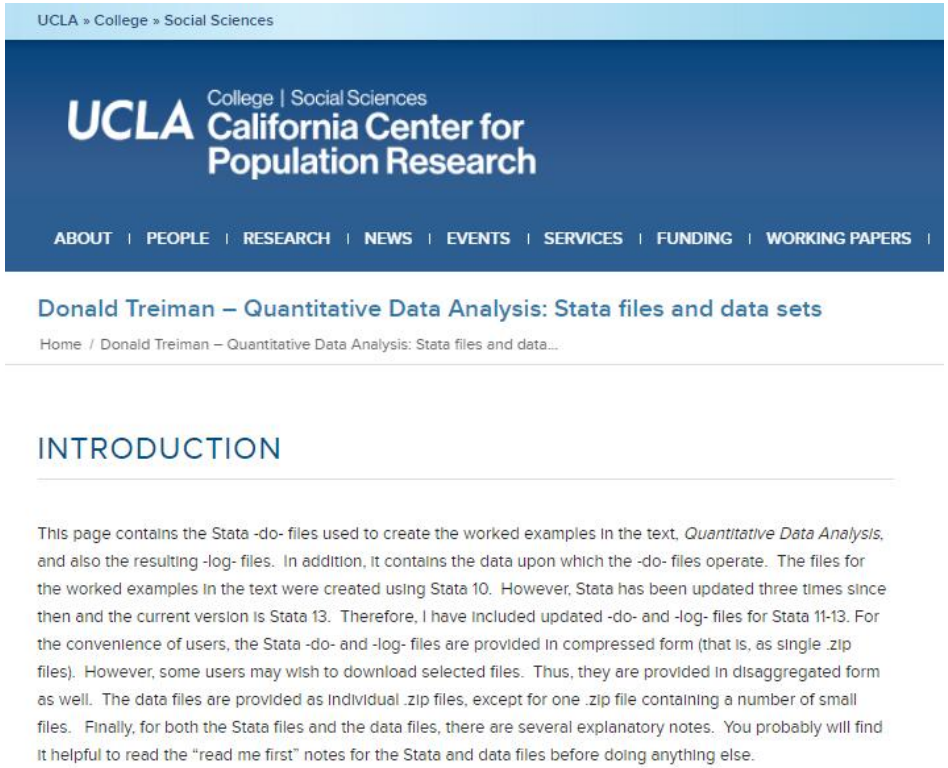
We live life in the network. We check our e-mails regularly, make mobile phone calls from almost any location, swipe transit cards to use public transportation, and make purchases with credit cards. Our movements in public places may be captured by video cameras, and our medical records stored as digital files. We may post blog entries accessible to anyone, or maintain friendships through online social networks. Each of these transactions leaves digital traces that can

ment agencies such as the U.S. National Security Agency. Computational social science could become the exclusive domain of private companies and government agencies. Alternatively, there might emerge a privileged set of academic researchers presiding over private data from which they produce papers that cannot be

critiqued or replicated. Neither scenario will serve the long-term public interest of accumulating, verifying, and disseminating knowledge.

What value might a computational social science—based in an open academic environment—offer society, by enhancing understanding of individuals and collectives? What are the

课程资料下载



http://www.ccpr.ucla.edu/dtreiman
https://ccpr.ucla.edu/donald-treiman-quantitative-data-analysis-stata-files-and-data-sets

CONTENTS

CONTENTS

第一讲内容

- 第1章第1-8页。
- 第9章第184-200页。

The word ”statistics” is derived from the Latin status, which led to the Italian word statista, first used in the 16th century, referring to a statist or statesman - someone concerned with matters of the state.

The Germans used Statislik around 1750, the French introduced statistique in 1785 and the Dutch adopted statistiek in 1807.

The system was first used in 17th century England by the London merchant John Graunt (1620-74) and the Irish natural philosopher William Petty (1623-97).

In the 18th century, many statists were jurists.

“变量”（变项）的英文名——variable, 意思是“可以变化的”。

1. 变量的本质就是“具有可变化特征的因素”（A variable is any factor that takes on a varying characteristic, Schwester, 2015）
2. 或者是“对某个一特征的测量”(A variable is an empirical measurement of characteristics, Babbie, 2013)

如何测量以下变量？

1. 年龄
2. 学习成绩
3. 教育获得（educational attainment）
4. 大学毕业生对学校满意度
5. 公司影响力
6. 社会分层中的阶层、地位和权力（class, status, and power），教材第230页
7. 态度研究中的失范、异化和权威主义（anomie, alienation, and authoritarianism）
8. 政治社会学研究中的自由主义和保守主义（liberalism versus conservatism）

测量的步骤

1. 概念化（conceptualization）：确定概念的定义。The first requirement for devising a valid measure is to be clear about what you are trying to measure.
2. 名义定义（nominal definition）
3. 操作化（operationalization）
4. 进行具体的测量：measurements in real world

SOCIAL CLASS

1. 概念化（conceptualization）：social class（教材第231页）
2. 名义定义（nominal definition）：define social class as representing economic difference, specifically, income.
3. 操作化（operationalization）：measure economic difference by income.
4. 进行具体的测量：”what is your income, before taxes, last year?”

1. We survey the sample by asking each individual in the sample a set of questions and recording the responses. The person being surveyed is asked to choose the best response from a list. （教材第3页）
2. Each response has a number associated with it, known as a code.
3. The term variable refers to each set of response categories and the associated codes.
4. Variables are characteristics of an individual or a system that can be measured or counted. These can be vary over time or between individuals.
   变量在测量以后，就会具有不同的数值或属性。

什么是变量值(variable value)?

1. 变量值，就是一个变量所描述的特征或者数量。
2. 一个变量总是对应着多于一个value（因为只对应一个value的叫constant, 常数）。

https://zhuanlan.zhihu.com/p/33589825

LEVELS OF MEASUREMENT

- **Levels of measurement （测量层次）**
  (1) nominal level of measurement (eg:婚姻状况、性别、民族背景)
  (2) ordinal level of measurement (eg:受教育程度)
  (3) interval level of measurement (eg: 收入、年龄、智商)
  (4) ratio level of measurement
- **Variables can be classified into two groups:**
  (1) categories
  (2) quantities

测量的VALIDITY和RELIABILITY

- valid（切实），是指被测量的属性确实是研究者意图测量的属性。举例：教育程度，用上学年数测量。
- reliable（可靠），是指测量结果经得起重复检验。

1. Sample: Pearson's closest friend, the Darwinian zoologist W.D.E. Weldon, began to use the term "sample" to refer to collections of observations of marine organism, through he wondered if his sample size was large enough.

2. Population: Pearson used the word "population" four years later to replace the term "normal group" and aligned population with sample in 1903.

back

1. A population is a technical term for the whole group of organisms or objects.
2. A population represents all conceivable observations of a particular type, whereas a sample is a limited number of observation from the population.
3. The best example of using an entire population is the decennial Census count.
4. In most studies, the population in which one is interested is far too large to measure each and every one of its members.
5. The statistician usually confines his analysis to a relatively small section of total population.

1. Statistics, which use Roman letters such as s and r (for the standard deviation and correlation) were developed primarily be Pearson.
2. Parameter,denoted instead by Greek letters such as μ (mu), σ (little sigma) ρ (rho) were introduced by Fisher in 1922 to estimate the mean, standard deviation and correlation in population.
3. Hence, statistics are to samples what parameters are to populations.
4. When these items measure a population, they are called parameters. If they are measures of a sample, they are called statistics.

1. Statistics ($\bar{y}$, s or r) are to samples what parameters (μ, σ, ρ) are to populations.

2. Every sample drawn from the population has its own statistics ($\bar{y}$, s or r), which is used to estimate the parameter (μ, σ, ρ) of its population.

问卷调查的方法

1. telephone surveys
2. mail surveys
3. web surveys

SAMPLING METHOD

1. Probability sampling
   (1) Random sampling
   (2) Systematic sampling
   (3) Stratified sampling

2. nonprobability sampling
   (1) Incidental sampling
   (2) Purposive sampling

- 1. Random: This is analogous to putting everyone's name into a hat and drawing out several names. Everyone in the original population has an equal and independent chance of being included in the sample. Although this is the preferred way of sampling, it requires a complete list of every element in the population, which isn't always possible. A table of random numbers in statistics books or those generated by computers and some phone systems can be used.
- 2. Systematic: This also requires the entire listing of a population, but here it's divided into blocks where every nth person is selected from the list (e.g. by taking every 10th person from an alphabetized list).

- 1. Random: This is analogous to putting everyone's name into a hat and drawing out several names. Everyone in the original population has an equal and independent chance of being included in the sample. Although this is the preferred way of sampling, it requires a complete list of every element in the population, which isn't always possible. A table of random numbers in statistics books or those generated by computers and some phone systems can be used.
- 2. Systematic: This also requires the entire listing of a population, but here it's divided into blocks where every nth person is selected from the list (e.g. by taking every 10th person from an alphabetized list).

- 1. Random: This is analogous to putting everyone's name into a hat and drawing out several names. Everyone in the original population has an equal and independent chance of being included in the sample. Although this is the preferred way of sampling, it requires a complete list of every element in the population, which isn't always possible. A table of random numbers in statistics books or those generated by computers and some phone systems can be used.
- 2. Systematic: This also requires the entire listing of a population, but here it's divided into blocks where every nth person is selected from the list (e.g. by taking every 10th person from an alphabetized list).

- 3. Stratified: The investigator selects a specific characteristic in the sample that he thinks is important for the research and then divides the sample into non-overlapping groups or strata, such as age-groups, gender, geographical areas or political affiliation. This can be used with the other four sampling procedures.
- 4. Incidental: Uses the most accessible and available sample by taking the most convenient set of subjects. It is the most reliable type of sampling.
- 5. Purposive: The experimenter chooses the subjects to be used because he thinks they are representative.

- 3. Stratified: The investigator selects a specific characteristic in the sample that he thinks is important for the research and then divides the sample into non-overlapping groups or strata, such as age-groups, gender, geographical areas or political affiliation. This can be used with the other four sampling procedures.
- 4. Incidental: Uses the most accessible and available sample by taking the most convenient set of subjects. It is the most reliable type of sampling.
- 5. Purposive: The experimenter chooses the subjects to be used because he thinks they are representative.

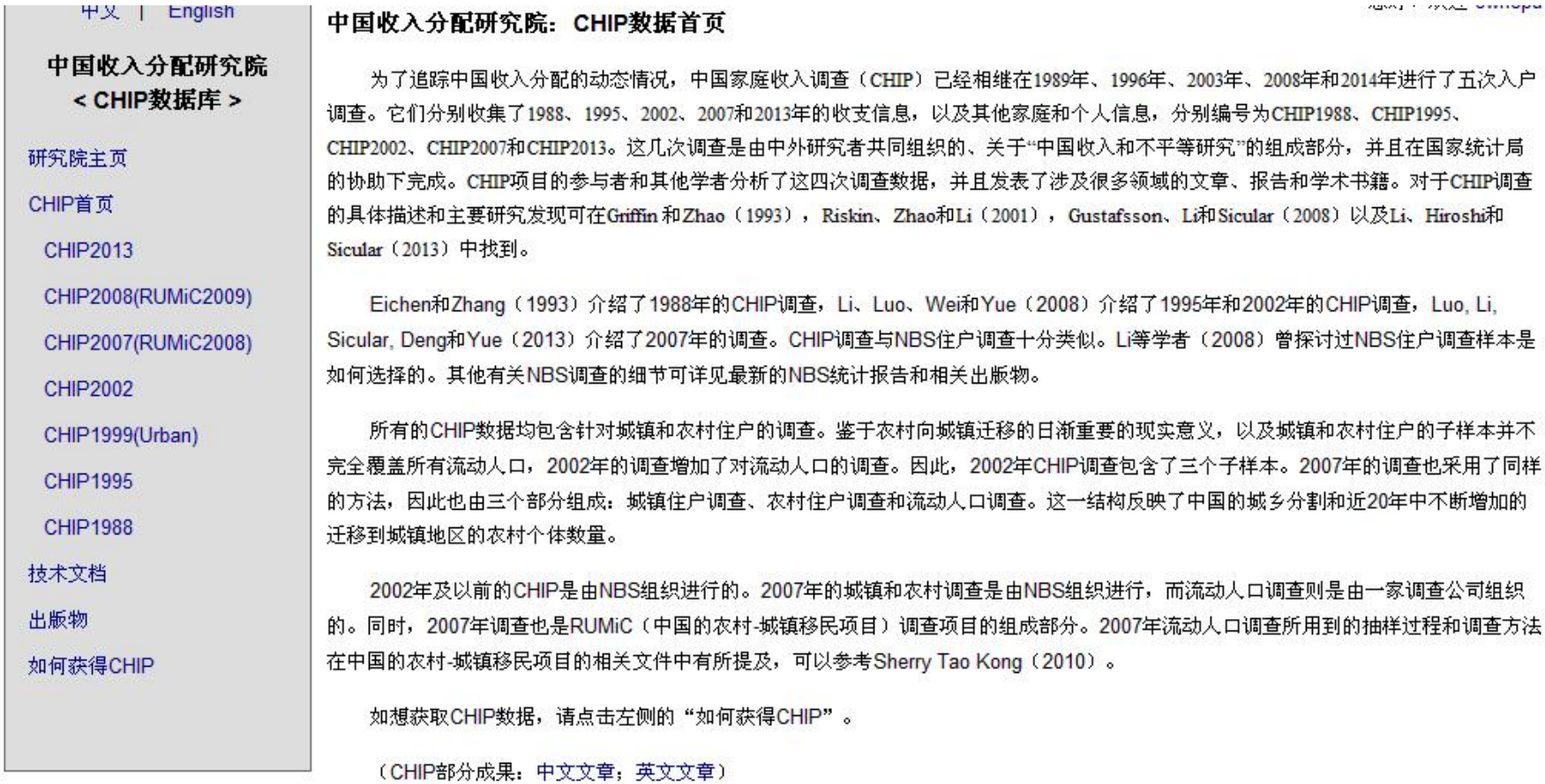
- 3. Stratified: The investigator selects a specific characteristic in the sample that he thinks is important for the research and then divides the sample into non-overlapping groups or strata, such as age-groups, gender, geographical areas or political affiliation. This can be used with the other four sampling procedures.
- 4. Incidental: Uses the most accessible and available sample by taking the most convenient set of subjects. It is the most reliable type of sampling.
- 5. Purposive: The experimenter chooses the subjects to be used because he thinks they are representative.

以下研究问题如何抽样？

1. 橘猫更容易发胖吗？
2. 咖啡喝太多会导致皮肤变黑吗？
3. 女生是否比男生拥有更强的语言能力？
4. 网络课程是否比教室授课更有效？

SAMPLING BREAKDOWN

1. Who do you want to generalize to? (The Theoretical Population)
2. What population can you access to? (The Study Population)
3. How can you get access to them? (The Sampling Frame)
4. Who is in your study? (The Sample)



中文 | English

中国收入分配研究院
< CHIP数据库 >

研究院主页
CHIP首页
CHIP2013
CHIP2008(RUMiC2009)
CHIP2007(RUMiC2008)
CHIP2002
CHIP1999(Urban)
CHIP1995
CHIP1988
技术文档
出版物
如何获得CHIP

中国收入分配研究院：CHIP数据首页

为了追踪中国收入分配的动态情况，中国家庭收入调查（CHIP）已经相继在1989年、1996年、2003年、2008年和2014年进行了五次入户调查。它们分别收集了1988、1995、2002、2007和2013年的收支信息，以及其他家庭和个人信息，分别编号为CHIP1988、CHIP1995、CHIP2002、CHIP2007和CHIP2013。这几次调查是由中外研究者共同组织的、关于"中国收入和不平等研究"的组成部分，并且在国家统计局的协助下完成。CHIP项目的参与者和其他学者分析了这四次调查数据，并且发表了涉及很多领域的文章、报告和学术书籍。对于CHIP调查的具体描述和主要研究发现可在Griffin和Zhao（1993），Riskin、Zhao和Li（2001），Gustafsson、Li和Sicular（2008）以及Li、Hiroshi和Sicular（2013）中找到。

Eichen和Zhang（1993）介绍了1988年的CHIP调查，Li、Luo、Wei和Yue（2008）介绍了1995年和2002年的CHIP调查，Luo, Li, Sicular, Deng和Yue（2013）介绍了2007年的调查。CHIP调查与NBS住户调查十分类似。Li等学者（2008）曾探讨过NBS住户调查样本是如何选择的。其他有关NBS调查的细节可详见最新的NBS统计报告和相关出版物。

所有的CHIP数据均包含针对城镇和农村住户的调查。鉴于农村向城镇迁移的日渐重要的现实意义，以及城镇和农村住户的子样本并不完全覆盖所有流动人口，2002年的调查增加了对流动人口的调查。因此，2002年CHIP调查包含了三个子样本。2007年的调查也采用了同样的方法，因此也由三个部分组成：城镇住户调查、农村住户调查和流动人口调查。这一结构反映了中国的城乡分割和近20年中不断增加的迁移到城镇地区的农村个体数量。

2002年及以前的CHIP是由NBS组织进行的。2007年的城镇和农村调查是由NBS组织进行，而流动人口调查则是由一家调查公司组织的。同时，2007年调查也是RUMiC（中国的农村-城镇移民项目）调查项目的组成部分。2007年流动人口调查所用到的抽样过程和调查方法在中国的农村-城镇移民项目的相关文件中有所提及，可以参考Sherry Tao Kong（2010）。

如想获取CHIP数据，请点击左侧的"如何获得CHIP"。

（CHIP部分成果：中文文章；英文文章）

Figure 4.1: 中国家庭收入调查

中国时间利用调查（2017）

“中国时间利用调查”是一项对中国城乡居民时间利用的全国性追踪调查，以知晓人们在何时从事何种类型的人类活动，在哪里从事这些活动，以及从事这些活动时与谁在一起。

- 《时间都去哪儿了？中国时间利用调查研究报告》，中国社会科学出版社，2018。
- 《时间都去哪儿了:2018年时间利用调查统计数据》，中国统计出版社，2019。

Example 3

2010年中国综合社会调查（Chinese General Social Survey，简称为CGSS），见《CGSS 2010 调查手册》第5页

第二部分、抽样说明

1. 中国综合社会调查（CGSS）第二期抽样方案简介

CGSS 第二期的抽样设计采用多阶分层概率抽样设计，其调查点覆盖了中国大陆所有省级行政单位。在全国一共抽取了 100 个县（区），加上北京、上海、天津、广州、深圳 5 个大城市，作为初级抽样单元。其中在每个抽中的县（区），随机抽取 4 个居委会或村委会；在每个居委会或村委会又计划调查 25 个家庭；在每个抽取的家庭，随机抽取一人进行访问。而在北京、上海、天津、广州、深圳这 5 个大城市，一共抽取 80 个居委会；在每个居委会计划调查 25 个家庭；在每个抽取的家庭，随机抽取一人进行访问。这样，在全国一共调查 480 个村/居委会，每个村/居委会调查 25 个家庭，每个家庭随机调查 1 人，总样本量约为 12000。其中，在抽取初级抽样单元（县区）和二级抽样单元（村委会和居委会），利用人口统计资料进行纸上作业；而在村委会和居委会中抽取要调查的家庭时，则采用地图法进行实地抽样；在家庭中调查个人时，利用 KISH 表进行实地抽样。

Example 4

2016年中国劳动力动态调查（Chinese Labor-force Dynamics Survey，简称为CLDS）见《中国劳动力动态调查手册20160615》第3页

CLDS 聚焦于中国劳动力的现状与变迁，内容涵盖教育、工作、迁移、健康、社会参与、经济活动、基层组织等众多研究议题，是一项跨学科的大型追踪调查。为保证样本的全国代表性，CLDS 的样本覆盖了中国 29 个省市（除港澳台、西藏、海南外），调查对象为样本家庭户中的全部劳动力（年龄 15 至 64 岁的家庭成员）。在抽样方法上，采用多阶段、多层次与劳动力规模成比例的概率抽样方法（multistage cluster, stratified, PPS sampling）。在追踪调查方式上，CLDS 在国内率先采用轮换样本追踪方式，既能较好地适应中国剧烈的变迁环境，同时又能兼顾横截面调查和追踪调查的特点。

Example 5

中国家庭动态跟踪调查（China Family Panel Studies，简称为CFPS），见《中国家庭动态跟踪调查(2010)用户手册（第二版）》第12页-19页、《中国家庭动态跟踪调查抽样设计》第14页-第15页

CFPS 最初目标样本规模为 16000 户，其中，有 8000 户从上海、辽宁、河南、甘肃、广东五个独立子样本框（称为“大省”）过度抽样（oversampling）得到，每个“大省”1600 户。另有 8000 户则从其他 20 个省份共同构成的一个独立子样本框（称为“小省”）抽取。5 个“大省”的子样本具有地区自代表性，可以进行省级推断以及地区间比较。5 个“大省”样本框在二次抽样后，与“小省”样本框共同构成具全国代表性的总样本框。[13]

附表 1. 10 个区县的 PPS 抽样示例

区/县 i	区/县 i 的人口数，M_i	累计人口数，T_i	抽取的编码 $R+(j-1)K$
1	1160	1160	
2	18160	19320	
3	8360	27680	22020
4	8840	36520	
5	12300	48820	
6	39440	88260	69486
7	12260	100520	
8	14680	115200	
9	10280	125480	116952
10	16920	142400	

$$T=\sum_{i=1}^{10}M_i=142400,\ n=3,\ \text{且}\ K=\frac{T}{n}=47466\ \text{(对}K\text{四舍五入取整)}$$

在 1 到 K 的范围内随机选取一个整数，比如 22,020，那么包含 $R=22020$、$R+K=47466+22020=69486$、以及 $R+2K=47466+44040=91506$ 的区县（见 T_i 列）将进入样本。换言之，区县 3、6、9 被抽中。

本节课作业

下载如下三个全国性抽样调查中的若干年份的数据。

1. 中国综合社会调查（Chinese General Social Survey，缩为CGSS）
2. 中国家庭动态跟踪调查（China Family Panel Studies，简称为CFPS）
3. 中国劳动力动态调查（Chinese Labor-force Dynamics Survey，简称为CLDS）

参考文献

- Magnello, Eileen and Borin Van Loon. 2009. Introducing Statistics: A Graphic Guide. London: Icon Books Ltd.